

ASSESSMENT: Common Fund Data Coordination Centers

*A Report Assessing The Readiness For
Accessing, Sharing and Analyzing Data Assets
Across the Common Fund Data Ecosystem*

C. Titus Brown
Amanda Charbonneau
Owen White

July 2019



Index

Index	1
Executive Summary - Common Fund Data Ecosystem Assessment	3
MAIN REPORT	6
Introduction	6
Our Assessment	6
General description of the DCCs; commonalities and differences	6
Deep dives	10
Opportunities and Challenges for Individual DCCs, summarized	13
Opportunities and Challenges across the DCCs, summarized	14
CFDE Resources	16
Recommendations for Resource Allocation and Investment by the Common Fund	17
Recommendation 1: Support the current DCCs with targeted investments.	17
Recommendation 2: Support the current DCCs with cross-DCC investments.	19
Recommendation 3: Support a shift in current CFDE activities to support transformative activities.	20
Recommendation 4: Invest in new transformative activities by the current CFDE Team	23
Recommendation 5: Invest in long-term ecosystem support with targeted RFAs.	26
Risks and Challenges	27
Acknowledgements	28
Appendices	29
Appendix A - Site Visit And Survey Methodologies	29
Appendix B - Descriptive Overview Of All Dccs	31
Appendix C - GTEx Site Visit	33
Appendix D - Kids First Site Visit	49
Appendix E - HMP Interview	69
Appendix F - LINCS Interview	79
Appendix G - CFDE tech team deliverables to December 2019	92
Appendix H - Cost estimators	93
Appendix I - GTEx and Kids First joint exercise	97
Appendix J - Single sign on and authorization assessment	99

Executive Summary - Common Fund Data Ecosystem Assessment

Preamble

In early 2019, we were charged with assessing the opportunities and challenges that Common Fund DCCs face with respect to making their data more accessible and usable within and between CF programs. Identifying and solving issues that inhibit data access and reuse will lead to enhanced utility of CF data, both during the CF support period and after a CF program has ended. Moreover, we wish to lay the foundation for interoperability that will enable work across one or more CF programs. Our effort identifies many of the vital elements needed to build a comprehensive Common Fund digital ecosystem.

In order to achieve this goal we initiated a comprehensive assessment of all of the CF DCCs specifically targeting their current FAIRness level, infrastructural elements such as data storage, support for clinical users, and their ability to provide access to human subjects data. This intermediate report summarizes the results of our initial in-depth assessment of 4 DCCs: Kids First, GTEx, HMP and LINCS. We make recommendations on how to support the DCCs, establish a Common Fund Data Ecosystem (CFDE) within and across the Common Fund DCCs, and lay the groundwork for integration with other NIH datasets.

This assessment was generated from a combination of systematic review of online materials, in-person site visits to the Genotype Tissue Expression (GTEx) DCC and Kids First, and online interviews with Library of Integrated Network-Based Cellular Signatures (LINCS) and Human Microbiome Project (HMP) DCCs. Comprehensive reports of the site visits and online interviews are available in the appendices. We summarize the results within the body of the report.

The Common Fund DCCs

The Common Fund DCCs store and provide data derived from hundreds of studies and samples collected from thousands of human subjects. An incredible diversity of datatypes has been generated at the genomic, expression, proteomic, metagenomic, and imaging levels, and the DCCs support a tremendous range of scientific discovery efforts.

However, the present ability of a clinical or biomedical researcher to use the resources generated by the Common Fund is poor. The resources are hard to search collectively and not readily usable in combination. *Rapidly* making use of these resources is particularly challenging because they are hosted across multiple Data Coordination Centers (DCCs). At our visit with the Kids First DCC, we heard the following story:

Members of Kids First were contacted by a doctor who had 24 hours to enroll a patient in a clinical trial. Using the Kids First platform, they compared the patient's

genome to variants hosted on their platform, *and reviewed these results with information hosted at GTEx*. As a result, the clinician was able to identify additional therapeutic avenues.

This was only possible because of unique circumstances. Kids First had already gathered the data from GTEx and reprocessed it in a way that made it possible to compare the results to their own data. This process required months of work, advance planning and significant bioinformatics expertise on the part of Kids First, and would have been an impossible task for the clinician they helped. This highlights the main result of our assessment: **the datasets hosted by the DCCs are not inherently interoperable, and placing their assets in the cloud does not intrinsically solve the problems of findability, accessibility, interoperability, and reusability**. Further, even if this combined dataset was made available, most researchers don't have the bioinformatic skills to use it: **researcher training is needed to support data use**.

It was apparent from our assessment that the DCCs have many needs, some of which are specific to one DCC and others which are shared among DCCs. These include needs for enhanced protected data access (GTEx), long term data storage support (GTEx, KF, LINCS, HMP), training (GTEx and KF), the ability to export collections of data from their portal (GTEx, KF), and support for their data and data portals past the end of the Common Fund Program lifecycle (HMP, GTEx and LINCS). We also found advanced capabilities at DCCs that could be reused by other DCCs (e.g. RNA-Seq and eQTL visualization at GTEx, and a strategy for accessing protected data at KF). We anticipate discovering additional needs and opportunities in our next set of interviews.

We also found that a transformative opportunity exists to operationalize FAIRness across the DCCs, to permit translational impact by improving data discovery, access, and reuse. While the individual Common Fund projects excel at making their data FAIR within each DCC, there is little cross-DCC dataset FAIRness. For example,

- There is no systematic way to identify what data is hosted at the CF DCCs, which makes individual datasets hard to Find.
- In the absence of a standard way to access protected datasets at dbGaP, much of the underlying phenotype data cannot be Accessed.
- There is no standard way to transfer collections of data from multiple DCC portals to a single analysis system like Broad's Terra or Kid First's Cavatica, which inhibits Reuse of data.
- Interoperability of data across DCCs cannot be evaluated in depth without actually performing analyses between DCCs, which relies on resourcing pilot studies.

Operationalizing FAIR principles across the DCCs would be transformative because it would permit researchers and clinicians to *rapidly* and *routinely* leverage multiple Common Fund datasets in their work, just as Kids First has done with GTEx.

Operationalizing FAIR will require cross-DCC solutions. For example, a cross-Common Fund portal would enable the Findability of related datasets in different DCCs. This portal would ideally rely on an asset inventory distributed by the DCCs in a common format, so that the portal could automatically update as new data is released. The portal, and its underlying standards, will accelerate discovery and decrease time to impact of Common Fund-supported research. Moreover, long-term sustainability of the datasets would also be enhanced by common descriptors and common access methods.

This tremendous opportunity comes with challenges. In order for cross-DCC data findability, accessibility, reuse and interoperability to work in practice, all of the DCCs must participate. This will involve a significant investment of time and energy that needs to be supported and incentivized by the Common Fund leadership. This transformation could be driven by using FAIRness evaluation as an organizational tool, and making investments in both incremental and transformative change at each DCC. There is also an important role for a continued trans-DCC effort that will engage with the DCCs, drive iteration of standards through FAIRness metrics, and implement missing technical solutions.

To address these challenges, our recommendations for resource allocation are as follows:

1. Support individual DCC needs with targeted investments, e.g. fund data storage.
2. Invest in common DCC needs, including a standard method for authentication/access to protected data, training, lifecycle support, and FAIRness metrics.
3. Continue the Common Fund Data Ecosystem's work to support transformation at the ecosystem level, including building a common portal, developing an asset specification, and supporting a pilot data reuse postdoc.
4. Invest in additional work by the Common Fund Data Ecosystem to drive transformational change with a common manifest format, FAIRness evaluation across multiple DCCs, and pilot data reuse projects between DCCs.
5. Consider RFAs for longer-term investment in supporting the CFDE ecosystem.

We have coalesced a consortium that is prepared to meet the challenges required to implement the CFDE. The flexibility of resource allocation and administrative oversight enabled by the OTA affords an unprecedented degree of effectiveness. This, in combination with an interdisciplinary group composed of NIH program representatives and technology experts -- and an engagement team who actively consults with the Common Fund DCCs -- has allowed us to adapt our original set of deliverables to rapidly meet the needs DCC community. Based on these experiences, and from what we have learned generating this assessment, we are confident we can operationalize the CFDE in the near future.

MAIN REPORT

Introduction

Common Fund has several large scale data coordination programs that include DCCs for the 4D Nucleome, GTEx, HMP, Kids First, LINCS, Metabolomics, HubMAP, MoTrPAC, and SPARC. One purpose of this report is to provide an overarching view of the content, status and maturity of the DCCs. Our specific charge was to evaluate the state of Common Fund DCCs to determine opportunities and challenges for increasing the value and usability of Common Fund datasets. The material in this report reflects several modes of engagement to better understand the Common Fund DCCs. This includes two site visits with GTEx and Kids First, two 3 hour teleconference interviews with HMP and LINCS, and webinar presentations to all groups. During Phase 1 of the DCPPC we also conducted working group calls on a weekly basis with the GTEx group. To gather additional data, we performed a thorough review of the websites and resources of all nine DCCs.

Below, we provide a summary of the results of our overall evaluation of the DCCs as well as our site visits. We then follow that summary with recommendations for investment in 5 categories:

- Supporting individual DCC needs
- Investing in common DCC needs
- Continuing the Common Fund Data Ecosystem project's work to support this transformation.
- Investing in additional work by the CFDE project to drive transformational change.
- Considering RFAs for longer-term work to support the Common Fund DCC at an ecosystem level.

We end the report with a summary of risks and challenges to the overall effort of improving FAIRness within and between the Common Fund DCCs.

Our Assessment

General description of the DCCs; commonalities and differences

The goal of this section is to compare and contrast the content, status and maturity of Common Fund DCCs. Information here was collected either by passive review of the websites and resources of the DCCs, or through personal contact with DCC staff (see [Appendix A, Methodology](#)).

First impressions. Collectively, the Common Fund DCCs:

- Range from just getting started to fully implemented;

- Possess a broad range of datatypes, many in common across all DCCs;
- Have many datatypes that are complementary with other coordination centers;
 - (e.g. variants and expression data)
- Are strongly motivated by FAIRness, but vary in their implementations;
- Each utilize a different data model to host their assets, which is vital to their operations;
- Largely do not use common metadata vocabularies;
- Mostly host data (four out of six sites), and also have data at dbGaP;
- Cannot perform queries across the data hosted at dbGaP;
- Use a variety of sophisticated and capable analytical tools specialized to their project;
- Provide training materials to their user community.

Other general information about all of the DCCs are the following:

Common Fund DCCs have vast data assets. The Common Fund DCCs possess data derived from hundreds of studies and samples collected from thousands of human subjects. A summary of the datatypes and studies hosted by each DCC is shown in Table 1: an incredible diversity of datatypes has been generated at the genomic, expression, proteomic, metagenomic, and imaging levels. At each DCC website, users can search through these data using a wide variety of facets, such as assay, project, tissue type, disease, patient variables. Information for individual genes is available for expression, epigenomics, variants, and chromatin organization.

	4D Nucleome	GTEx	HMP / iHMP	Kids First	LINCS	Metabolomics
Studies and Bulk File Content	730 experiment sets, 2107 experiments, 6823 files, 28.99 TB, 166 external datasets	53 Tissues, 960 donors 30,000 samples, data set size expected increase to a total of 600GB by next release	21 studies, 48 primary body sites, >32,000 samples, >118,000 files, 9.75TB (iHMP), 7.22TB (HMP)	11 studies, each with 250 - 2000+ subjects, 927.1TB	398 datasets, 100TB genomic data, >1PB imaging data	920 studies, 6.4TB (compressed zip files), 233 studies with restricted access
Dataset types	DNA FISH, RNA FISH, SPT, 2-stage Repli-seq, in situ Hi-C, single cell Hi-C, RNA-seq, ChIP-seq, DamID-seq, ATAC-seq, NAD-seq, ChIA-PET, DNA SPRITE, PLAC-seq, MARGI, RNA-DNA SPRITE, Micro-C, TSA-Seq	De-identified annotations, RNA-seq, single-tissue cis-eQTLs, multi-tissue eQTLs, single-cell data	Reference microbial genomes, whole metagenomic sequence, 16S metagenomic sequence	BAM, CRAM, fastq, VCF, clinical measurements	Binding, imaging, transcriptomics, proteomics, epigenomics	Raw/unprocessed NMR data, MS data, Processed data (general)

Table 1: Summary of bulk assets hosted by each DCC who host data publicly, and their relevant studies.

Common Fund DCCs exist on a continuum. Figure 1 shows approximate start and end dates of Common Fund funding for each of the DCCs; it is recognized that funding for some programs is continued by another NIH IC. One implication of the range in start dates is that DCCs vary widely in their data assets, depending on their stage of maturity. For example, HuBMap was launched in November 2018 and is not expected to be in production phase until 2022. At the other extreme, the HMP/iHMP DCC has completed 10 years of operation and is generating no

new data - its funding has been discontinued. The DCCs also vary in terms of their readiness to be operationalized on cloud-based systems. This is reflected in Table 2, which provides a short summary of data hosted at each site, their number of users, and whether each DCC is using a cloud-based system.

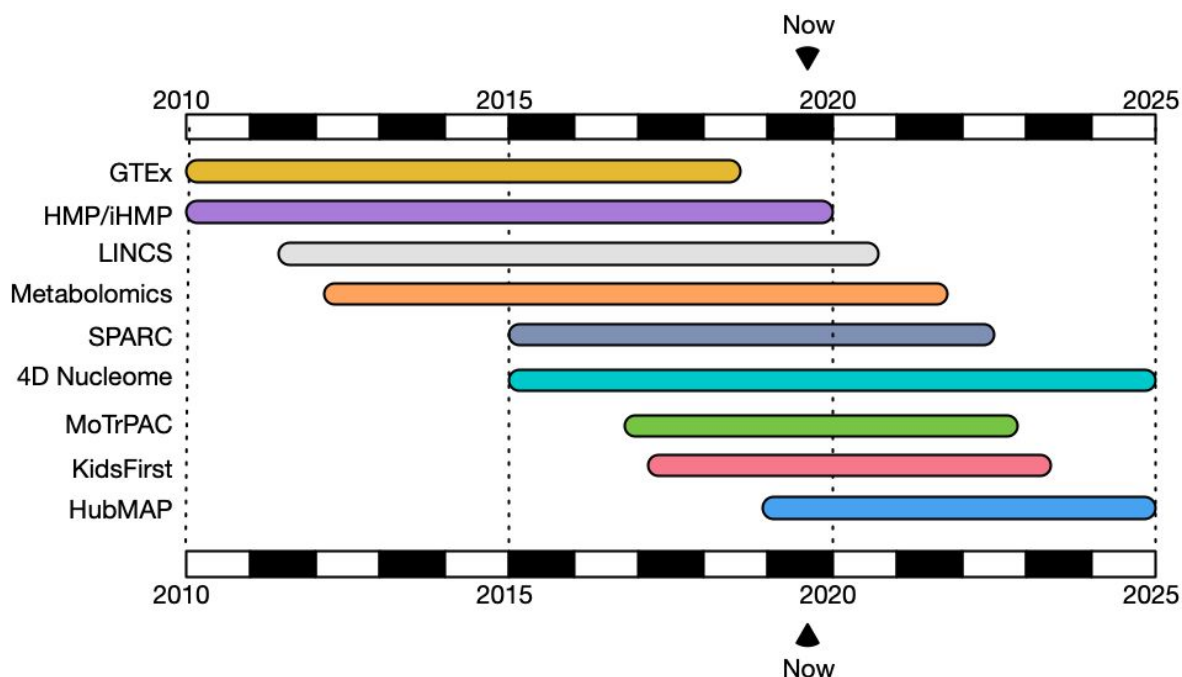


Figure 1: Approximate start and end dates of Common Fund funding for the Common Fund DCCs. Some dates may be inaccurate due to differences between funding approval and program start.

	4D Nucleome	GTEx	HMP / iHMP	Kids First	LINCS	Meta-bolomics	Hubmap	MoTrPAC	SPARC
API available	Yes	Yes	Yes	Yes	Yes	Yes	N/A	N/A	N/A
Data model documented	Yes	Yes	Yes	Yes	Yes	Yes	"	"	"
Training materials online	Yes	Yes	Yes	Yes	Yes	Yes	"	"	"
website user visits / month		~15,000	~20,000	~2000	~7,000	~1500	"	"	"
Total volume of data	28TB	600GB	10TB (iHMP) 7TB (HMP)	927TB	100TB genomic, >1PB imaging	6.4TB	"	"	"
Linked to cloud workspace	Yes	Yes	No	Yes	Yes	No	"	"	"
Cloud or local storage	Cloud	Cloud	Local	Cloud	Local	Cloud	"	"	"
Protected data hosted at dbGaP	No	Yes	Yes	Yes	Yes	No	"	"	"

Table 2: Approximate dataset size, numbers of users, controlled access usage, and additional resources hosted by Common Fund DCCs

Commonalities and complementarity of DCC assets. A comparison of data types across all DCCs is presented in Table 3; these data indicate that the same types of data are hosted between sites, and that data found between sites could be useful in combination. Whole

genome sequence, exome sequence, and transcriptional data were among the datatypes most frequently hosted by the DCCs. Several sites host data associated with human genes, which means that if properly linked, users could obtain expression, epigenetic, and variant information associated with specific gene regions. Metadata categories are also frequently similar across multiple sites. For example several sites host data associated with a body site, suggesting that queries such as "retrieve all datasets associated with skin samples" would return multiple datatypes from multiple DCCs. At least four DCCs host clinical information which suggests that CFDE users could obtain sets of different datatypes associated with disease and patient variables (e.g. body mass index, blood pressure) from across the Common Fund DCCs.

A.									
		4D Nucleome	GTEx	HMP / iHMP	HubMap	Kids First	LINCS	Metabolomics	MoTrPAC
Clinical Data		X	X		X	X			
Whole Genome/Exome Sequence		X	X		X			P	
Transcriptomics	X	X	X	P	X	X		P	P
Histology Images					X				
Radiology Images					X				
Metatranscriptomics			X					P	
Metaproteomics			X						
Marker Sequence Metagenomics			X					P	
Microbial Reference Genomes			X					P	
ChIPseq	X					X			
FISH	X			P					
ATACseq	X			P		X			
Hi-C	X								
ChIA-PET	X								
Proteomics			X	P		X		P	P
KINOMEscan						X			
Metabolomics			X	P			X	P	
Lipidomics				P					
scDNAseq				P					
Epigenomics			X	P		X		P	

B.				
		Systems	Organs	Cells
MoTrPAC	X	X		
SPARC	X	X		
HubMap		X	X	
LINCS			X	X
4D Nucleome			X	X
GTEx				X
KidsFirst				X
HMP/iHMP				X
Metabolomics				X

Table 3: Summary of DCC datatypes and subject matter. **(A)** Datatypes found across all sites. Assets that are currently available are represented by 'X'. Assets that are planned are indicated by 'P'. **(B)** Level of resolution, in terms of anatomy, cellular or molecular level, that are represented at each site.

FAIRness of the DCC assets is reasonable, but has room for improvement. We evaluated each Common Fund DCC that publicly hosts data for FAIRness. We employed a set of metrics developed by Wilkinson et. al, Scientific Data, 2016 and the results are shown in Table 4. The FAIR Principles described by Wilkinson place specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. Our analysis of the DCCs demonstrates that even though they do not follow a common set of FAIRness practices, each individual group has been very effective at improving the FAIRness of their data. All groups did very well with *findability* (hosting data in searchable

resources), supply free and open electronic protocols for *accessibility*, and support *reusability* by describing their data with accurate and relevant attributes. Interoperation measures scored reasonably well, but there was more room for improvement in this category than in the other metrics.

Additional overview information for all of the CF DCCs can be found in [Appendix B](#).

		4D Nucleome	GTEx	HMP / iHMP	Kids First	LINCS	Meta- bolomics
Findability. The DCC:	assigns globally unique and persistent identifiers	Yes	Partially	Yes	Yes	Partially	Yes
	enables discovery through rich metadata	Yes	Yes	Yes	Yes	Yes	Yes
	associates metadata with persistent identifiers	Yes	No	No	No	Partially	Partially
	registers or indexed (meta)data in a searchable resource	Yes	Yes	Yes	Yes	Yes	Yes
Accessibility. Electronic protocols at the DCC:	retrieve (meta)data using a standardized communication protocol	Yes	Yes	Yes	Yes	Yes	Yes
	are open, free and universally implementable	Yes	Yes	Yes	Yes	Yes	Yes
	allow for an authentication and authorization procedure	Yes	Yes	Yes	Yes	N/A	Yes
	access metadata even when the data are no longer available	Partially	Partially	Partially	Partially	Partially	Partially
Interoperability. (Meta)data at the DCC:	use a formal, accessible, shared, and broadly applicable knowledge representation	No	Yes	Yes	No	Partially	Yes
	use vocabularies that follow FAIR principles	No	Yes	Yes	No	Partially	Partially
	include qualified references to other (meta)data	Yes	Yes	N/A	N/A	Partially	N/A
Reusability: (meta)data at the DCC:	are richly described with accurate and relevant attributes	Yes	Yes	Yes	Yes	Partially	Yes
	are released with a clear and accessible data usage license	Partially	No	No	No	Yes	Partially
	are associated with detailed provenance	Yes	Yes	Yes	Partially	Yes	Yes
	meet domain-relevant community standards	Partially	Partially	Partially	Partially	Yes	Yes

Table 4: FAIRness assessment performed by manual, subjective review of each DCC. The CFDE tech team is currently developing objective and fully automated measures to be applied to each site going forward.

Deep dives

To date, the CFDE engagement team has interviewed four DCCs, which are at varying points in their funding lifecycle. Two of these engagements, Kids First and GTEx, were two-day, in-person interviews. Our other two interviews, with HMP and LINCS, took place via teleconference. (Additional information about our engagement methodology is in [Appendix A](#)). One important result of these engagements was to re-enforce how personalized visits play an essential role in building an effective working relationship with the DCC staff. Our visits were a very effective mechanism in establishing trust, understanding the goals of each DCC, creating incentives for the DCCs to participate with CFDE, and discovering important resources developed by the DCCs that could be utilized by other groups.

DCC site visits are a crucial part of creating an effective working relationship with the DCC staff. During our visits, we operated in a mode of listening, avoiding any discussion of specific CFDE implementation details; staff were then much more receptive to sharing how they imagined this project would unfold. There were also more emboldened to be honest about where the CFDE might fail. Because we are researchers who are independent from the NIH, staff were more open to communicate about challenges with NIH policies. Each visit helped us recalibrate our understanding of the goals and technical expertise of each DCC. Synthesizing this information after each visit pushed us to align our concept of the CFDE with the true needs of the DCCs. Overall, the site visits increased trust between the DCCs members and our group, which will be vital to successful operationalization of the CFDE.

While we were able to get baseline data from our interviews regardless of format, we found our in person visits to be much more productive. Our face to face conversations, and the level of trust they established, allowed us to discover important blockers and challenges faced by the DCCs that we could not have learned from any other means. Further, visiting over two days allowed us plenty of time to cover topics in-depth, while still allowing for breaks, downtime and review. DCC staff did much of the talking on day one of our meetings, and would be quite fatigued by early afternoon. We took this as a natural breakpoint, and moved our engagement team off-site to discuss what we had learned, and what CFDE might do to support the challenges the DCC had presented to us. On the second day, we were able to offer well-thought out suggestions for CFDE involvement, and to get feedback in real time. These interactions were vital to creating our vision for CFDE, and many of the recommendations in this document were shaped in those meetings. We believe it is essential to continue meeting on-site with DCC staff.

One clear lesson from our site visits is that new DCCs will face a wide variety of challenges. Each center we talked to had deep knowledge of challenges that are likely shared by other DCCs, and had solved many with innovative solutions. Unfortunately DCCs frequently operate in isolation from each other due to the initial burden of getting started, and continued pressure of serving teams of data generators. Staff at these sites are experts in the niche domain of running a DCC, and between them there is a vast wealth of institutional knowledge. The CFDE could leverage that knowledge, and increase cross-DCC interactions, by continuing outreach and engagement to the DCCs, serving as a centralized resource of information that is generated over time, and building a DCC community.

GTEx. Over 15,000 users come to their site each month and GTEx works hard to ensure that those researchers can easily accomplish their goals. GTEx is heavily focused on user experience. GTEx has done a tremendous amount of work to ensure that all of their data is uniformly processed, and has undergone rigorous quality assessment before being displayed in the portal. The GTEx portal allows a user to do a wide array of analyses, such as compare expression levels and plot PCAs, in a rich, interactive point and click interface. The richness and complexity of the GTEx dataset combined with their massive number of users results in an incredible support burden. Although some inquiries are about GTEx data specifically, most of

their time is spent answering questions such as ‘How do I compare my RNA-Seq with GTEx data?’, that are really questions of basic bioinformatics skills rather than about GTEx in particular. A planned update to their portal later this year will allow even more complex analysis of eQTLs, and GTEx is committed to increasing user accessibility, even if it increases their support burden. They told us that enabling user queries across datasets, for example between GTEx and Kids First, and improving user access to data hosted at dbGaP would be among their highest priorities when participating in CFDE. GTEx is also working to enable their users to perform analysis on Terra, the Broad’s data analysis platform. The complete GTEx report can be found in [Appendix C](#).

Kids First. The overarching goal of the Kids First DCC is to accelerate the pace of translational research, in order to impact the lives of children *right now*. Kids First takes a very pragmatic approach to every aspect of their DCC, and is interested in essentially any program, standard or collaboration that will increase the pace of research. Although they do not currently have a user training program, Kids First is eager to create one. The users for Kids First range from clinicians with no computational knowledge to bioinformaticians with no medical knowledge, and so require an equally broad training program. Getting such a training program up and running will allow more researchers to use the Kids First data, and to use it to ask more sophisticated questions, faster; which puts it near the top of Kids Firsts priority list. Kids First also expressed interest in making all of their data FAIR. They are very interested in a cross-DCC search capability and making their data interoperable with outside datasets. Again, broadening access and reducing barriers to finding and using their data is a top priority. Assuming CFDE can operationalize FAIR principles, Kids First is excited to improve their FAIRness and participate in assessments. Finally, Kids First would like to participate in a DCC community. Running a DCC is a specialized skill shared by only a handful of people. However, because the demands of daily operations far outweigh other priorities, it is rare when one DCC will be incentivized to work with another. The Kids First team pointed out that while our goal of de-siloing the data had merit, that people at DCCs need to be de-siloed as well. The complete Kids First report can be found in [Appendix D](#).

HMP. Funding for the HMP expired earlier this year, so while the portal and data are still accessible, there is no active work on the DCC. As such, much of our discussion was centered on how fundamental concepts in data sharing apply to retired data centers. For instance, in the first phase of HMP, they devoted a great deal of resources to ensuring that their pipelines were well documented and could be replicated by users, and they worked hard to build a consistent and complete metadata model. While the term FAIR and its attendant definitions didn’t exist until well after phase 1 of HMP was complete, the HMP was using many of the same principles to build their data center. However, the data at HMP is due to be moved to archival storage in the next few months as they no longer have funding to host it, and although the pipelines are well documented, many of their processing pipelines are based on software that should be updated. The HMP group has exerted significant resources towards metadata curation and documenting how their data was processed for the research community but it is unclear what will happen to that information once their data is moved to an archive. The HMP was also

interested in the idea of an avenue for cross-DCC collaborations and discussions. In fact, the HMP told us that they have been offering informal advice and mentoring to younger data centers, though not ones on CF's priority list. Given their lack of funding, the HMP indicated that moving their data to a stable, professionally managed file system (such as Google or Amazon cloud) was a top priority. They also noted that updating their pipelines to use more modern technology will be key to providing quality datasets to users going forward. The complete HMP report can be found in [Appendix E](#).

LINCS. The primary concern of the LINCS DCC is that their funding ends on June 30th, 2020, and in conjunction with their Program Officers, they are working hard to mitigate the effects of this impending hard stop. Common Fund DCCs are limited to ten years of funding, and receive their awards in equal sums across all years. LINCS told us that this made their first year difficult, as they couldn't onboard staff or ramp up projects quickly enough to use the funding, and were not allowed to carry forward an unobligated balance. Currently, LINCS supports over 50 portals and analysis apps, tens of terabytes of data, and a robust training program, all of which are popular with the community. No one wants all of those resources to just disappear, but the path forward is uncertain. LINCS' top priority is trying to get approved for a no-cost extension, or to use their old unobligated balance to extend their funding past June 2020. They have also begun work on the Signature Commons project, a philosophical continuation of the LINCS portal that emerged from their technology, and that could accept new funding under the new name; however, none of these solutions were guaranteed at the time of our meeting. All of this highlights that different stages of the DCC lifecycle have varying challenges, and that DCCs could benefit from targeted support in their early and late years. The complete LINCS report can be found in [Appendix F](#).

Opportunities and Challenges for Individual DCCs, summarized

The challenges described in this subsection, as well as the opportunities we have identified, are those that will require targeted funding to individual DCCs in addition to broader CFDE efforts. These challenges may be shared across DCCs, however the specific details of the challenge at each DCC are different enough that they require individualized implementations. This is as opposed to challenges that can be addressed by a generalized Common Fund-wide solution, and which are addressed in the next subsection. In some instances, such as user support, there are challenges at both the individual DCC and Common fund level.

Data maintenance and access. All of the DCCs reported having various funding issues regarding hosting, analyzing or providing access to their data. Although they have used different, non-cross-compatible implementations, user access to unprotected data is solved at each DCC. However, for example, GTEx reported that their current protected data access solution is non-functional. Targeted support is needed to make GTEx data accessible. Similarly, all of the DCCs reported being un-funded or under-funded for fees associated with cloud computing, however their specific needs vary. This challenge is addressed in Recommendations 1, 3, and 4.

User training needs. All four interviewed DCCs see user-focused training as a straightforward way to onboard more users to their datasets, speed discovery for clinical researchers, and improve the skill level of bioinformaticians. Users at each DCC require specific training tailored to that DCCs datasets. KF and GTEx have few ongoing training efforts and see it as a substantial need, while both LINCS and HMP have invested in training and see it as a path to enhanced use of their data. KF was particularly enthusiastic about the opportunities for enhanced clinical impact through training clinicians to use their portal. This challenge is addressed in Recommendations 1, 3, and 5.

Ageing infrastructure. HMP is no longer funded, and LINCS funding ends on June 30, 2020. Both DCCs use primarily local storage, and both told us that they are unsure about ongoing data hosting. In particular, the local servers that hold their data are being retired in early 2020. Unless they receive immediate funding to migrate that data to the cloud, and ongoing funding for cloud storage, their data will become inaccessible. This challenge is addressed in Recommendations 1, 3, 4, and 5.

Opportunities and Challenges across the DCCs, summarized

These challenges represent issues that are both widely faced by DCCs *and* that require a single Common Fund wide-solution or buy-in from multiple DCCs.

Asset specification incompatibilities. Each of the DCCs host many files (e.g., genomic sequence, metagenomic, RNA-Seq, physiological and metabolic data) and it is hard to discover these files across DCCs. Moreover, information describing the contents of the files is not available in a standard format. This prevents DCCs from making use of each other's data, makes the data less discoverable by others, and challenges interoperability. If the DCCs adopt a standardized Common Fund asset specification format, these problems could be solved. This challenge is addressed in Recommendations 1, 3 and 4.

Data access barriers. Both Kids First and GTEx reported that a critical need is to reduce the barriers associated with their users accessing dbGaP data. Many types of relatively simple data retrieval are made impractical by the structure of dbGaP, FISMA compliance is a major challenge, and the administrative burden of obtaining access to multiple dbGaP studies is prohibitive. Another significant concern that was expressed was that there are no methods for DCC users to achieve signon and authorization that could be used by all of the DCCs. This challenge is addressed in Recommendations 3 and 4.

Asset transport. DCCs saw a need for an interoperable mechanism for transporting datasets from DCC portals to analysis resources, such as Terra or Cavatica. This would support combining dataset cohorts across DCCs, and will also facilitate the use of a range of analysis pipelines. An example workflow from Kids First is: a clinician builds a synthetic cohort on the Kids First portal, and the portal would create a list of the relevant data files. The list would then

be transmitted to an analysis platform where a biomedical data scientist can analyze the data in response to the clinician's needs. Asset transport mechanism already exist, but a significant issue is no single standard is used by the DCCs, greatly impeding data sharing, movement of assets to common analysis systems, and sharing lists of data assets between users. This challenge is addressed in Recommendation 3.

Life stage challenges. DCCs noted various challenges as their center were initially ramping up, interacting with data producers during the active period of the project, and sustaining data and tools after the primary funding is over. These challenges include recruiting engineers, making infrastructure technology choices, creating a data submission, validation, and processing pipeline, providing data in standard formats, developing robust user-facing software, and transitioning data to long-term storage. This results in slow ramp-up, suboptimal infrastructure choices, delays in getting robust data pipelines in place, fragile software, and lost opportunities for data reuse. This challenge is addressed in Recommendations 1, 4 and 5.

Expertise silos. Each DCC has developed tools and strategies to support their mission, and these tools could be used by other DCCs. For example, GTEx has built a mature visualization portal for RNA-Seq data, with tremendous data exploration capabilities that could be used in other DCCs; Kids First has a pragmatic strategy for protected data access that could be readily adopted by new DCCs; and LINCS has a number of powerful analysis tools and approaches that could be reused. There is strong interest in enabling DCCs to reuse existing resources and tools and to learn from each other, in addition to interoperating. A great opportunity for CFDE is to promote cross-DCC interactions, to maintain institutional memory across the DCCs and increase re-use of infrastructure developed by each site. This challenge is addressed in Recommendations 2, 3, and 4.

Increased data analysis and problem complexity. Operationalizing FAIR principles will result in less expert users being able to complete complex, cross-DCC analyses. This increase in researcher access and ability will also dramatically increase both the number and complexity of user support requests at all participating DCCs. In particular, both KF and GTEx are already challenged by the support needs for their datasets. While both DCCs are eager to see more reuse of their data, and KF in particular sees tremendous opportunity for training clinicians to use their portal to find data relevant to patient care, they are wary of bringing on more users without a better on-ramp system and a way to increase the skill level of biomedical scientists using their data. Solutions include obtaining access to inexpensive, large-scale computing, as well as improving interoperability between computational platforms so that users can use familiar tools, and developing helpbot technology to direct users to appropriate documentation resources. This challenge is addressed in Recommendations 3, 4, and 5.

Support burden. All four interviewed DCCs saw a strong need to reduce their support burden. More users and more data reuse will overwhelm their already significant activities in user support, both by creating more support requests and by diversifying the requests they do get as user expertise grows. This challenge is addressed in Recommendations 3, 4 and 5.

Obstacles to cooperation. DCCs have limited time to participate in group exercises, and are wary of spending resources to participate in creating a common set of best practices unless they know *all* of the DCCs will adopt them. Without facilitation by the CFDE, cross-DCC participation is difficult because interactions between the groups are limited, each group has few incentives to adopt the practices of another, for some data and metadata assets no standard exists, and for other assets competing standards may be used by two different DCCs. This challenge is addressed in Recommendations 2, 3, 4, and 5.

Lack of common practices. The acronym "FAIR" is now a popular term frequently appearing in conference presentations, whitepapers, peer-reviewed literature and NIH RFAs. However, there are no specific criteria for DCCs to operationalize FAIRness, or criteria that should be followed in order to determine if data is more FAIR or not. The absence of rigorous FAIRness criteria also greatly reduces each DCCs motivation to participate in the use of common practices. This challenge is addressed in Recommendations 2, 4 and 5.

Insufficient DCC engagement. Technical cooperation between the DCCs and identification of reusable technical solutions is challenging because DCCs are largely unaware of what technical approaches are being used by other DCCs. More generally, as noted by Kids First, running DCCs is a specialized skill shared by only a handful of people. Regular sharing of challenges, approaches, and solutions between DCCs, with the CFDE, and with Common Fund program officers, could have a significant operational impact. This challenge is addressed in Recommendations 2, 3, 4 and 5.

CFDE Resources

The CFDE was formed In April 2019, and has two main components. The engagement team, headed by C. Titus Brown, serves as the point of first contact with DCCs, and provides ongoing engagement between the DCCs, the Common Fund, and the CFDE technical team, as well as public outreach. The engagement team is also providing DCCs with some training and outreach resources. The technical team is a cross-institution collaboration between the teams of Owen White at the University of Maryland, Ian Foster at the University of Chicago, Carl Kesselman from the Information Sciences Institute, Avi Ma'ayan from the Icahn School of Medicine and Susanna-Assunta Sansone from the Oxford e-Research Centre. Additionally, the University of Maryland team provides overall project management of the technical team.

At the inception of the CFDE, the technical team proposed to create a number of tools and interfaces to support cross-DCC search and ecosystem integration. These included a data dashboard, which would monitor DCC data upload to the cloud as well as usage statistics; a unified portal for searching across DCC metadata; a metadata model to facilitate the search portal; and tools to support harmonization, FAIRness assessment, data staging and pipeline sharing. These deliverables are outlined in [Appendix G](#).

Throughout the engagement process, we presented these proposed technical products and solicited DCC input on their usefulness, and feasibility. While several of our technical deliverables are continuing in their original form, others have shifted in scope and focus in response to the needs and concerns expressed by the DCCs. The activities described in Recommendation 3, below, are the result of these changing priorities.

Recommendations for Resource Allocation and Investment by the Common Fund

We have identified potential for investment in 5 categories. Projected costs in these sections are based on reported costs from DCCs where possible. All other costs are estimated using figures from [Appendix H](#). Estimations for salaries do not contain fringe benefit costs, nor do estimations include costs associated with institutional facilities and administrative rates (F&R).

Recommendation 1: Support the current DCCs with targeted investments.

These recommendations focus on investments into individual DCCs that would leverage and expand their existing capabilities.

Data storage and computing. Costs associated with storage and computing will vary significantly for each project. The proportions of storage versus computing will also be different for each group because some centers require more or less computational analysis or quality control improvements to their data. For example, KF, GTEx and HMP currently have no NIH support for storage or computing, and LINCS will have no funding for storage or computing June 2020. It is impossible to know the exact costs, however we can bound the range of costs for data centers that are similar in size to those we have interviewed. DCCs with data obligations similar to that of GTEx (~600gb of data) could incur up to \$250k of cloud-based storage and computing a year. Larger projects use considerably more resources, and cost significantly more: Kids First reports that their current AWS storage bill is about \$70,000 per month (\$840,000 per year). Note that while GTEx, HMP and LINCS are end of life-cycle DCCs, and will have relatively static yearly costs, younger DCCs will require increasing investment each year. Kids First projects their data growth at 200TB a year, which suggests that their cloud storage costs will double in four to five years.

Given their lack of funding, the HMP indicated that moving their data to a stable, professionally managed file system (such as Google or Amazon cloud) was a top priority. They also noted that updating their pipelines to use more modern technology will be key to providing quality datasets to users going forward. Addresses: [Data maintenance and access](#).

Targeted training programs. Several DCCs would like to develop training materials to upskill their user community to make more and better use of their data. There are two types of costs for targeted training: creating/maintaining the materials and the cost of running workshops.

Creating workshop materials is best done in pairs or small groups, as to benefit from multiple experts. We estimate that creating an entirely new, two day workshop would require the equivalent of about three months of full time work for two people, or \$30,000 - \$57,500 for a Bioinformatics Analyst and Bioinformatics Engineer pair.

Running workshops requires supporting event hosting, paying instructor travel and accommodation costs, and the administrative costs associated with registration and event organization. If staff instructors or volunteer instructors are not available, costs could also include temporary instructor pay. Optionally, workshops can include sponsoring some or all of the travel and accommodations for learners. We estimate that a minimal workshop for ~30 learners, with volunteer instructors or onsite staff and no learner sponsorship, would cost \$10,100 - \$19,100 per workshop. We expect a fully sponsored workshop for 30 learners to range from \$57,100 - \$99,300 per workshop.

Webinars typically have approximately the same the administrative overhead of a two day workshop, but require less material generation, and no travel. They primarily serve already expert users, however, and are not effective at stimulating data reuse in our experience. The number and type of people required will vary greatly depending on the topic of the webinar, however we estimate that a 3 experts would spend at about 40 hours each on prep work, and that there would be an additional 40 hours of administrative work, for a range of \$5000-\$9000 per webinar.

Kids First is interested in developing training for several different user communities, including training for clinicians in using their portal and training for biomedical data scientists who are new to working with clinical data. Additional funding for a training-focused hire (1 FTE) would help them develop this capacity and run events.

Likewise, GTEx would like to develop several sets of training materials, including:

- A short intensive workshop on using the Broad Institute's Terra compute analysis platform to analyze RNA-Seq data with the GTEx pipeline. This would enable users to compare their RNA-Seq data with GTEx results.
- A webinar series demonstrating the correct way to use their upcoming eQTL analysis data release.

GTEx training could be addressed with a combination of funding a training-focused hire (1 FTE) and burst funding to support the creation of individual workshop materials. Addresses: User training needs.

End of lifecycle support. The HMP group has expended significant resources towards metadata curation, value addition to their data (such as generation of assemblies, gene catalogs, and annotation of reference bacterial genomes); however, the HMP DCC no longer receives support from the Common Fund. It is unclear what will happen to the HMP resources. This situation is

an example of how the CFDE can serve as a steward for data and tool sustainability. The processing pipelines used by the HMP group are well documented, but these pipelines are based on software that should be updated. We propose to work with the HMP group to modernize their data processing systems, ensure they are documented, and reprocess the data prior to making it available on the CFDE. An estimated 2 FTEs at the HMP DCC would be required to migrate their data to a cloud-based platform using the Common Fund best practices, collect their documented standard operational procedures, transition their analysis pipelines to re-usable Docker containers, and transition the capability of processing their data to an external team. Costs are estimated to be no more than \$260,000 for this effort. Addresses: Ageing infrastructure, Asset specification incompatibilities, and Life stage challenges.

Recommendation 2: Support the current DCCs with cross-DCC investments.

These recommendations focus on investments that would benefit multiple DCCs. Costs for personnel, travel and workshops found in this section are based on estimations outlined in [Appendix H](#). Estimations for salaries do not contain fringe benefit costs, and no estimations include costs associated with institutional facilities and administrative rates (F&R).

Participation in CFDE best practices. We anticipate that each DCC hosting public data will require personnel to participate in CFDE Best Practices in the first year; these FTEs will assist with generation and refinement of the Common Fund Asset Specification, and participate in building Common Fund Asset Manifests described below. These systems serve as the basis to inventory the data assets at each DCC so they will be findable and interoperable. Best practices will also require that DCC staff aid in developing GUIDs, work on training materials, and attend CFDE face-to-face meetings. Oversight for these personnel will be performed in part by the CFDE tech team. This work will likely require 2 Software or Bioinformatics Engineers, or Bioinformatics Analysts averaging no more than \$260,000 per site, per year. Addresses: Lack of common practices, Insufficient DCC engagement and Obstacles to Cooperation.

DCC cross-pollination events. Individual DCCs have significant expertise in complementary areas, and we recommend that the Common Fund support a two day conference to bring DCC personnel together in person to discuss their technological challenges, approaches, and solutions. Kids First and HMP were particularly interested in sharing technical solutions and connecting their teams with other DCCs, with the goal of de-siloing not only their data but their people. In fact, the HMP told us that they have been offering informal advice and mentoring to younger data centers, though not ones on CF's priority list. Annual conferences would serve as an avenue for building cross-DCC collaborations and discussions, and identifying complementary expertise and technologies across the DCCs. The anticipated costs for a 2-day conference for 30 people (20-25 attendees and 5-10 DCC staff) are between \$49,000 and \$84,000. Addresses: Expertise silos, Obstacles to cooperation, and Insufficient DCC engagement.

DCC-to-DCC joint exercise. We suggest Common Fund support joint exercises between DCCs to engage in development, analysis, or training that involves using datasets from each site. We recommend starting with GTEx and Kids First, because their data assets are complementary, and both groups expressed enthusiasm to be more closely aligned during our interviews. A proposal for the GTEx/Kids First joint exercise can be found in [Appendix I](#). Estimated costs for this exercise would be two FTEs at each site to work on data management issues and harmonization, and to deploy analyses to the computational platforms. This exercise would also serve as a pilot project to demonstrate the power of data reuse and illuminate challenges in increasing data interoperability. Additional time to support involvement from PIs will also be needed to assist with coordination and project development. Costs are estimated to total \$897,600. Addresses: [Expertise silos](#).

Infrastructure re-use. The DCCs have a significant amount of infrastructure capability, and have implemented many analytical tools that could be leveraged for re-use by other DCCs. Unfortunately the primary reason DCCs are unable to share resources such as these is due to resource limitations, e.g., the DCCs were not originally funded to make tools or infrastructure into exportable products. We recommend that additional funding in the form of small project grants be offered to incentivize the DCCs to share these tools and resources beyond their project. Awards are optional, and should be allocated based on requests from each site. Activities are likely to require 1-2 Software or Bioinformatics Engineers, or Bioinformatics Analysts averaging no more than \$260,000 per site per year. Addresses: [Expertise silos](#).

Recommendation 3: Support a shift in current CFDE activities to support transformative activities.

These recommendations focus on activities performed by the CFDE tech team and will be completed by leveraging our current set of deliverables, outlined in [Appendix G](#). These efforts require no new funding and will involve relatively minor course corrections using our current set of deliverables. These activities will need to continue after December 2019, and funds should continue to be reserved to support them at or above our current level.

Communication within Common Fund program. A number of steps can be taken to accelerate operationalization of CFDE. First, we strongly recommend distribution of this report to the DCC principal investigators (PIs), and that Common Fund leadership meet with the PIs to review and discuss the contents of the report with this group. We recommend this be conducted in combination with the CFDE engagement team (to review the details of the project), the Common Fund Program Officers, and other members of Common Fund leadership. Addresses: [Insufficient DCC engagement](#)

Engagement with Common Fund. The CFDE group recommends meeting with senior leadership of Common Fund for an extended (~3 hour) discussion in order to better understand their priorities, their goals for the coming year, and how the CFDE can assist with those goals. This

will better enable us to create strategies for Data access barriers, Data maintenance and access, Ageing infrastructure, Expertise silos, and Obstacles to cooperation.

Common Fund Data Asset Specification. Each of the DCCs host files (e.g., genomic sequence, metagenomic, RNA-Seq, physiological and metabolic data) and we can greatly simplify discovery of these assets by creating a specification for a set of descriptors for each of these files. The specification will contain a small number elements such as:

- Global Unique Identifier (GUID)
- Originating institution (e.g., "Broad Institute")
- File type (e.g. "RNA-Seq", "GWAS")
- Tissue source and species name for the sample

The data asset specification will be encoded in a computer readable format, and will enable us to use readily available internet technologies to get additional information for each asset, such as the metadata (e.g. patient variables, project name), and to resolve access issues such as files being hosted on the cloud or local servers. While many implementations for electronically encoded data assets of biomedical resources have been proposed in the literature, no single standard has been adopted by the Common Fund DCC community. However, there is a high likelihood of achieving adoption across several groups if a consensus-building process is carefully managed by the NIH and the CFDE team. Addresses: Asset specification incompatibilities

Common Fund Data Asset Manifest. The ability to bundle a list of CFDE data assets into a machine-readable file will greatly facilitate finding datasets among DCCs, and effectively transporting these datasets to resources such as cloud-based analysis tools. We refer to the list of Common Fund assets as *manifests*. The DCCs will generate manifests for all of their data assets at their center, enabling both comprehensive inventories for all of their files and the use of that information to find and access all of their data in a CFDE portal. Manifests are similar in function to users collecting a shopping list on a commercial web site, and manifests for subsets of data located at multiple Common Fund DCCs will be used to transport files to analysis resources, such as analysis pipelines hosted at Terra or Cavatica. While a standard for manifests has not been developed or adopted by the broader community, the CFDE project represents an excellent opportunity to drive creation of a standard for all of the Common Fund DCCs. Addresses: Asset transport.

Use of FAIRness as an organizational tool. An incentive to motivate the DCCs to converge on inter-DCC compatible standards to represent their data will be to how DCC assets are hosted on the internet (e.g., CFDE Best Practices), and by verifying that each DCC is participating in these best practices by measuring their compliance with specific, concrete FAIRness metrics.

These activities will serve to break through the dilemma presented to each DCC who want to participate in CFDE but would only do so if *all* of the DCCs also participate in a common set of best practices. Addresses: Obstacles to Cooperation.

CFDE Portal. Construction of our portal is underway. Demonstrations for usage of the site will occur around October. Data collection is still preliminary, but depending on the success of the DCCs generating manifests of their data assets, it is likely we will demonstrate query capability across most of the DCCs hosting data. Usage of assets combined between DCCs, as well as passing those assets to computational platforms, will be possible prior to the end of the year. Address issues associated with finding Common Fund data assets that were described in Asset specification incompatibilities.

Computational analysis platforms. Our team will investigate potential computational platforms to perform analyses of CFDE data in the next few months. We will review the total list of platforms with Common Fund for approval and recommend these sites should be considered include: Terra, Cavatica, and other stack providers from the Data Commons Pilot Project (e.g., 7 Bridges and others). Coordination with these providers will involve testing utilization of the Common Fund Data Asset Manifest system, and developing pipelines that can be deployed on their platform. The analysis platforms will be used for the DCC joint exercises described in the Cost Estimation section, as well as training. Addresses: Increased data analysis and problem complexity.

Continued engagement. One result of this engagement was to reinforce how essential a role site visits play in building an effective working relationship with the DCC staff. These visits were very effective for establishing trust, understanding the goals of each DCC, exploring incentives for the DCCs to participate with CFDE, and discovering important resources developed by the DCCs that could be utilized by other groups. These activities will continue to gather additional information from the remaining DCCs. Sustained, high-level engagement of all DCCs may require an increase in post-December CFDE funding. Addresses: Insufficient DCC engagement

User training. The realization that multiple DCCs were tremendously interested in training programs was an unexpected result of our close engagement with them. Since training is closely linked to *actual* data reuse, and the best (and perhaps only) way to evaluate actual interoperability is to demonstrate it, we see training as one key to operationalizing FAIRness. Training can also reduce the substantial user support burden common to all of the DCCs we interviewed.

There are a variety of training modalities already being used by the DCCs we interviewed. These include webinars on data analysis and portal use (LINCS and GTEx), MOOCs (LINCS), and in-person workshops (HMP). Despite intense interest in continuing these activities, relatively little training is currently happening, and most training materials are at least somewhat out of date.

CFDE can play a catalytic role in data reuse by bringing expertise in designing user engagement and training to DCCs, helping create training materials for new users, and supporting evaluation and assessment activities. CFDE can provide support for training by supporting existing training activities such as webinars, helping develop new materials to support training within and across DCCs, and working with DCCs to pilot new workshops for their user communities. CFDE also has significant expertise in evaluation and assessment of training activities, as well as pedagogical techniques and strategies for user-focused training and engagement. Depending on the level of support DCCs require for development of training programs, this effort may require an increase in post-December CFDE funding

Based on the results from our DCC engagement, CFDE plans to engage in the following pilot training efforts through the remainder of this year:

- Kids First is intensely interested in developing a hands-on workshop series to train clinicians to use their portal. We will work with KF to outline, write, and pilot workshops focused on clinicians.
- CFDE will work with GTEx to pilot a webinar series on eQTL use and interpretation. These webinars will be recorded and made available through the GTEx site for users.
- We will also work with GTEx and the Terra platform to build and pilot a curriculum for a two-day hands-on data analysis workshop that teaches users how to use the GTEx analysis pipelines on the Terra platform for their own data, so that they can reuse GTEx data for their own analyses.
- We will work with KF to build and pilot a similar two-day curriculum for the Cavatica data analysis platform.

The DCCs are strongly motivated to participate in these training activities because they see it as serving their user community and reducing their support burden. Incentives for user communities to attend these trainings will be integrated into the materials as we develop them. Addresses: [User Training Needs](#), [Support burden](#), [Increased data analysis and problem complexity](#)

Recommendation 4: Invest in new transformative activities by the current CFDE Team

We encourage Common Fund to consider reserving funds for activities described in this section. The activities could be achieved through a combination of repurposing existing deliverables by CFDE technical team, and/or through addition of new funds. Cost estimates are not provided here but will be generated upon request from the NIH Common Fund.

User Helpbot: a first point of contact help desk for all CFDE users. Staff at the DCCs have generated a considerable amount of documentation for their users, but unfortunately users rarely read these materials prior to contacting the DCC for additional help. This observation led to the realization that creation of CFDE is likely to *increase* the help desk burden of all DCCs. In

addition to creating new, potentially confusing, ways for users to combine data, CFDE will draw additional users to all sites, many of whom lack bioinformatics expertise. It would be of significant positive impact to create a common front-end “helpbot” service that would be available at all DCC web sites created for CFDE. This service could potentially use AI approaches to first filter user questions, to see if questions could be answered from FAQ content, and to handle general bioinformatic questions that are not related to any particular DCC. This “user helpbot” could significantly lower the support burden for each DCC, and unify the bioinformatic educational information supplied to CFDE users. Addresses: Increased data analysis and problem complexity and Support burden

A CFDE search plugin. It will be straightforward for CFDE to create a web based plugin to be used by all of the DCCs, to assist with accessing data between sites. The advantage to creating a common plugin for all of the sites is that it would reduce costs associated with multiple DCCs creating interfaces that perform the same function, and would simplify the user's experience by creating a single search tool. By having a single development team create the plugin, it would also mean we could rapidly respond to changes to the underlying implementation of the minimal metadata standards and CFDE manifest system. An interface component that is used at the websites of all the DCCs will provide a common “look and feel” of the websites across the CFDE. The search plugin address issues associated with finding Common Fund data assets described in Asset specification incompatibilities and reduces unnecessary re-development of similar technologies of the DCCs described in Expertise silos.

Improved sign on and authorization. Among the most impactful, disruptive, capabilities that could be achieved by CFDE would be to provide DCC users with a single sign on step to access protected data across the Common Fund, and to enable researchers to query on those data using a user-friendly interface. [Appendix J](#) reviews many important considerations relevant to protected data access, some considerations are briefly discussed here. Authentication refers to an electronic means to verify a user's identity. Authentication is similar in function to using a passport when entering a country. Single sign-on (SSO) is similar to the using *the same* passport to enter different countries. In the case of the CFDE, it would be desirable to have an authentication system whereby a user logs in with a single username and password, and is granted access to multiple data from multiple DCCs. Many authentication and SSO systems exist but a single system that works across all of the CF DCCs has yet to be adopted, it is not clear which system should be adapted to the CFDE portal. Whichever system is selected, it will also be valuable use this common SSO and auth system to allow users to pass data assets to computational analysis platforms such as Terra and Cavatica. The CFDE technical team can review the Kids First SSO and auth strategy to see if that can be applied more broadly to work for the CFDE. Alternatively, members of the CFDE technical team have developed effective production system that is used by a number of hospitals, research institutions, government agencies, and NIH centers -- this system could be readily applied to the CFDE. Addresses: Data access barriers.

Reduce costs for FISMA compliance. The Federal Information Security Management Act (FISMA) defines a set of controls to protect computer information and operations from security breaches. Requirements include maintaining an inventory of IT systems, categorizing data systems by risk level, maintaining a security plan, utilizing security controls, conducting risk assessments, obtaining certification, and conducting continuous monitoring. These activities often incur more than \$100,000 in costs to institutions obtaining FISMA compliance. However, the benefit of obtaining FISMA compliance is that each DCC would increase its ability to host human protected data, obtain trusted partner status for managing data (even just metadata) that are currently hosted at dbGaP, and to share data with other FISMA-compliant DCCs. This would enable each DCC to provide a much wider range of important services to its users. The CFDE tech team is engaged in evaluation of FISMA compliance under our current set of deliverables; however, we recommend that Common Fund examine more ambitious mechanisms to enable the CFDE tech team help reduce the cost burden of DCCs seeking to become FISMA compliant. Addresses: Data access barriers.

Develop CFDE Best Practices. In order to achieve its goals CFDE will encourage the DCCs to participate in adopting a series of best practices in order to operationalize FAIRness and promote interoperation between datasets. In the first year, the best practices will require implementation of the Common Fund data asset specification and the Common Fund data asset manifests at each DCC. Other best practices will be developed over time in close collaboration with the DCCs, and disseminated to all groups. Future best practices will include recommendations for single sign on, authorization methods, FISMA compliance and other important implementation elements of CFDE. Addresses: Lack of common practices and Obstacles to cooperation.

Lifecycle support. CFDE has the potential to provide comprehensive sustainability of Common Fund data that addresses two needs: helping newly formed DCCs to join CFDE, and participating in the best FAIRness practices for CFDE to thrive. These best practices (e.g., Common Fund data asset specification) will also enable findability, accessibility and reusability long after DCCs are decommissioned, provided that information continues to be hosted on cloud systems. We will also produce software (such as the CFDE search plugin or helpbot) that will reduce costs of individual DCCs implementing similar systems. In addition to this, CFDE will develop a robust end of life cycle program to manage continued stewardship of data as DCCs are deactivated. One example project for end of lifecycle support was listed in the 'Recommendation 1: Support the current DCCs section with targeted investments'. Similar projects should be considered in future years. Addresses: Life stage challenges.

Incentivize participation. Once the rules for engagement are clearly operationalized by CFDE Best Practices and enforced through open FAIRness evaluation, Common Fund has several options to incentivize the DCCs to abide by CFDE Best Practices. This can be achieved by supplying resources to each center to assist in the development of improved data asset specifications, granting access to services (such as cloud-based workspaces) that would rely on the use of CFDE Best Practices, and reducing storage costs associated with CFDE compliant

data. Compliance with CFDE Best Practices, can also be linked to cost reductions associated with the STRIDES program. Offering RFAs to external to groups that use CFDE Best Practices should also be considered, as well as providing short to medium term grants for DCC-to-DCC projects like increasing the harmonization level of data between sites and demonstrating cross-dataset usability. Addresses: [Obstacles to Cooperation](#).

[Engagement to spread best practices](#). Ideally, development, maintenance and dissemination of CFDE Best Practices will be done in close consultation with the DCCs, while still being managed by an independently operating engagement team to ensure no single DCC dominates development of CFDE Best Practices. The CFDE engagement team currently serves this role, and can continue to regularly engage the DCCs, promote DCC-to-DCC interactions, develop documentation, participate in training, and perform joint exercises across the DCCs to test the effectiveness of interoperation. The engagement team will also disseminate all documented best practices to the DCC over time. Addresses: [Obstacles to Cooperation](#), [Expertise silos](#), and [Insufficient DCC engagement](#).

Recommendation 5: Invest in long-term ecosystem support with targeted RFAs.

CFDE will rapidly grow into a Common Fund resource that will accelerate new biomedical discoveries by providing a cloud-based platform where investigators can store, share, access and compute on biomedical resources. Several key funding opportunities that leverage CFDE should be considered.

[Training opportunities](#). The training needs for the CFDE will grow in concert with the CFDE resources and usage, and training can also spur adoption of CFDE resources by new users. Training needs include basic training for cloud computing, technical execution of workflows on the Terra and Cavatica platforms, building bioinformatics competency among users of the Common Fund resources, and clinician-focused training programs that integrate with specific CF resources. We recommend that the Common Fund consider issuing RFAs to develop open educational resources and curricula and deliver workshops to target user populations. We also suggest that the Common Fund invest in a “data use” postdoc program that would support deep cross-DCC data use by biomedical researchers funded specifically for this purpose. Depending on the size and scope of the RFAs, a training coordination center may also be important for connecting and facilitating training activities. Addresses: [Increased data analysis and problem complexity](#), [User Training needs](#), and [Support Burden](#).

[CFDE sharing of best practices and sharing](#). The CFDE will need a coordination center, to maintain institutional memory across multiple DCC lifecycles. This coordination center would maintain engagement activities, coordinate and iterate on standards and implementations, maintain beginning and end lifecycle documentation, coordinate events, and track common practice as it emerges. Addresses: [Obstacles to Cooperation](#), [Insufficient DCC engagement](#), [Life stage challenges](#), and [Lack of common practices](#).

Analytical development. CFDE will adopt common systems to find biomedical resources (such as datasets from multiple cohorts) and then easily transfer lists of those datasets to other resources. We recommend the Common Fund consider issuing RFAs that would incentivise analytical tool developers to create and adapt analytical tool to take advantage of CFDE environment, and to link these tools to cloud based computing environments. Addresses: Increased data analysis and problem complexity.

Cloud workspace support. Several groups are well-positioned to offer computational workspaces to enable CFDE users to perform analysis. These infrastructures could be rapidly adapted to CFDE data standards, and specially tailored to support the types of data hosted by the Common Fund DCCs. The cloud workspace providers could develop novel query tools, computational pipelines, and create social-network based sharing systems to support consortiums. Addresses: Increased data analysis and problem complexity and Data access barriers.

Risks and Challenges

Several key concerns regarding the formation and success of CFDE have been noted throughout this document. These concerns are restated here and listed in relative order of priority:

dbGaP access. While CFDE could easily result in significant advances in querying and accessing datasets hosted at each DCC, this information will be far more useful if can be used in combination with data hosted at dbGaP. There is a critical need to reduce the barriers associated with accessing dbGaP data, and to enable access to data *across* studies hosted there.

Single sign-on and authorization. Two of the centers most intimately familiar with protection of human data (GTEx and Kids First) are skeptical that the solutions provided by the NCBI Virtual Directory Service will meet their needs. This solution is also expected to take another 18 months for implementation. In the absence of a single-sign on solution it is unlikely that CFDE will be able to adopt a system that will readily apply to all Common Fund DCCs; this will impede our ability to share restricted data.

Building expertise within and across the DCCs. There is a large potential loss of opportunities if the DCCs continue to operate in isolation of each other. We strongly recommend increasing their interactions with each other to avoid duplication of effort, capturing institutional knowledge from the mature DCCs, promoting the re-use of technologies, and encouraging shared standards development.

Support burden. All DCCs reported that user support is vital to their mission, and requires a lot of investment of their time. It is likely CFDE activities will increase the number of users seeking help as we enable the user community to make use of datasets across all of the DCCs. This increased support burden can be addressed by creating a centralized help desk, and offering additional training in the form of documentation, webinars, and conferences.

STRIDES and CFDE role clarification. Confusion on the part of the DCC staff about the difference between STRIDES and CFDE was evident during our visits. It is likely the lack of clarity is shared by other sites, and this could potentially reduce the interest of the DCCs, as well as their NIH Program Officers, in participating in CFDE. Another concern is that storage costs are significant; regardless of the funding source, or the reductions made available through STRIDES, costs to host data on cloud-based systems are going to be considerable in size.

Ensuring CFDE compliance. While we have proposed several measures to promote usage of CFDE Best Practices, the concern remains that DCCs will not be motivated to participate in the standards for use across the DCCs. Ultimately DCCs are independent entities, and answer to their own NIH Program Officer; encouragement from Common Fund leadership for DCCs to participate in CFDE will be vital to the success of this project.

Acknowledgements

We are grateful to Brian Osbourne and Brian O'Connor for their assistance in preparation of this document. Many members of the CFDE tech team participated in data gathering; in particular we would like to cite Lee Liming for his review of the report and assistance with data collection, as well as Theresa Hodges for her assistance in data collection.

Appendices

Appendix A - Site Visit And Survey Methodologies

DCC SITE VISITS.

Two separate meetings were held with GTEx and Kids First, each meeting was for a day and a half. Prior to the meetings we presented an hour long presentation by teleconference to give an overview of the CFDE, answer questions, and establish the goals of the visit. An agenda was exchanged with DCC personnel, who assisted with arriving at the final material to cover at the meeting. During the meeting, the agenda was used as an informal guide for structuring the day. Four CFDE personnel which always included Amanda Charbonneau (UCD), Brian O'Connor (Bionimbus), Titus Brown (UCD), and Owen White (UMB). Primary representatives from each DCC included the PIs and their senior technical staff. In most cases were able to meet with the entire DCC staff during the visit.

Discussions were initiated with short introductions from the engagement team and attending DCC members. We reviewed that the goal for the engagement team was to collect information about the DCC, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for the DCC group. The DCC group was then asked to provide an overview of their operation, with topics including: mission, vision, goals, stakeholders, and challenges. Most of the CFDE members took separate notes, all of which were recorded and stored as google documents. Notes were then reviewed and collated into the reports appearing in the appendices in each report. Another goal of the site visits was to establish how relevant the activities planned by the CFDE were to each DCC. This was achieved by hearing about the general challenges faced by the DCC in their presentations, informal conversations and a final summary discussion held on the second day. Summaries from both site visits were reviewed with NIH staff members within one week of each visit.

TELECONFERENCE INTERVIEWS WITH LINCS AND HMP/iHMP

Separate 3 hour teleconferences were held with LINCS and the HMP/iHMP DCCs. Agendas for each meeting were circulated prior to each call. Teleconferences were initiated with short introductions from engagement team members and attending DCC members. Goals for the meeting were reviewed and focussed on the DCC team values for things like: mission, vision, goals, stakeholders, and challenges. Reviews of types and formats maintained, tools and resources owned by the DCC were also performed. Prior to each meeting participants were asked to sort their goals for the DCC using an online resource known as FunRetro which enables users to create comments in responses to questions, and to prioritize those comments.

Each DCC was requested to provide a short 20 minute overview presentation, and team members were asked to cover topics such as: the vision for your organization, problems it is solving, goals for the next year, issues that are taking up their bulk of their time, challenges blocking implementation, and a number of user engagement questions. The remaining time for the teleconference calls were then dedicated to review the results of the goals assessment to ensure the engagement team accurately reflects the DCCs answers, motivations and goals.

PASSIVE EVALUATION OF EACH DCC

Members from three separate CFDE technical teams performed independent evaluation of each DCC resource. The information presented in Tables 1-3 and Figure 1 was initially gathered by our team by reviewing the NIH Common Fund Programs website (<https://commonfund.nih.gov/initiativeslist>), checking data in NIH RePORTER, and visiting the current websites and data portals for each of the DCCs. (We registered for access to data portals when necessary to gain access to details.) We confirmed and added to technical pieces (e.g., data types, use of cloud resources and cloud expertise) during our interviews with four of the DCCs.

OVERALL FACT CHECKING

Information from all tables, appendices and this report have been reviewed by CFDE technical team members from at least two different institutions. Staff from NIH have also participated in review of material appearing in all report documents. Reports for the two DCC site visits have been vetted for accuracy by the DCC staff from those institutions. FAIRness measures were collected and summarized by CFDE technical team members from two different institutions.

Staff from the DCCs have NOT reviewed information collected by passive evaluation.

Appendix B - Descriptive Overview Of All Dccs

	General description	Cohorts / supported data sources	Degree of Operationalization	Bulk file content	Dataset types	Dataset search facets	Protected data hosted at dbGaP?	Cloud expertise / usage
Kids First	Integrates genomic and clinical data from different disease types to accelerate discovery using cloud-based analyses systems.	Approximately 20 data contributors generating whole genome sequence data and, in some cases, whole exome and transcriptome data for tumors or affected tissue to discover genetic variants that contribute to pediatric conditions. Utilizing a variety of approaches may be taken for sequencing of tumors or affected tissue, such as 30X whole genome sequencing, combined with 100X whole exome and 100X RNA sequencing.	18 months. However significant volume of data and infrastructure has been generated during this time thanks in part to additional funding from independent sources.	11 studies, each with 250 - 2000+ subjects. 927,11B	BAM, CRAM, fastq, VCF, clinical measurements	Study, diagnosis, clinical phenotypes, patient events, pedigree, gender, race, tissue type, data type	Yes	Fully deployed on cloud-based system
GTEX	The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific community a resource with which to study human gene expression and regulation and its relationship to genetic variation.	V7 includes ~1000 post mortem donors, 53 tissue sites, 11 distinct brain regions, and 2 cell lines.	Established in 2010. Globally used resource with roughly 50% of their users from the US or the UK, with the rest being spread throughout the world. Portal receives roughly fifteen thousand users per month.	53 Tissues, 960 donors 50,000 samples, data set size expected increased in a total of 600GB by next release	De-identified annotations, RNA-Seq data, Single-Tissue datasets, Multi-Tissue eQTL Data, Reference Files, and Single-Cell Data.	Tissue type, individual gene expression level, expression comparison across tissues, comparisons based on eQTLs, histological images	Yes	Fully deployed on cloud-based system
HMP / IHMP	Characterization of the human microbiota to further our understanding of how the microbiome impacts human health and disease.	Healthy Cohort of 300 healthy individuals, each sampled at 5 major body sites (oral, airways, skin, gut, vagina) at up to three timepoints. Each body site consisted of a number of body subsites, for a total of 15 to 18 samples per individual per timepoint. Disease Cohorts: 18 projects with one or more cohorts aimed at studying specific health conditions.	Completing 10 years of operation, funding discontinued.	21 studies, 48 primary body sites, >32,000 samples, >118,000 files 9.75 TB (IHMP), 7.22 TB (HMP)	Reference microbial genomes, Whole metagenomic sequences, 16S metagenomic sequence	Unique subject ID, Body site, Sex, Studies, Visit number	Yes	Local servers for data served from the website. Have performed demonstrations w/ orkshops with cloud-based data.
LINCS	The Library of Integrated Network-Based Cellular Signatures (LINCS) Program aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents.	Studying effects of perturbations on gene expression in ~1,175 cell lines, 60 primary cells, 32 iPSCs, 31 differentiated cells, and 2 embryonic stem cells.	The initial phase ended in FY 2013. Phase II has been funded since 2014. Funding will end on June 30, 2020.	396 datasets, 100TB genomic data, >1PB imaging data	Binding, imaging, transcriptomics, proteomics, epigenomics	Method, Subject area, Center, Assay, Process, Project	Most data is not protected. Some information may be in dbGaP.	Not as yet, effort towards using cloud based systems underway.
4D Nucleome	The 4D Nucleome (4DN) project aims to understand principles underlying nuclear organization in space (three dimensions) and time (the fourth dimension), the role nuclear organization plays in gene expression and cellular function, and how changes in nuclear organization affect normal development as well as various diseases.	24 different labs are helping to generate quantitative models of nuclear organization in human and mouse genomes in diverse cell types and conditions, including in single cells. Five cell lines have been designated as Tier 1, which will be a primary focus of 4DN research and integrated analysis. 12 other lines that are expected to be used by multiple labs and have approved SOPs for maintaining them have been designated Tier 2.	Stage 1 has been operating since 2015. Stage 2 is expedited to launch in 2020.	730 experiment sets, 2107 experiments, 6823 files, 28.99 TB, 166 external datasets	DNA FISH, RNA FISH, SPT, 2-stage Repli-seq, in situ Hi-C, single cell Hi-C, RNA-seq, ChIP-seq, David-seq, ATAC-seq, NAD-seq, ChIA-PET, DNA SPRITE, PLAC-seq, MARqI, RNA-DNA SPRITE, Micro-C, TSA-Seq	Organism, Experiment type, Biosource type, Lab, Center, Treatments, Assay details, Comments	No	Fully deployed on cloud-based system
Metabolomics	The Common Fund's Metabolomics Program serves as long standing national public repository for metabolomic data. Users are enabled to analyze and interpret metabolomics data, including the ability to determine metabolite identities. This project also has developed best practices and guidelines to promote accuracy, reproducibility, and re-analysis of metabolomics data.	The Data Repository and Coordinating Center (DRCC) accepts metabolomics data for small and large studies on cells, tissues and organisms via the Metabolomics Workbench. It accommodates a variety of metabolite analyses, including, but not limited to MS and NMR. Processed data (measurements) may be in the form of quantitated metabolite concentrations, MS peak height/area values, LC retention times, NMR binned areas, etc. Raw data in the form of MS and NMR binary files and associated parameter files may also be uploaded. Data from both targeted and untargeted studies are accepted.	Launched in 2012. Approved for a second stage of support from FY18-2015.	920 publicly available studies, 6.4TB (compressed) 210 files, 233 studies with restricted access	Raw/unprocessed NMR data, MS data, Processed data (general)	Project (Study groupings), Study, Sample source (site on body), Species, Disease (from study) Human Pathways (metabolic process) Metabolite class, PUBCHEM CID, Name (Common, Systematic), Formula, Exact mass, Tolerance (daltons), LIPID MAPS ID, KEGG ID, InChIkey, Gene Name, Gene Symbol, Synonyms, Alternate names, HMDB Pathway, Reactome Pathway	No	Cloud-hosted data hosted at the San Diego Supercomputer Center.

	General description	Cohorts / supported data sources	Degree of Operationalization	Bulk file content	Dataset types	Dataset search facets	Protected data hosted at dbGap?	Cloud expertise / usage
MoTPAC	The Molecular Transducers of Physical Activity in Humans program aims to extensively catalogue the biological molecules affected by physical activity in people. Identify some of the key molecules that underlie the systemic effects of physical activity, and characterize the function of these key molecules.	Artrial studies involved acute and exercise training of 6-month and 18-month rats. Eighteen tissues were collected at 7 post-acute exercise time points for acute cohort. For rats that underwent a training regiment ranging from one to eight weeks, tissues were harvested 48 hours after last exercise bout. Human studies will involve a multi-center clinical cohort of approximately 2,700 healthy human volunteers (male/females), 10-80 years of age, all fitness levels. They will collect blood, muscle, and fat samples from active and sedentary volunteers who will perform resistance or aerobic exercises.	Awards issued in Sept. 2016. Six-year program, through 2022. Pre-clinical animal studies finished in May 2019. Recruiting for human clinical participants is expected to begin in Fall 2019.	No data available yet.	Genomic, transcriptomic, epigenomic, metabolomics and proteomics data are expected.			
SPARC	The Stimulating Peripheral Activity to Relieve Conditions (SPARC) program aims to transform our understanding of nerve-organ interactions and ultimately advance the neuromodulation field toward precise treatment of diseases and conditions for which conventional therapies fall short.	SPARC projects are expected to use a multi-expertise approach to comprehensively understand the neuroanatomy and neurobiology of both afferent and efferent innervation of a major organ, as well as characterize nervous system regulation of function of that organ. The targets of the projects encompass 10 major organs and 7 other tissues/organs, such as adipose tissue, bone marrow, esophagus etc. Data generated by these projects will be provided to the SPARC data coordination center, which in turn will generate detailed functional and anatomical neural circuit maps for major organs and their functionally-associated structures.	Program was launched in FY 2015. Funding for the Data Coordination, Mapping, and Modeling Center started in September 2017 and will end in August 2022.	No data available yet. Data portal expected to be launched in Summer 2019.	Imaging and omics data, such as proteomics and transcriptomics are expected. Functional and anatomical neural circuit maps will be also be generated.			Blacklyn, Inc. has been awarded funding for five-years to develop a cloud-based scientific data management platform tailored to the needs of SPARC investigators.
HUBMAP	The Human BioMolecular Atlas Project (HUBMAP) aims to catalyze development of an open, global framework for comprehensively mapping the human body at the level of individual cells.	The HUBMAP Tissue Mapping Centers (TMCs) will collect and analyze a broad range of largely normal tissues, representing both sexes, different ethnicities and a variety of ages across the adult lifespan. These tissues include: 1) discrete complex organs (kidney, ureter, bladder, lung, breast, colon); 2) distributed organ systems (vasculature), and 3) systems comprised of dynamic or mobile cell types with distinct microenvironments (lymphatic organs: spleen, thymus, and lymph nodes).	Launched in November 2018 and entered a scale-up phase in Fiscal Year (FY 2019) that will continue through FY 2021. A production phase will run during FY 2022–FY 2024, and a transition phase will occur in FY 2025.	Data portal is currently being developed. No public data until 2020.	Imaging data, single cell omics datasets (scRNAseq, scATACseq, scDroptseq, SNAREseq, scTISseq), and MS-based proteomic, lipidomic, and metabolomic datasets are expected.			Cloud based data hosted at Pittsburgh Supercomputing Center (PSC) and the University of Pittsburgh.

Appendix C - GTEx Site Visit

Broad Institute – 415 Main Street Cambridge MA 02142

Monday June 3, 2019

Room Location: 75A-11-Teton (11081)

Meeting Logistics

We held a meeting with GTEx group at the Broad Institute on Monday June 3, 2019 for a day and a half. GTEx collaborated with us before the meeting to build the agenda at the end of this document, which we used as an informal guide for structuring the day. Representatives in attendance from the CFDE were: Anup Mahurkar (UMB), Amanda Charbonneau (UCD), Brian O'Connor (Bionimbus), Titus Brown (UCD), and Owen White (UMB). Primary representatives from GTEx were Kristin Ardlie (PI), Jared Nedzel (lead software developer) and Francois Aguet (lead computational biologist). Several sessions were held where we were able to meet with the entire GTEx staff of eight people.

Discussions were initiated with short introductions from the engagement team and attending DCC members. We reviewed that the goal for the engagement team was to collect information about GTEx, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for the GTEx group. The GTEx group was then asked to provide an overview of the GTEx operation, with topics including: mission, vision, goals, stakeholders, and challenges.

Meeting Logistics

GTEx Overview

Harmonized data

Cross cutting metadata models

Self-governed metadata standards

FAIR assessment

Findability and Accessibility

Interoperability and Reusability

Authentication/Authorization

[Data Dashboards](#)

[CF Data Portal](#)

[Data Platform](#)

[Data Hosting](#)

[Data Analysis](#)

[Training](#)

[Outcomes](#)

[CFDE Targets](#)

[Major Use Cases:](#)

[Game Changers](#)

[Agenda](#)

GTEx Overview

GTEx is a data resource and tissue bank to study the relationship between genetic variation and gene expression in multiple human tissues. It is a valuable tool for exploring the genetic basis of complex human diseases. GTEx is also examining sex-based, and cell-based differences in how genes are turned on and off and how they are regulated. Samples were collected from deceased adult donors, with multiple tissue samples collected per donor (e.g. lung, brain, pancreas, skin, etc.). The GTEx project differs from most other Common Fund DCCs in that instead of being dedicated to disease, they are striving to build a 'reference normal' database. While all of their tissue is necessarily from deceased donors, each sample is screened by pathologists, and tissues showing signs of specific diseases are excluded from their pipeline. As such, the GTEx data collection is our best current resource for representing the typical range of human gene expression.

GTEx started in 2010 and was funded for seven years. During that time, over 650 papers were published using data from the GTEx data resource. The portal that provides access to GTEx resources was launched in 2013, and is still active. Although the project is officially over, the portal still has about 15,000 visitors a month, and is currently funded with a genomic resources grant. The current release of GTEx, V7 includes: ~1000 post mortem donors, 53 tissue sites, 11 distinct brain regions, and 2 cell lines. Due to their rigorous tissue screening process, the number of samples for each tissue varies greatly. For example, reproductive, bladder, and some brain tissues are under-represented in the dataset, which is unsurprising given the high mean age of their typical donor. Data types hosted by their group include: summaries of gene expression based on RNA-Seq data, quantitative trait loci data, and histology images. Their

current dataset size is roughly 395 TeraBytes (TB), and will soon increase to a total of 600TB. GTEx also has a large number of datasets that have protected status and are stored at dbGaP. This information includes raw RNA-Seq files, whole genome shotgun sequence, genetic variants, and donor phenotypes.

Harmonized data

GTEx has already harmonized with ENCODE and will be closely harmonized with MoTrPAC; their RNA-Seq pipeline is also implemented and used by TOPMed. These collaborations were primarily due to opportune circumstance: in addition to her work at GTEx, Kristin Ardlie sits on MoTrPAC advisory board, and collaborated with three MoTrPAC PIs who are experienced with the GTEx datasets (Mike Snyder and Steven Montgomery from Stanford, and Tuuli Lappalainen from the New York Genome Center). Both Kristin and Francois are also engaged in implementing best practices for RNA sequencing pipelines for TOPMed.

Cross cutting metadata models

GTEx expressed concern that broad metadata harmonization may be practically unfeasible for more than a small number (likely fewer than 10) of core terms per data type, and that it may be of limited value. Instead, they favor creating a data definition language for metadata such that a user with sufficient knowledge of the ontology of two datasets can use the same API to navigate both and build correspondences as needed. The ontology could have a small number of core terms that are always applicable, as well as a more articulated set that cannot always or often be mapped between datasets. For instance, RNA-Seq datasets might always be required to have terms for metadata such as sample collection dates, tissue type, lane or other batch sequencing batch effects, and other data that would need to be incorporated into any model that was used to analyze RNA-Seq data taken from many experiments. In their view, an overall cross-cutting data model is not practical, or even particularly desirable: “The important thing is for the metadata to be clear and defined, so it can be clear if you’re comparing apples to apples or not.”

One potential solution to this is to introduce ‘metadata levels’, where a particular dataset with little or no metadata might be a ‘level zero’, one with the core metadata a ‘level one’, one with some level of harmonization is a ‘level three’, and so on. GTEx pointed out that a core problem with this approach is that once the ‘levels’ are linked to the amount of harmonization, the problem of assigning a level becomes extremely complex, because there is no single, defined ‘center’ of metadata to harmonize to. This means that a given dataset, A, might be 80% harmonized with a second dataset, B. However dataset A might not share any terms with a third dataset, C, even though 50% of the terms in datasets C and B are harmonized. The larger the

number of datasets, the more impossible it will be to assign a single harmonization level to any given dataset. Instead, harmonization level would need to be described in terms of a given pair of datasets, and any given dataset would be very unlikely to be at the same level of harmonization with every other. This also has implications for FAIRness metrics, as assigning a dataset a degree of 'Interoperability' will have the same challenges.

Self-governed metadata standards

GTEEx would be interested in contributing to defining a metadata standard, assuming that standard is based on something like the ontology described in the Harmonized Data section. That is, a method that encodes each ontology into a data definition language. In this way, two *different* ontologies can still be in the *same* common schema, and users that have knowledge of the ontology can use the same API to access that information. GTEEx is not interested in participating in an attempt to broadly harmonize across every Data Commons dataset. This would take a substantial amount of resources, and they are skeptical that it can be achieved in any practical way.

FAIR assessment

Findability and Accessibility

GTEEx made a clear distinction between the different parts of FAIR and how, and whether, they are useful. Findability and Accessibility are important and actionable. As is further discussed in the Data Platform section, the raw data from GTEEx is currently not readily accessible by users. As part of the DCPPEC, GTEEx moved all of their raw data into the cloud, however, when the program ended, the system for providing access to that data ended with it. Accessibility will continue to be a problem until there is both a system in place for data access, and also a long-term plan for funding the cloud model. The GTEEx team are working on hosting of their upcoming V8 release with ANVIL.

Both GTEEx and their users want their data to be easier to find, i.e. browse and combine with other datasets, and useable by more people, for more projects. Recently, GTEEx surveyed their user base, and found that their users are extremely interested in locating data that can be used together. They asked:

If we were able to integrate GTEEx protected (dbGaP) data into the GTEEx portal, would you find it useful to query and visualize sets of

data grouped by phenotype (e.g. actual age, medical history, genotype, etc.)

Would you find it useful if we were able to integrate other protected datasets, such as TOPmed, with the visualizations on the GTEx Portal?”

Of their 112 respondents, ~87.5% said yes to each of these questions. Although adding extra metadata into their portal is a straightforward task, GTEx has no immediate plans to incorporate these other protected metadatasets. This is because 1. dealing with the protected metadata aspect is currently intractable and, because, 2. they cannot financially support new staff to do the work. Onboarding the protected metadata would require integration with FireCloud/Terra, which is a FISMA Moderate platform. This means GTEx itself would have to become FISMA Moderate, and the portal currently does not have this level of access control. Bringing the portal up to FISMA Moderate standards would require a great deal of documentation and some re-programming of the portal. For example, currently the metadata used to draw the interactive plots in the portal is readable by any user that views the webpage source code and understands html. So, while the portal is fine for displaying publicly available metadata, it would unacceptably expose any restricted metadata that runs through it, and the codebase would require extensive re-writing to adequately protect private data. There are teams at the Broad who have the ability to do all of this work, however GTEx doesn't have the resources or funding to support them. GTEx also noted that additional funding through the OT mechanism would not help, as they cannot make short contracts for engineering hires (<12 months). They would require a stable contract of at least one year to be able to hire people to improve this, or indeed any, aspect of their operation.

Interoperability and Reusability

GTEx is somewhat less concerned about the other aspects of FAIR: Interoperability and Reusability. The concern with these concepts is similar to the one they raised about metadata 'levels': whether a dataset or data center is 'interoperable' is based on what other dataset or data center you are comparing it to, and these pair-based comparisons are difficult or impossible to generalize. They will always be project-interaction-specific and context dependent. Similarly, whether a dataset is 'reuseable' is extremely difficult to judge, and contingent on people actually having done it, i.e. "data is not reusable unless it has been reused."

Authentication/Authorization

Covered in [Appendix J](#).

Data Dashboards

All of GTEx's raw data is currently hosted on Google Cloud, so they have little use for this utility. If it could monitor access (which is currently a problem) or help with reporting usage statistics, they might find that useful.

CF Data Portal

GTEx has a refined and interactive portal that allows their users to explore their publicly available metadata and data. Any user can access the GTEx portal without logging in and use it to:

- Search for data with specific metadata terms
- Find what genes are expressed, and at what level, in any tissue
- Compare gene expression between tissues
- Look up expression for a gene or genes across tissues
- Generate a PCA of expression subset by any public metadata
- Find and view eQTLs and make basic comparisons
- View histological images for over 25,000 samples
- Users with a Google login can submit requests for GTEx biosamples.

In the next few months, they will be rolling out additional features to allow their users to do more complex eQTL analysis.

Their portal is extremely well used by their user community, and is popular because it requires little expertise to navigate and create useful figures. They see about fifteen thousand users per month on GTEx portal, with over a hundred and forty thousand page views. They are a globally used resource with roughly 50% of their users from the US or the UK, with the rest being spread throughout the world. They serve a full spectrum of users on the portal, from bioinformaticians to clinicians and genetic counselors. Their portal is written in D3, and Python, with a Javascript front end, and REST back end. The database is MongoDB and Google Datastore, and uses the Google App Engine.

Although GTEx is interested in a CFDE portal to connect users to new datasets, they suggested that the specificity of each DCC will make a single data analysis portal impractical. While some of their portal features are generic, most are specific to this project and their data organization. GTEx has a very specific picture of how their users want to interact with their data, and has tailored not only their portal, but their underlying analysis pipelines, to that vision. They expect all the other DCCs to do this as well.

All data coordinating centers are well-equipped to know the perils of their data, how users misinterpret data, and the most common ways that their data can be mis-used. GTEx reports that while they don't know the experience level of their users, they have reason to believe that most of their users do not have the technical sophistication level to join GTEx data to other datasets in a statistically appropriate way. Many of the help requests at GTEx are low level bioinformatics questions that are not GTEx specific. Their users almost never ask about the type of statistical or technical issues that they would need to correctly combine their own data with GTEx results, which is a common use of GTEx data; and a recent paper found that most papers, 35 of the 50 sampled, did not use an appropriate method to compare the GTEx eQTLs with GWAS data (<https://www.nature.com/articles/s41588-019-0404-0>). This will only be made more problematic if users are able to link between GTEx datasets and other resources through the CFDE, without appropriate guidance and/or training.

This example suggests that these type of data (GWAS, eQTLs) should be more readily available as harmonized resources. Although it would take a lot of time and resources to generate this kind of data, it would facilitate more sophisticated and statistically sound analyses for less expert users, who tend to look up individual SNPs. GTEx would not support a portal that allows an end user to naively combine processed datasets across DCCs unless they could be certain that all of the data a given user might choose had been run through the same pipeline and could be trusted to give scientifically valid answers. However, they would be in favor of a system that has expertly harmonized data for popular GTEx features (GWAS, eQTLs), and one that allows the individual interfaces of each DCC to pass data back and forth in a more facile way. They would also support a system that allowed users to find datasets across data centers, and think it would be of value to document issues concerning how to properly build and analyze new cohorts and to make those tutorials available to the wider community.

Data Platform

Data Hosting

GTEx has historically been the most frequently downloaded data from dbGaP, however that may no longer be true: due to their participation in the DCPPEC, all of the GTEx raw data is now cloud based, and was set up to be accessed via the DCPPEC access solutions. GTEx itself uploaded a copy of all the raw data to the Google Cloud, while a second copy of the GTEx data was made available on Amazon Web Services (AWS) by NHLBI. GTEx is still paying to host all of this data in the cloud, but it is inaccessible to end users.

The DCPPEC solution for the Google Cloud involved hosted pointers to data storage 'buckets'. Unfortunately, when the DCPPEC dissolved, the pointer hosting went with it, so although the data is in the Google Cloud, there is no way to access it. The AWS solution used signed URLs, and was set up with incorrect authentication. When the DCPPEC ended, work on the signed URL

system halted, and the authentication issue was never fixed. As such, users can see that the data exists on AWS, but cannot download it because the signed URLs don't authenticate correctly, and users can't get to the Google data bucket at all. The SRA database does not have the complete GTEx dataset, so at the time of our meeting, a large proportion of GTEx data was completely inaccessible to anyone outside GTEx.

Even once access is fixed, GTEx has concerns about data storage in commercial clouds. First, GTEx currently only has funding to maintain their portal. They currently pay cloud hosting fees and all their other costs, using money leftover from the original grant. However, those funds will not last much longer, and it's unclear who will pay for the cloud storage long term, or what will happen to the GTEx data when their funds run out. There are also issues with user costs. As difficult as dbGaP is to navigate, it is free for users, and that's something users really like. Cloud buckets cost money for data storage, data processing, and data movement, i.e. downloads. Currently the GTEx buckets are set to be 'requester pays', that is, the user must pay download fees, which is expensive for the user. However the only current alternative would be for GTEx to incur the user download costs, which they don't have funding for. A solution to the download problem would be to encourage users to work entirely in the cloud, and to simply access the data rather than downloading it. However, this has its own problems. Most GTEx users don't know how to work in the cloud, or how to move their analysis pipeline there, and would need a great deal of basic bioinformatics training that GTEx doesn't have the resources to provide, and wouldn't fit their mission. It also simply moves the cost from 'downloads' to 'compute time', and there is no good way to mitigate those fees. Supplements to NIH grants, for example, do not work for non-NIH fundees, and as nearly half of GTEx users are international, they would likely not be covered by any US based supplement system. It will also be difficult to get users to move to a cloud-based model when many are at universities with on-premise High Performance Computing centers. HPCs are typically partially supported by overhead, and so are "free" for compute time and storage. These users have very little incentive to move to a cloud system that costs much more out of pocket and gives them only transient access to the data.

Data Analysis

GTEx are working with ANVIL to enable the next dbGaP release of the data to be available in ANVIL/Terra.

Training

Helpdesk tickets at GTEx are handled by everyone on the team as they come in, via their internal request tracker. This is an important feature because users can ask questions and get answers from the person who best knows a given topic. Many of the questions are "one off", and range from simple to sophisticated, however they are skewed towards the basics. Many of the questions are not particularly GTEx specific, and instead are about how to conduct general

bioinformatics analyses, for e.g. how to do RNA-Seq. This is something that training could easily address, however GTEx does not have the time or resources to create or deliver such training. It's also not clear that it is something they should put resources into, when there are so many GTEx specific data analyses they could be teaching.

GTEx also has several other ways they reach users. Their portal has documentation and a small FAQ, they have held a small number of in person workshops, and they host three videos on YouTube with a combined total of about four thousand views:

- GTEx: Genotype-Tissue Expression 1062 views
- GTEx Portal: Introduction to the Gene eQTL Visualizer 2731 views
- GTEx Portal: Viewing Gene Expression Data on the GTEx Portal 469 views

GTEx has had mixed results from these training efforts. On the one hand, they know their users crave training and outreach. Their workshops are always over-subscribed, and their help desk system is always filled with new questions, however their user base does not seem to be getting more sophisticated. As previously discussed, much of their help desk traffic is users who need basic bioinformatics training, and another large segment is users asking questions whose answers could already be found in the portal. For these latter questions, it's not clear whether the problem is more about the site, i.e. the answers are difficult to find, or about the user, i.e. it's easier/faster to email the helpdesk than to look for the answer. In either case, it is clear that a great deal of GTEx resources are going into supporting naive users who might be better served by a more general bioinformatics training resource.

This is an issue as GTEx's main goal with providing training is to increase the usage, and depth of usage, of their data. They want to enable their users to deeply understand the data, to increase their ability to use it on their own, and to increase the sophistication of the questions they answer with it. To that end, GTEx is very interested in running a hands-on API workshop. This would require some kind of library implementation, but would result in users with much more flexibility in how they query the data. As GTEx already has the expertise to build this workshop, this could be a potential burst funding investment compatible with the OT mechanism.

Their other workshops would likely benefit from some restructuring. GTEx is concerned that perhaps one reason they don't see an increase in user abilities after workshops is that they try to cover too much, too fast, and that learners leave having learned many things, but at too low a depth to be useful. Rather than try to cover all topics, they would like to try having more workshops, but with each one covering a narrower set of topics. This is a workable goal, but GTEx doesn't currently have the time or resources to spend on restructuring their lessons.

Outcomes

GTEX hosts a widely used dataset for a community with a broad spectrum of both biological and computational abilities. They are interested in expanding their user base, but even more keen to expand their user's level of bioinformatic sophistication. They are constantly striving to offer tools that allow a deeper and more nuanced view of their data without demanding additional computational expertise on the part of their users, and everyone at GTEX contributes to giving personalized user support.

CFDE Targets

- Mailing List: GTEX has a mostly defunct mailing list set up through Mailchimp. The CFDE could help revive and increase the reach of this mailing list to help with training and reaching out to users.
- Workshop Development: GTEX needs more targeted training modules, and some new materials. The Brown lab from CFDE can provide resources to rapidly develop the training materials, as well as some personnel for running workshops.
- Reliable funding: GTEX cannot hire for 6 month projects. Steady, longer-term funding through the CFDE would give GTEX the ability to fix many of their blockers.

Major Use Cases:

- Query by gene, query by variant, by tissue
 - A user can query and visualize sets of data
 - Ideally, this would include data grouped by protected variables (e.g. actual age (instead of age in 10yr intervals))
- Data download
 - A user can use the results of a visual query to download the raw data, or link it to an analysis platform
- GTEX API (usage is untracked currently) - REST API
 - A user can run the GTEX visualization tools on any dataset with the appropriate data types, independent of the GTEX platform
 - Ideally, this would be used not only by end users, but other DCCs and would foster interoperability between data centers

Game Changers

During our discussions with GTEx we collectively came to realize that there are several ways the CFDE might significantly advance the state of NIH biomedical data management across all of the Common Fund DCCs by accomplishing a targeted set of goals that were not originally included in our proposal. If these goals are shared by most DCCs, we could positively impact their internal operations and make important new advances in data access and analysis for the NIH research community. These targets are labeled as "game-changers" because by accomplishing these innovations, the CFDE has the potential to dramatically improve the access and usability of an organized set of resources hosted at the Common Fund DCCs.

1. Reduced barriers to search across all data at dbGaP. GTEx has been very successful in working with their users to understand the challenges that users experience, and a continuous theme throughout our discussions was the frustration experienced by users accessing data at dbGaP. For example, in a recent survey they conducted users were asked:

If we were able to integrate GTEx protected (dbGaP) data into the GTEx portal, would you find it useful to query and visualize sets of data grouped by phenotype (e.g. actual age, medical history, genotype, etc.)?

Of the 112 users who replied to their survey, 87.5% answered yes to this question. Users face many challenges in using dbGaP:

- Applications for data are cumbersome, and users must submit a separate application for access for data at each Common Fund DCC
- dbGaP data associated with individual dataset is very poorly documented
- Data file names from DCCs are idiosyncratically changed on upload such that neither users nor DCCs can tell whether they're using the same file
- User interfaces for viewing that data are quite superficial
- There is no uniformity in the way that data is described from one project to another

The GTEx staff were adamant that a critical need for their users would be reduce the administrative barriers associated with users to be able to access dbGaP data, to improve the user interfaces that enable researchers to query on those data, and to reduce or eliminate the barriers for users to compare dbGaP data across Common Fund projects. One relatively simple, but impactful, step towards this goal would be simply making public what fields of private data exist for each dataset.

2. CFDE compliance, component 1: establish the minimal metadata for a resource object. GTEx staff recommended that an important development would be to create a standard

that describes a minimal set of metadata for an individual file, such as an RNA-Seq file associated with their project. Implementation for this standard would not be complicated, and would simply involve an agreed upon electronic format and fewer than 8 metadata terms that are relevant to each file. The minimal metadata terms would likely include: originating institution (e.g., "Broad Institute"), file type, tissue source and species name for the sample, a unique identifier, and a checksum for the file. While many implementations for electronically encoded metadata for biomedical resources have been proposed in the literature, no standard has been adopted by the broader user community. The GTEx group suggested that given the position of the CFDE project among several DCCs, and the funding incentives that could be offered by the Common Fund, the CFDE has a high likelihood of achieving adoption across several groups.

3. CFDE compliance, component 2: resource object collection. The ability to bundle a list of CFDE resource objects into an machine-readable file was also highlighted in our discussions. These lists are similar in function to users collecting a shopping list on a commercial web site. In this case it would be useful for objects (such as RNA-Seq files) to be identified by user queries at the CFDE portal, and for the object list to be passed to other resources such as the GTEx web site (to retrieve the data), or to a computing analysis resource such as Terra. We refer to these lists as *manifests*, and similar to the minimal metadata description described above, implementation of the manifests will not be complicated. Potential implementations have been proposed in the literature, and the Kesselman and Foster teams also proposed utilizing the bdBag system for manifests during Phase 1 of the DCPPEC. While a standard for manifests has not been adopted by the broader community, the CFDE project represents an excellent opportunity to create a standard for all of the Common Fund DCCs. CFDE Cost Incentives: All senior GTEx members in attendance at the meeting concluded that an extremely important incentive to get their group, as well as other DCCs, to participate in the two compliance components described here would be for the NIH to offer CFDE cost incentives - e.g., reductions in storage costs on the cloud for DCCs where were compliant with those technologies.
4. A proof of concept: enabling other DCCs to use GTEx visualization tools. One potential use for adopting a CFDE manifest is that a collection of files could be transferred to other resources such as RNA-Seq analysis tools. The GTEx staff have previously developed several visual analysis systems that are available as downloadable tools for their users. These tools were originally created for users who wanted to combine their own data with GTEx and be able to perform analyses that were similar to information presented at the GTEx website. Users are able to install the tools, link them to their own data, and create high quality figures that can then be used in publications. One suggestion that emerged from our discussion was to link these tools to RNA-Seq files from another Common Fund DCC, using the CFDE manifest as mechanism.

5. A CFDE search plug in. Another benefit of all Common Fund DCCs adopting a single set of minimal metadata for a research object, and a common manifest for describing lists of CFDE data is that it would then be relatively simple to create a web based plug-in that could be provided to all of DCCs, to assist with accessing data between sites. The advantage to creating a common plug in for all of the sites is that it would reduce costs associated with multiple DCCs creating interfaces that perform the same function, and would simplify the user's experience by creating a single search tool. By having a single development team create the plug-in, it would also mean we could rapidly respond to changes to the underlying implementation of the minimal metadata standards and CFDE manifest system.
6. User Helpbot: a first point of contact help desk for all CFDE users. The GTEx group has been systematically tracking helpdesk requests over the past few years using an RT tracking system. The GTEx staff have also generated documentation, lists of Frequently Asked Questions (FAQs), and other materials. GTEx staff report that somewhere between 25-50% of the help desk questions reflect requests from users that are more related to basic bioinformatics questions, than they are associated with the GTEx dataset. The GTEx staff also noted that many of the submitted questions would be answered if users simply read the FAQs. This observation led to the realization that creation of the CFDE is likely to *increase* the help desk burden of all DCCs. In addition to creating new, potentially confusing, ways for users to combine data, the CFDE will draw additional users to all sites, many of whom may lack significant bioinformatics training. During our discussion the group realized that it would be of significant positive impact to create a common front-end "helpbot" service that would be available at all DCC web sites created for the CFDE. This service could potentially use AI approaches to first filter user questions, to see if questions could be answered from FAQ content, and to handle basic bioinformatic questions that were not related to any particular DCC. This "user helpbot" could significantly lower the support burden for each DCC, and unify the bioinformatic educational information supplied to CFDE users.
7. Reducing costs for FISMA compliance. The Federal Information Security Management Act (FISMA) defines a set of controls to protect computer information and operations from security breaches. Requirements include maintaining an inventory of IT systems, categorizing data systems by risk level, maintaining a security plan, utilizing security controls, conducting risk assessments, obtaining certification, and conducting continuous monitoring. These activities often incur more than \$100,000 in costs to institutions obtaining FISMA compliance. However, the benefits to obtaining FISMA compliance is that each DCC would increase it's likelihood to host human protected data, obtain trusted partner status for managing data (even just metadata) that are currently hosted at dbGaP, and to share data with other FISMA-compliant DCCs. GTEx staff encouraged us to examine mechanisms that might be supplied from a centralized service that would help reduce the cost burden of obtaining FISMA compliance for these reasons.

Agenda

MONDAY June 3, 2019 - Broad Institute – 415 Main Street Cambridge MA 02142

Room Location: 75A-11-Teton (11081)

9:00am(ish)-9:30am - Introductions

DCC Attendees: Kristin, Jared & Francois

Short introductions from engagement team members and attending DCC members.

9:30am-10:00am – GTEx DCC overview

DCC Attendees: Kristin, Jared

Short overview of GTEx: Results from GTEx user surveys; User statistics; Data statistics; Current metadata harmonization; Current training and outreach

10:00am-12:00pm - Goals Assessment

DCC Attendees: Kristin, Francois, Jared, Katherine Huang, Duyen Nguyen, Shankara Anand, Aaron Graubert

GTEx to rank these goals by importance and potential timeline, followed by discussion around why they rank goals the way they do,

Goals List:

- | | |
|------------------------------------|-------------------|
| • Self-governed metadata standards | • Data Dashboards |
| • Harmonized data | • CF Data Portal |
| • Cross cutting metadata models | • Data Platform |
| • FAIR assessment | • Training |
| • Authentication/Authorization | |

12:00pm-1:00pm – Lunch

1:00pm – 4:00pm Open discussion (with breaks)

DCC Attendees: Kristin, Jared & Francois

Using the results of the mornings exercise and a collaborative format, iteratively discuss goals, blockers, etc., such that the DCC agrees that the engagement team can accurately describe their answers, motivations and goals.

Topics:

Infrastructure (KA, JN, FA):

- What has been your experience with uploading to the cloud?
- What challenges have you faced?
- How have you dealt with those challenges?
- How well does dbGaP work for GTEx?

Review of metadata (KA, FA):

- What's metadata is important for your org? For your users?
- Do all of your datasets have approximately the same metadata? Or do you have many levels of completeness?
- Do you have any data already linked to outside resources?
- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- How do your users obtain metadata and raw data?

Use cases (KA, FN, JN):

- What if any, use cases do you have documented? Undocumented?
- What things are users currently able to do with your data?
- What things would people love to do with your data, but currently can't (or can't easily)?
- What are the challenges associated with those desired uses?

Training (KA, FA, JN):

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?
- What challenges keep you from offering the training you'd like?

User Interactions/HelpDesk (KA, FA, JN):

- How do you interact with your users? Who on your team does these interactions?
- What kinds of issues to you most frequently handle?
- What lessons do you wish you could impart on users, but currently can't?
- Does your org have any dream initiatives that could be realized with extra resources? What resources would you need?

Policies (KA, JN, FA optional if need the time):

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR (JN, KA):

- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

TUESDAY June 4, 2019

Room Location: 75A-5-Biscayne (5021)

9:00am-12:00am Review of goals and CFDE involvement

DCC Attendees: Kristin, Jared & Francois

Review of what topics are priorities for GTEx, and discussion of where and how they would utilize extra personnel/money/other resources

Appendix D - Kids First Site Visit

Center for Data Driven Discovery in Biomedicine, The Children's Hospital of Philadelphia
2716 South Street, 12th Floor, 12312

Tuesday June 25, 2019

Room Location: 12105

Meeting Logistics

We held a meeting with Kids First group at the Children's Hospital of Philadelphia on Tuesday June 25, 2019 for a day and a half. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the day. Representatives in attendance from the CFDE were: Amanda Charbonneau (UCD), Brian O'Connor (Bionimbus), Brian Osbourne (BioTeam), Titus Brown (UCD), and Owen White (UMB). The primary representatives from Kids First were Adam Resnick (PI), Allison Heath (Director of Data Technology and Innovation), Jena Lilly (Director of Operations & Strategic Planning), Bailey Farrow (Technical Project Manager), Yuankun Zhu (Bioinformatics Engineer Supervisor) and Tatiana Patton (Clinical Research Program Manager). We also met briefly with Amanda Haddock, a patient advocate who is the president and co-founder of the Dragon Master Foundation. After the death of her young son from brain cancer in 2010, Amanda started the DMF to help speed biomedical discovery and empower cancer researchers and she works closely with the Kids First staff.

The engagement team began by reviewing their goals for the meeting. These goals include learning about the structure and goals of Kids First, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for their group. In turn, Kids First provided us with a wide-ranging overview of their organization, and gave us a great deal of insight about the intersection of big data and patient care.

Meeting Logistics

Kids First Overview

Harmonized data

Cross cutting metadata models

Self-governed metadata standards

FAIR Assessments

Findability and Interoperability

[Accessibility](#)

[Reusability](#)

[Authentication/Authorization](#)

[Data Dashboards](#)

[CF Data Portal](#)

[Data Platform](#)

[Data Hosting](#)

[Data Analysis](#)

[Training](#)

[Outcomes](#)

[CFDE Targets](#)

[Major Use Cases:](#)

[Game Changers](#)

Kids First Overview

The Kids First Data Resource Center (KF DRC) was created to improve collaborations across disease communities and to help translate research into personalized medicine, with a focus on childhood cancers and structural defects; primarily by building a portal that would connect and coordinate the huge amounts of data being generated by the Gabriella Miller Kids First Research Act. This act, signed into law in 2014, was in response to advocacy groups pressuring congress, most notably, the eponymous Gabriella Miller, a ten year old girl dying from an inoperable brain tumor. During her illness, Gabriella was an outspoken activist for childhood diseases, and shortly before her death, she challenged Congress to stop talking and start doing something, in a widely circulated video.

Gabriella Miller's call for action is clearly evident in all of the activities at the KF DRC. Their *raison d'être* revolves around the speed at which they make more data more accessible to more researchers because it is vital to improving outcomes for their patients. They were encouraged by the concept of the CFDE, because "...a commons accelerates progress and increases accessibility, and is a more efficient way to drive from data to knowledge to impact." Until meeting with KF staff, our thinking involved reducing the time needed for gathering data, or running analyses in terms of money saved and convenience for the researcher. However, for the people at KF, ***time is the obstacle to translational research***. As the KF DRC staff made clear multiple times during our visit: they are measure the passage of time in children's lives.

The Gabriella Miller Kids First Research Act reappropriated the Presidential Election Campaign Fund to instead fund a 10 year, \$12.6 million research program into childhood diseases. As the research program was rapidly created by Congress, there was almost no organizational framework or direction for the project. The NIH put out RFAs for a sequencing center and data coordinating center in 2016, but researcher-led sequencing began almost immediately. Structural birth defect researchers started to sequence trios, while the cancer researchers focused on tumor vs. normal comparisons. Data was collected across all pediatric cancers and structural defects, by a wide variety of people with very different goals, funding and sophistication levels. For example, structural birth defect researchers often want to identify phenotypes that are underpinned by genetics and rely on standards, ontologies, etc. while this strategy is rarely used in pediatrics. Many of the data generators are small groups, funded by philanthropy, or other sources, including the NIH and investigator driven grants. The differing research norms between disciplines as well as the overall lack of cohesion created tension between these various research communities, and a wide-ranging, but poorly overlapping sets of data.

The Kids First Data Resource Center, based out of the Children's Hospital of Philadelphia (CHOP) officially received funding in mid-2017, and has made an enormous amount of progress, especially given the resources available and the complexity of the situation they were charged with managing. In less than a year, Kids First was able to deploy the alpha version of their portal, and could support access to data. At the time of our meeting, the DRC was not quite two years old, but already had a robust data access infrastructure, a sophisticated query system, and a responsive help team which supports about 2000 users per month interacting with almost three petabytes of data.

Harmonized data

Harmonized data is vital to the operations of Kids First because the clinical researchers using their data must be able to access patient phenotypes and clinical variables associated with information derived from a diverse set of data generation sites. This represents a significant challenge for them, because when the KF DRC started, data collection had already been in progress for about three years by researchers on 23 different X01 grants as well as four sequencing centers, with no clear standards for metadata, analysis or storage. Melissa Haendel's group was brought in to handle metadata curation, and there are five to ten people on the clinical and phenotypic side of the operation who work on harmonization. KF have re-curated the metadata multiple times; for instance, for diagnosis phenotype they have added new terms to the Human Phenotype Ontology and Mondo (the Monarch Disease Ontology) three times over the last year.

They face several challenges in the type of metadata that they collect, that are partly a result of harmonization, and partly a result of the complexity of their clinical domain. For example, datasets at KF frequently incorporate time. Clinical phenotypic data has "events", such as doctors visits and updated diagnoses, and clinicians may add in data after samples are included

in the database, so there is longitudinal information and diagnoses can change, or expand, over time. The difficulty is further compounded by the broad, rather unfocused mandate of the Gabriella Miller Kids First Research Act. The research scope is “anybody and everybody” with a pediatric cancer or structural birth defect, and so encompasses hundreds of distinct modalities. Due to both a given patient’s needs and a researcher’s interests, clinical records for a given patient may have incredibly rich and deep metadata about the specifics of their focus area, and only passing reference to others. For example, in Epilepsy studies, coordinators who are collecting brain tumor data may just say “seizure” whereas Epilepsy researchers will detail the type of seizures. So specialists from different communities generally do not speak in the same level of detail about the same symptom.

Datatype compatibility is also a concern. For example, while the KF data only includes RNA-Seq data from two cancer research groups, who use the same strategies, combining their data is not trivial. To account for batch effects and allow for differing downstream analyses, the KF DRC carries over all metadata from these studies. There is a constant need to revisit pre-analytic variables, so as to remove batch effects as well as iterative refinement of what metadata information to collect, and to pass on to users.

The KF DRC handles the idiosyncrasies of metadata with these issues by taking a practical, minimalistic approach, and have used it as an opportunity to build an ecosystem that is vertically and horizontally integrated. The KF metadata model uses data definitions to make data findable, and allow the user to see as much metadata as is available. The KF DRC data curation tool imports/loads from user spreadsheets, and serves as a first-level data ingestion tool for the team. KF has taken the time to discuss metadata terms with their users, and so in most instances, the appropriate terminology is used in the portal. In these cases raw spreadsheet data from a data generator can be computationally imported, but all data still requires some manual curation. When metadata is harmonized, KF tends to err on the side of preserving the original metadata terms, e.g. the text from the source Excel file, as well as the harmonized data.

They noted that legacy data is will also be a significant problem; since data collection pre-dated the curation effort, the early KF datasets tend to have idiosyncratic pipelines and metadata terms. Even data generated since mid-2017 has these issues, as the KF DRC has no authority to (or interest in) dictating a metadata standard. Pediatric oncology is full of rare diseases, and research is often driven by single point labs. These labs often have their own specialized metadata requirements and pipelines dictated by patient care needs which are not amenable to standardized metadata models. Further, for some data generators, there is still a barrier for how much work people are willing to do to supply them with metadata. A data generator will give KF a spreadsheet, or raw forms, but are often unwilling to do any harmonization, or supply clarification of terms. Generally data providers are not funded to curate that data for KF, and so they have no incentive to increase their effort.

KF minimizes technical harmonization variables by ensuring that everything uses hg38 as a reference and they suggest the Broad Institute’s Genome Analysis Toolkit as a best practice.

Any data they receive that is processed using other tools is reprocessed by KF before being added to the collection. They run the genomic or RNA-Seq workflows using CWL, have Dockerized tools for each step in their those pipelines, and they use the NCI Genomic Data Commons data dictionary to capture other technical variables like read group and read length info. For genomic data, which accounts for the vast majority of their portfolio, these practises make harmonizing technical variables relatively straightforward.

Cross cutting metadata models

Harmonization is a difficult, multifaceted, and just as the CFDE technical team would agree, the KF staff stated that a cross cutting model will not serve as a magic bullet to solve all harmonization problems. The KF team agrees that a common metadata model will be useful for searching across CF DCCs, but it would not inherently solve the types of harmonization issues they face. Simply stated, the C2M2 is likely to just bring a lot of incongruous metadata together into a single collection.

For clinical variables, the KF DRC relies heavily on Fast Healthcare Interoperability Resources (FHIR), a standard for health care data exchange based on the HL7. FHIR can handle 60-70% of the KF DRC use cases right now, is an open source and flexible framework that they can add data resources to as needed, and the most difficult part of implementing it was to create a FHIR server. The core data model of FHIR is Argonaut, which Google, Apple, and Microsoft have bought into, along with Epic, the system that manages more than two-thirds of all medical records in the United States. For hospitals and communities that want to participate in research, FHIR is a bare-minimum point of contact, and it is even patient accessible. For instance, you can use FHIR to integrate your personal medical records with the Health App on your iPhone.

The KF DRC staff were very excited about FHIR and its long term implications for managing patient metadata. Currently, FHIR doesn't really work for non-human data, and the specimen and genomic sequencing data in the FHIR system is shallow, and based on what a clinician would see in an electronic health record. Typically, hospitals will report only a digest of sequence analysis. However, FHIR is very focused on clinical reporting of known data, and KF is sure that FHIR could be extended to deal with the research space. In fact, there is already some work in that area. FHIR genomics can deal with whole genome sequences, and GA4GH seems to be adopting it as a standard. KF staff also pointed out that within two years, they expect every patient at CHOP, nearly 30,000 children a year, to be routinely sequenced, and stated that "the minute clinical data will routinely include whole genome data, the FHIR community will immediately adopt it." Using FHIR also aligns well with the overall goal of the KF DRC, which is to improve patient outcomes, and that relies on keeping -omic data integrated with its phenotypic data.

Self-governed metadata standards

Due in a large part to the overwhelming variety of metadata in their datasets and the fast pace of their field, KF DRC staff did not see much value in adopting overall metadata standards, regardless of who governs them. Instead, they stressed the need to be forward thinking and build resources for the way the world actually works rather than the way we might like it to be. “We cannot build rules and infrastructure that prevent you from doing all the things users think they want to do.”

They also pointed out that even simple, seemingly easy sounding standards can be complicated in practise. For example, we suggested that one minimal standard might be that metadata has to include a Global Unique Identifier (GUID), so that a user could be sure they had a specific dataset. Allison Heath rightly pointed out “when you talk about files, this all works...but when you talk about rows in a database... how does that work?” Datapoints in the KF dataset aren’t a single, static genome sequence, they’re children who are often under aggressive treatments for deadly diseases, and their phenotypes are constantly being updated. Given these kinds of scenarios, GUIDs would have to be handled carefully, and GUIDs might need to be assigned to individuals rather than datasets. At the scale of the overall dataset, like records from FHIR, changes might happen daily, and even each individual person might ‘version’ with every appointment.

While uninterested in top-down standards, KF is extremely interested in the idea of building a self sustaining community, both for researchers and for DCC staffers themselves. Right now there is a gulf between staff at different DCCs to communicate with each other to talk about many important topics such as harmonization, data models, and dealing with protected data. Creating spaces for DCCs to engage with each other would go a long way towards simplifying both the current effort to improve interoperability and future efforts to integrate between datasets. In the short term more socially integrated DCCs will be more likely to find compromises and shared ground on topics like authentication. In the long term, they will be more likely to choose compatible solutions to future shared technological problems, because they will have been able to share ideas. Creating a DCC network will also improve the overall sustainability of Common Fund projects, as older DCCs can share institutional knowledge with up and coming DCCs, and perhaps even serve as talent reservoirs. For closely aligned projects, seeding new DCCs with experienced, onboarded personnel from sunseting DCCs would allow them to get up to speed more quickly, and to benefit from prior Common Fund investments. One observation from the meeting was the recognition that building a true community of DCCs might will be most successful when all the Common Fund DCCs participate, it will be far less successful if an isolated set (or just two) DCCs work to cross-reference their data or build common standards.

FAIR Assessments

The KF DRC is very interested in FAIR principles, primarily due to their belief that making the data more accessible to more researchers, faster, is the key to improving outcomes for their patients. The KF staff are dedicated to every element of making data more accessible. As an independent entity, their DRC dedicates an enormous amount of time and energy to all elements of making their data assets findable, accessible, interoperable, and reusable -- what was evident from talking to them however was that there are no clear guidelines to operationalize FAIRness. Other than their own expertise, there is a vacuum of guidelines for them to operationalize FAIRness improvement.

Findability and Interoperability

Findability and Interoperability are inextricably linked in most of KF's use cases and both the KF DRC and CHOP are interested in supporting environment search and interconnectivity. This reflected by the KF portal, which potentially has the most advanced query capability among DCCs, enabling users to search across a wide array of variables and see the data in real time. Still, in keeping with their 'we have to accept the world for what it is' philosophy, they would like to replace it with an even more sophisticated system. "We've got to break out of facets, they just don't work!" They think a free text google-like search is necessary, especially for metadata terms that may be recorded in multiple ways: RNAseq vs RNA-seq vs rna seq. Search at KF mostly starts with diagnosis and phenotype, but depending on the intended cohort, they may also start with "genomic" terms or biospecimen information, for e.g. tumor descriptions. Researchers want to drill down into the data, and even very complicated boolean searches can't always build a cohort that corresponds to their study question in the way a free text search might.

Cohort building is also an iterative process. Researchers want to be able to share their current query state and send it to collaborators, who should also be able to edit those queries. They want to support saving queries and supporting sharing with other researchers: e.g., operations like save cohort, share cohort, and manipulate cohort. There are also longitudinal aspects of cohort building, family structure, and complex pedigrees - information KF would like to conserve. And they want to allow users to search for information about events over time. True to KF's primary focus on quality of care, they are aware that the end user must be supplied with all information about the patient as it changes over time.

KF also discussed several forward thinking ideas about genomics search, again in the spirit of not hamstringing the creativity of future researchers. These included 'layering', the idea of using entire datasets as a way to think about findability. Here, a clinician might start with genomic data to look for a specific feature. Then, based on their interpretation of that data realize that they need another specific layer of data to overlay onto it: "to check this SNP, now I need single cell data". Similarly, they suggested that one way to deal with interoperability would be to set some sort of interoperability thresholds that are computed as a user sets up an analysis. For example,

when a user saves an analysis pipeline file, Cavatica, Kids First's data analysis platform, might issue a warning that use of RNA-Seq dataset A with pipeline B **might** result in overestimation of expression differences. This way a naive user isn't left to their own devices to perform a potentially ill-advised experiment, but expert users still have the freedom to explore. They also suggested other user alerts that might push ahead stalled research, such as alerting users when the files they have downloaded have new metadata. Or that files that look like ones they have previously used are now available. KF is also very interested in fostering social communities that may be formed by users of their portal. Essentially, KF is thinking about ways to allow researchers to search for exactly the data they need, for the specific problem at hand, rather than trying to build systems based on constrained metadata terms.

Finally, KF talked about FAIRness of citations. The KF DRC, as well as their patients and data contributors would like to be able to answer questions like:

- If I find an interesting cohort, how many of the individuals have been used in publications?
- If I own this data as KF, how do I know it was used?
- How many people are using my child's data?

Accessibility

Kids First has essentially solved accessibility from the standpoint of their users being able to get KF data. Their portal allows anyone to query the unprotected data and requires only a simple sign up, and the portal itself has sophisticated logic for advanced queries, based on real use cases. A user with access to protected data can quickly add their credentials (during our visit their demonstration took about thirty seconds), and users can push cohorts selected on the portal directly to Cavatica, their data processing platform provided by Seven Bridges.

However, Kids First stressed that accessibility in terms of understandable metadata as typically described by FAIR is only part of the true problem of accessibility. Researchers have few problems finding and accessing data at KF, however they have little or no ability to effectively engage with the resources. So, while the data is **technologically** accessible, it is not **intellectually** accessible. Small datasets (like for rare diseases) require the cloud so they can be compared and analyzed; and clinicians have specific, sophisticated questions they want to ask with the data, as well as an intimate knowledge of the datasets and their metadata, but they don't have the bioinformatics knowledge to do an analysis on the cloud. KF points out that the difficulty with 'Big Data' is not *just* that we have too much data, it's that the cloud is providing new opportunities that are only currently accessible to a handful of specially trained people.

The KF DRC is also interested in accessibility in the other direction. That is, how accessible is their system to people who want to input data. They discussed this in two contexts. First, they want to ensure their system is accessible to data creators, such as sequencing centers and researchers, who currently have little incentive to upload complete metadata (as discussed in

Cross cutting metadata models). The easier the process of inputting data, presumably, the more likely a given researcher will comply.

Second, KF talked about making their infrastructure accessible to patients. Patients and their families should have the ability to upload their own data to KF, and “get the data out of the hospital”. Whereas the NIH is often focused on keeping human metadata private, KF and the patient advocate we spoke with stress the importance of liberating metadata. Patients at Kids First often have terminal diseases, and interventions like experimental treatments might be their only hope. Families are therefore desperate to get their child's DNA sequence into the hands of as many researchers as they can, as quickly as possible. Currently, many families organize their own DNA file sharing via Facebook, a practise that KF is very concerned might become predatory, for e.g. pitches like ‘for only \$10,000 I’ll analyze your kids genome’. Similarly, as noted in the introduction, embargo periods for X01 awardees, while short for research, seem like an eternity to families. Staff at KF related having to face parents asking questions like: “Why did you sequence my son’s tumor if it did not impact his treatment?”. Giving patients the ability to drive the accessibility of their own data empowers patients who have no hope or comparators.

KF has already done a great deal of work in liberating metadata. Weight, for instance, is not identifiable, but is hidden by dbGaP. The KF DRC was able to show that other NIH data repositories, such as the Genomic Data Commons, showed weight and a number of other metadata terms without requiring a log-in. By examining the differential between what metadata KF was initially exposing and what was available elsewhere, KF was able to work with their PO to update the KF system to align with the GDC. Similarly, KF noted that institutional certificates often don’t match the consent forms signed by patients. There are four sequencing centers where the many KF contributors send samples. The genomic data flows through the sequencing centers, while phenotypic and clinical data comes from the original contributing clinicians, and in many cases, KF had to reach back to the clinicians to get access to the raw data. Some of the early consent agreements had prohibitions on multi-cohort use, however when KF went back and revisited data use agreements they were able to remove the prohibitions. By systematically reviewing documents, and working with their program officer, KF was able to retroactively change several institutional certificates to allow more metadata to be made public, in better accordance with the agreements the patients actually signed. As of our meeting, the default agreement at KF with their X01 grantees is that all of their phenotypes are public.

Reusability

Kids First staff discussed the idea of reusability in several ways. First, they addressed the idea of licensing. As of our meeting, there are no NIH-wide standards or guidelines for licensing. While they did not think that the NIH should choose a single license everyone *must* use, they suggested that it would be useful to have a framework. For instance, the Common Fund might issue guidelines that some small set of pre-existing licensing options are compatible with CF principles, so that researchers can choose among them.

Generally speaking, and in keeping with their goals, Kids First data is very reusable, however their infrastructure is less so. Data at KF is processed in a reproducible way, using documented workflows. Between GitHub hosted documentation and CWL workflows, an interested and sophisticated user can find all the internal metadata such as program parameters and versions of software. However, these parameters are not exposed by default to the end user. This decision is still an open question at KF, as they try to strike a balance between transparency and confusion. As previously discussed, most of their users are already overwhelmed with just their own data analysis, so adding all of these parameters to the default view would likely make the data *less* accessible overall. KF also noted that exposing their pipelines by default may also imply a level of authority they are uncomfortable with “My foremost concern is that somehow we are seen as defining what is best practice.” They want to be sure they have a light touch and encourage community building around what pipelines they are running.

In terms of the reusability of their infrastructure, various components ranged from ‘not at all’ to ‘completely reusable’. Their metadata model as a whole is not reusable. As discussed previously, they took a very practical approach to metadata as necessitated by operational needs, so their model is tailored to their data. However, FHIR, the standard they’ve adopted for patient variables, is very reusable, as is the KF data portal site. They also pointed out that their authentication system is universalizable already, it’s just burdensome to set up. However, the biggest challenge is contracts not technology. If NIH built the authentication system, there would be no need for an ATO, but right now everyone has to build their own system, and that requires FISMA moderate and an ATO, which are expensive and time consuming sign-offs.

Authentication/Authorization

Covered in [Appendix J](#).

Data Dashboards

Kids First already has all of their data in the cloud, and has a system in place for ingest, so they were not initially interested in the Data Dashboard as proposed by the CFDE. However, they imagine that a dashboard with a few added features would be very useful. In particular, one that manages patient uploaded data, or one that tracks usage or processing type statistics such as:

- What processing stage data is in
 - e.g. “harmonized”, “in the process of being harmonized”
- How the data is being used by others
 - E.g. “downloaded”, “used in virtual study”
- How many people are using my child's data?

CF Data Portal

In keeping with their mission to get more data, to more researchers, faster, Kids First is very interested in a CFDE portal to connect users to new datasets. The illnesses faced by the children that come to CHOP and KF are mostly rare diseases, and a single hospital might only see one case every few years. This means that research is limited by numbers of samples, and researchers want to interconnect with more and bigger datasets and cohorts. There are also 'failure of development' diseases, where researchers want to interact with normal comparators as well as make novel connections to data that might be relevant, but that exists in another disease space. No single institution or single dataset has enough information, so there's a prebuilt requirement to collaborate in the pediatric space to both search across platforms and to be able to integrate the results in order to compete with more common diseases (prostate, breast, etc.) However, as previously noted, KF doesn't think investing in one overarching metadata model entirely solve this problem. Instead, they picture something more like portable queries. Rather than each DCC changing to a shared overall model, DCCs could provide a query guide to their internal model, and engage with each other to map queries to other DCCs. This might be done through a single plugin shared across DCCs that is connected to each underlying database in such a way that the user can query any site on a given set of search terms and get back the sensible response for that dataset. Or it could be done by creating a utility that takes a given user query at one DCC and translates it into the corresponding queries for other DCCs.

The Kids First DRC already has a simple to use, and yet very sophisticated portal that allows users to interactively explore a wide range of publicly available metadata. The portal requires a login, however, joining is simple. Users have the option to create a stand-alone login, or to simply connect an existing Google or Facebook account. Upon logging in, a user can:

- Access their previous queries or previously viewed files
- Apply controlled access credentials to their account
- Explore community data plots such as 'member research interests'
- View pathology and histology images
- Push cohorts to Cavatica for analysis
- Explore the KF datasets using a number of metadata terms by:
 - Entering queries into a boolean search
 - Filtering by a wide range of clinical and technical variables
 - Clicking on data points in cBioPortal plots that dynamically respond to selections

As was discussed in FAIR: Findability and Interoperability, Kids First also has a number of plans to make their portal more specific to their own datasets. They hope to eventually add functionality to search by events and other clinical variables, as well as add free-text search. They also plan to add a notification system that notifies users if a saved query result changes over time.

Data Platform

Data Hosting

As previously discussed, the KF DRC was funded several years after data collection began, and the primary goal of the KF DRC was to summarize, harmonize and collect partially processed data sent to them by the sequencing centers. However, the Sequence Read Archive (SRA) stopped taking on high volume datasets, and the KF had to take on hosting, storing, and distributing the raw data for the program. Essentially they began to act as the SRA for Kids First data. Initially this was difficult, as the DRC was not funded for data storage or cloud computing, however CHOP contributed in-kind funding as part of the grant mechanism which helped support this. As of our meeting, the KF DRC still does not have NIH funding for cloud resources.

At the time of our meeting, Kids First was hosting about three PB of data, about one petabyte of which is harmonized. With four active sequencing centers as well as other sources of data, they are growing at a pace of about 6,000-10,000 whole genomes (about 1800 terabytes) per year. They project that this figure will soon increase exponentially, as they expect that within two years, every patient at CHOP will have a WGS as a matter of routine. “Right now, it’s not clear on how we’ll keep up on data”. KF told us that in the short term, liability will require that they keep all data, so it can be analyzed against new variants, and tested in new ways. They warned that at some, not too distant, point getting sequencing might transition to be more like getting a blood draw, at which point it may be cheaper to keep the sequencing summary and re-sequence as needed. However, this will likely not be true for specific samples, like lesions or microbiomes, which will be much more precious.

Data Analysis

Cavatica, developed by Seven Bridges, is the Kids First Cloud computing environment. Users choose cohorts via the Kids First portal and push those UIDs to Cavatica. Cavatica then pulls the correct data. Since this operation only pushes data IDs rather than moving or copying data, users can begin their analysis immediately. New users get \$100 worth of credits when they register their Cavatica account, and Kids First subsidizes continued compute and storage for their researchers. Users can also directly link their own AWS buckets to Cavatica and do combined analyses. Kids First would be interested in APIs or other solutions that would allow their users to link to other data sources as well. The Cavatica workspace ensures that a users analysis is reproducible by recording details on tools, as well as input and output parameters, and advanced users can use the API or ‘Docker pull’ to access extra details about the pipeline.

Training

The Kids First Data Resource Center serves a wide variety of users. Their portal gets about 2000 users per month, and their Cavatica space has somewhat less than 500 total accounts,

about 200 of which are active, repeat users. The majority of their users have role of research, but there are also families, patients, patient advocates and clinical users. A number of these researchers are Kids First awardees, either from X01 or R03 grants.

At the time of our meeting, Kids First had not run any training sessions, however they were considering starting to do monthly office hours on site. They frequently go on 'listening tours' to all their awardees to survey their needs, and have a standard presentation for their X01 grantees that explains the overall process of managing data. KF also have attended several conferences for cancer, as well as more specialized disciplines, and recently one of the KF staff gave a Gordon Conference talk. KF also has a Twitter, Facebook page, and mailing list, however these are aimed at the general public, not at researchers

KF discussed two user groups with complementary training needs. They have clinicians and clinical researchers who have little or no bioinformatics training and struggle with computational issues, but there are also a number of genomic researchers who are accomplished bioinformaticians, but don't know how to deal with the vast amount of clinical data tied to each genome. Kids First reported that their support burden is mostly from the first group. Frequently, KF is being asked by local people to perform analyses for them, and noted there are a lot of constraints on resources and insufficient bioinformatics support. These circumstances cause Kids First to wonder what role they should be playing; are they just a data provider or are they also a bioinformatics core? And just how many of their resources should be devoted to basic bioinformatics training? The vast majority of their users require basic computational training before they can begin to interact with the datasets, and so it would technically support their mission to increase data use. However, supporting basic bioinformatics training takes away from support for more experienced users, and from many other activities that are much more obviously Kids First's responsibility.

KF is most interested in training the clinical researchers. The NIH funds approximately ten new X01 KF awards each year which are handled by the KF DRC. All of these awards have different timelines, readiness to begin the project, and bioinformatic aptitude. Once these researchers receive their files from the sequencing facility, they have a six month embargo period before the data is released to the public. However, in most cases, these awardees don't know how to get started, and need a lot of help. Very few of them finish, or even in some cases start, their analysis before the embargo period is over. This is frustrating for the researchers, but also for the families of the patients involved with the study, who are expecting fast results and hoping for miracles.

There are three main reasons that KF has not begun work on a training program: time, money and expertise. First, they haven't had time. Although Kids First is a very mature DRC in terms of their portal and volume of data, they had only been in operation for a little under two years when we visited. Getting all of their infrastructure to its present level of maturity took precedence over developing training. Building a training program also takes money. While KF has some backing from CHOP as well as several Foundation partners, most of this money has gone into

supporting infrastructure needs that are still unsupported by the NIH. Finally, training is simply not in their expertise. In particular, they worried that they don't know how to capture the messy underside of data analysis, which is essential when their audience is going to use their new skills on patients. "How do you take data and new abilities to do science and give it to more people?"

The Kids First DRC staff were very excited about several of our training suggestions, and suggested that with the right proposal, they may be able to secure additional funding from non-NIH sources to supplement their training goals. In particular, patient advocates are usually present at KF events, and have heard firsthand from researchers about the training they lack. Since patients and their families are interested in getting their data used by more researchers, they would likely be amenable to funding those training programs.

Outcomes

There were a number of outcomes, both technological and social, that Kids First would like to see from the Common Fund Data Ecosystem:

CFDE Targets

- Help with building a KF specific training portfolio:
 - A webinar with curated questions about how a type of data is best analyzed, access questions, and other higher level issues. This would allow users to get help, serve as a resource for new users, and help to find new use cases to pursue
 - A two-day clinician-focused, hands-on intro to the Kids First portal, and how to explore already-done analyses. CFDE would provide initial work to create materials, and eventually hand off training program to KF when sustainable
 - A series of clinician-focused workshops + remote touch-ins to help newly awarded X01s go from data to analysis. This would help X01 awardees get their data analyzed inside their 6 month embargo period.
 - A two-day to 1-week hands-on portal+workflow platform for bioinformatics, and how to work in the cloud. This might be better suited as a CFDE product than an individual DCC project
- A 'Proposal in a Box' for training that they could use to attract additional funding from foundations and sponsors: outlines the purpose and costs to train X clinicians
- Unified authorization: A CF wide authorization solution
- Reduce risk of sharing data
 - Risk can arise when clinical data is selected for research projects
 - And has protected patient health information
- Inform KF on what other datasets "their" users are accessing
 - "If you like this dataset then you will also like ..."

- For instance, it would be interesting to know if GTEx is more or less useful than TopMed
- Related: analysis on data use as categorized by users
 - e.g. data use trends within dbGAP
- Research on how to “do a commons”:
 - Fund discovery of what a data ecosystem *is* “The social science of data science”. A data commons is analogous to the concept of a commons in the economic sense and it would be useful to have a model of the benefits and requirements of a science commons, for e.g. the total cost of operations, or how to build, or exploration into narratives of success of a science commons.
 - A DCC is the “perfect” subject for an experiment on the value of a science commons
- Accelerate discovery and improved care for children.
 - More data, into the hands of more researchers, faster
 - Steer more children to appropriate cancer treatment
 - Involve children in clinical trials when necessary
- Basic bioinformatics training: Workshops for clinicians who have access to tissue and can get sequence, but lack bioinformatics exposure.
- Facilitate collaborative or “crowd sourced” publications
 - Crowd sourced analysis
 - “Hackathon-y”
- Create local champions who have the mandate of chasing down answers to questions

Major Use Cases:

- Automated data ingest:
 - A user take their FASTQ or BAM file and push a button to put it through KF pipeline. It’s automatically harmonized and comparable with their KF cohort.
- Shared workspaces:
 - A user can build a cohort and start an analysis on Cavatica and share that workspace with another user. Have to support dbGaP _and_ other auth/access. E.g. if you want to share data from dbGaP with others, you need to make sure everyone has dbGaP access. But then there are PI-owners who want to have direct ability to share.
- Patient driven access
 - A patient can upload their own data to KF, have it automatically harmonized, and be able to track how and whether their data is used
- Anti use case: allow naive data recombination between DCCs
 - Allowing users to combine any data creates the possibility for data misinterpretation. There are perils of joining data where the details are concealed from the users, or the users are not savvy enough to recognize the problems.

Game Changers

During our discussions with Kids First we brainstormed several ways the CFDE might address the problems posed by Kids First in a more holistic way across the entire Common Fund. This is a targeted set of goals that were not originally included in our proposal, however, if the concerns from Kids First are shared by most DCCs, we could positively impact the entire ecosystem and make important new advances in data access and analysis for the NIH research community. These targets are labeled as "game-changers" because by accomplishing these innovations, the CFDE has the potential to dramatically improve the access and usability of an organized set of resources hosted at the Common Fund DCCs.

Building a DCC community. Kids First pointed out that they have few opportunities to collaborate with other DCCs or to share ideas. In other words, there is currently no simple way for the people who are the most expert in how to successfully operate a Common Fund Data Coordinating Centers to talk to one another. Unsiloing the data is not enough. We need to work to unsilo the staffs of DCCs, and provide avenues for discussion and collaboration. The CFDE could help by

- Defining bottom up efforts that can be connected
- Creating opportunities for collaborative proposals between DCCs
- “Bubbling up” the knowledge and the solutions present in all the DCCs
 - Providing opportunities for cross training such as ‘how to commons for Program Officers’ so other DCCs can benefit from KF’s work in liberating metadata
- Hosting DCC conferences
- Create DCC mentoring space so solved problems can be shared
 - Based on maturity of the DCC and the specific needs

Pipeline tracking. Even with CWL and parameter tracking in Cavatica, Kids First is still plagued by the difficulties of tracking pipelines by metadata, and what makes a pipeline sufficiently ‘different’. When the near infinite number of tool combinations to process data is combined with the ever evolving nature of Kids First metadata, it is difficult to apply GUIDs or other standard ways of thinking. This could potentially be solved with an algorithm that looks at Cavatica and Terra execution metadata and determines what pipelines they were run with, and rather than report back to the user whether the pipelines are ‘the same’, which they almost never will be, reports back whether the pipelines are likely to be incompatible, based on how similar the pipelines are.

A cross-DCC query system: This idea is a variation on the CFDE shared portal proposal. Kids First imagines that most DCCs will have metadata models like their own: idiosyncratic and highly specialized. However they still want to be able to query data across sites. A solution to this might be that rather than each DCC changing to a shared overall model, DCCs could provide a query guide to their internal model, and engage with each other to map queries across DCCs. This might be done through a single plugin shared across DCCs that is connected to

each underlying database in such a way that the user can query any site on a given set of search terms and get back the sensible response for that dataset. Or by creating a utility that takes a given user query at one DCC and translates it into the corresponding queries for other DCCs.

Agenda

Day 1

9-9:30am Introductions

Short introductions from engagement team members and attending DCC members. The overarching goal for the engagement team is to collect value and process data about the DCC. Values data will include things like: mission, vision, goals, stakeholders, and challenges. Process data includes: data-types and formats maintained, tools and resources owned by the DCC that they would like to have broader use, points of contact for follow up on technical resources, etc.

9:30-10am DCC overview

Short overview of DCC. Can be formal or informal, choose 1-5 topics to cover. Suggested topics: What is your vision for your organization? What big problems are you trying to solve? What are your big goals for the next year? Who do you see as your most important users/stakeholders? What project(s) is currently taking up the bulk of your effort/time? What areas of your organization are you putting the most resources into? What is the rough composition of your user base in terms of discipline? Do you have any challenges that are blocking implementation of your current goals? What skill set would you like to add to your project? How do you engage with your users? What kind of sustainability issues are you confronting? Can you currently do combined analyses with external datasets?

10am-Noon Goals Assessment

An exercise to get an idea of what types of things are important, what types of things are challenges, what do you dedicate your time/resources towards, and what types of things are not current priorities. Given a list of common goals provided by the engagement team, plus any additional goals the DCC would like to add, DCC members will prioritize goals into both timescale: "Solved/Finished", "Current-Input wanted", "Current-Handled", "Future-planned", "Future-unplanned", "NA to our org" and for desirability: "Critical", "Nice to have", "Neutral", "Unnecessary", and "NA to our org". The engagement team will work to understand the reasons for prioritization, but will not actively participate in making or guiding decisions.

Goal List

- | | |
|--|------------------------------|
| ● Increase end user engagement X% over Y years | ● Metadata harmonized with |
| ● Move data to cloud | ● Metadata harmonized across |
| ● Metadata harmonized within DCC | Common Fund |

- Implement new service/pipeline
- Increase number of visitors to your site
- CF Data Portal
- Single Sign On
- Pre-filtered/harmonized data conglomerations
- A dashboard for monitoring data in cloud
- User-led training for end users (i.e. written tutorials)
- Webinars, MOOCs, or similar outreach/trainings for end users
- In-person, instructor led trainings for end users
- A NIH cloud playbook
- Full Stacks access
- Developing a data management plan
- Increased FAIRness
- Governance role in CFDE

Lunch

1 - 3:30pm Open discussion (with breaks)

Using the results of the morning exercise and a collaborative format, iteratively discuss goals, blockers, etc., such that the DCC agrees that the engagement team can accurately describe their answers, motivations and goals. Topics don't need to be covered in order, we'd just like to touch on these types of questions.

Topics:

Infrastructure:

- Do you intend to host data on a cloud service?
- Have you already started using cloud hosting? If yes:
 - Approximately how much of your data have you uploaded? How long did that take? How are you tracking progress?
 - What challenges have you faced?
 - How have you dealt with those challenges?
- What potential future problems with cloud hosting are you watching for?
- Does your org use eRA Commons IDs? Do the IDs meet your sign on needs?
 - If yes, did you have/are you having challenges implementing them?
 - If no, what do you use? What advantages does your system provide your org?

Use cases

- What is the rough composition of your user base in terms of discipline?
- What if any, use cases do you have documented? Undocumented?
- What things do people currently love to do with your data?
- What things would people love to do with your data, but currently can't (or can't easily)?
- What pipelines are best suited to your data types?
- What are the challenges associated with those desired uses?
- What other kinds of users would you want to attract to your data?

Review of metadata:

- What's metadata is important for your org? For your users?

- Do all of your datasets have approximately the same metadata? Or do you have many levels of completeness?
- Do you have any data already linked to outside resources?
 - Did you find the linking process easy? Challenging? Why?
- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- What automated systems do you currently have for obtaining metadata and raw data?

Training:

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?
- What challenges keep you from offering the training you'd like?

Policies:

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR:

- Has your org done any self assessments or outside assessments for FAIRness?
- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

Other:

- What search terms would make your data stand out in a shared DC search engine?
- Does your org have any dream initiatives that could be realized with extra resources? What resources would you need?
- If you had free access to a Google Engineer for a month, what project would you give them?
- Any other topics/questions the DCC would like to cover

9-10am Review of goals and CFC involvement

A quick review of what topics are priorities for the DCC with suggestions from engagement team on how we can help.

10-noon Open Discussion

DCC reflection on suggestions, open discussion to find shared solutions.

Lunch

1-2pm Thoroughness checking

Touch on any questions not covered previously, ensure we have:

- Action Items for us, and rough timelines for getting back to DCC on them
- Tools / resources the DCC thinks might be useful for the overall project

- Points of contact “Who is the best point of contact for your metadata schemas, your use cases, the survey of all your data types?”
- Who would like to be added to our governance mailing list?
 - Or contact info/instructions on how to get that information offline.

Appendix E - HMP Interview

Wednesday July 17, 2019

1pm to 4pm

Meeting Logistics

We held a meeting with the HMP via Zoom on Wednesday July 17, 2019 for about three hours. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the meeting. Representatives in attendance from the CFDE were: Amanda Charbonneau (UCD), Ben Carr (BioTeam) and Titus Brown (UCD). We spoke with several members of the HPM team: Michelle Giglio, Heather Creasy, Victor Felix, Kemi Ifeonu, and Jonathan Crabtree. The PI, Owen White, joined us for the last half an hour or so. Anup Mahurkar was not available so he held a one on one conference with Amanda to ensure his input made it into the report.

The engagement team began by reviewing their goals for the meeting. These goals include learning about the structure and goals of the HMP, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for their group. In turn, the HMP gave us an overview of their organization, and talked to us about the varying goals and mandates of their two HMP grants. They also gave us a great deal of insight on how DCCs navigate the end of their lifecycle.

Meeting Logistics

HMP Overview

Harmonized data

Cross cutting metadata models

Self-governed metadata standards

FAIR Assessments

Findability

Interoperability

Accessibility

Reusability

Authentication/Authorization

Data Dashboards

CF Data Portal

Data Platform

Data Hosting

Data Analysis

Training

Outcomes

CFDE Targets

Major Use Cases:

Game Changers

Agenda

HMP Overview

The Human Microbiome Project (HMP) had two phases. Phase 1 (HMP1), began in 2007 with the goal of building a shared community resource: a representation of the healthy human baseline normal microbiome. They recruited over 300 people to the project and sampled 15 body sites from men and 18 from women. Most of these were analyzed using 16s sequencing, and about 2200 of the samples were used for whole genome sequencing. The HMP Data Analysis and Coordinating Center (DACC) worked with four sequencing centers, the Broad Institute, the Baylor College of Medicine, Washington University School of Medicine, and the J. Craig Venter Institute, to create a single, integrated data resource. They also worked with some researchers to do 'demonstration projects', in which twelve groups were funded to examine correlations between changes in the microbiome communities and a disease. A subset of these projects received additional funding for follow up studies. Finally, the HMP was involved in an effort to find and sequence 3000 bacterial reference strains, which eventually became part of the NIAID BEI resource database. Funding for phase 1 of the HMP ended in 2012.

Phase 2 of the HMP had a different focus, with some new sequencing centers, and a different overall organizational principle. Beginning in 2013, the DCC shifted to serving as only a Data Coordinating Center -- with no analysis support -- for the 'Integrative HMP', variously referred to as iHMP or HPM2. The Phase 2 goal was to characterize the microbiomes of three study groups: pre-term birth (both of the pregnant women and the child), Inflammatory Bowel Disease (IBD) patients, and Type II diabetes patients. These are all longitudinal studies that aim to study how the microbiome changes during progression of each condition. Data collection was split between sequencing centers by project. Virginia Commonwealth University collected the pregnancy and pre-term birth data; Stanford University and Jackson Laboratory collaborated to build the onset of Type 2 Diabetes dataset; and Harvard Medical School and The Broad Institute worked together on characterizing the onset of IBD. Taken together, these changes resulted in

a dataset that has much deeper sequencing, but a narrower scope, and overall is a much less integrated dataset compared to phase 1.

Harmonized data

Surprisingly, the harmonization effort at HMP is tightly linked to the terminology in their originating grants. Although the DCC was a ‘center’ for both HMP1 and HMP2 grants, the mandates of the RFAs, and by extension, the overall structure of the consortia, were very different. In HMP1, the team performed *data analysis*, that is, they were charged with ensuring that the data from the four sequencing centers was integrated and interoperable. The DACC captured all of the protocols and did a lot of cross-validation to ensure data quality. They created a single set of metadata terms that mapped to a single ontology. They also had the authority to impose those standards on the data generation centers. As such, all of the data from this phase acts as a single, harmonized dataset. It is comprised of many smaller datasets, but it can basically be treated as one enormous study: all the data was built with common analysis pipelines, using common outputs.

In HMP2, the coordinating team was designated a *Data Coordinating Center*, and that dramatically changed both their mandate and their ability to make a cohesive dataset. As a coordinating center, they were mostly charged with building a portal and making sure the data was accessible, but had no funding or power to make the data interoperable. HMP2 was essentially a loose federation of related sequencing efforts: every center used their favorite pipelines and their own ontologies. In order to upload their data to the DCC, they all did agree on a shared format for data, but that is where the similarities end. At the end of the project, it was effectively impossible to combine data across the various parts of HMP2, or to compare anything from HMP2 to HMP1.

Although their funding has ended, the HMP DCC staff have been working to increase harmonization between the datasets, and to make all of the data from the two projects at least partially integratable on the portal. However, the center does not have access to the protected metadata associated with either phase, and so has no ability to harmonize it. As such, although the HMP2 studies should all be easily interoperable because they used many tissues in common, they are not.

Cross cutting metadata models

The data structure used by the HMP, for both phases, should make them easily integratable with the C2M2. HMP data is organized within the Open Science Data Framework (OSDF) and uses a RESTful API for storing, retrieving and modifying JSON documents. Their JSON-Schema imposes structure on the data, and allows for validation of completeness or correctness of data by defining required properties, data types and formats using controlled ontologies. The HMP schemas are freely available at <https://github.com/ihmpdcc/osdf-schemas>.

Self-governed metadata standards

The HMP group noted that during the course of their project they were able to work closely with the data generators to develop a preliminary set of metadata, but data collected still required further curation. For example, during the first phase of the HMP only four centers were generating sequence information from a single study, and terms such as body site still varied based on the center performing the sequencing. During the first round of the HMP several demonstration projects were supported, and while subject variables were collected (e.g., race, weight, smoking status), none of the data generators used common controlled vocabularies to describe those variables. In the second half of the project, the iHMP data generators also gathered subject variables in a free form manner.

FAIR Assessments

Findability and Accessibility

At the time of our meeting, processed data from both phases of the HMP was easily findable and accessible through the portal. Users can use faceted search to find a particular subset of data, and are provided with a manifest and instructions for downloading their selections. However, in the absence of funding, the portal will be shut down in mid 2020, and data from the HMP2 will be functionally impossible to find or access.

Interoperability and Reusability

Although FAIR terminology didn't yet exist during HMP1, the consortium was working according to FAIR principles. They were very focused on metadata consistency and completeness, and have detailed SOPs for their workflows. However, they told us that the average user could not easily integrate their dataset with the HMP, or reuse the data: "There's no way in the world that this data could be integrated with a dataset that comes along from a new project."

This is largely due to the age of their data. For example, HMP1 did their gene calls using the version of blasttools that was cutting edge in the late 2000's. Their process and pipeline is well documented, and includes relevant variables such as the version of blasttools they used. However, a user couldn't replicate the HMP results unless they could also access the same blast database. Perhaps more importantly, it is not clear that you could recreate the results even if the HMP could provide a copy of the old database. The pipeline is outdated at this point, and has not been containerized. So while the pipeline is reproducible in principle, it would not be possible to reproduce the results because databases that are used for assigning gene calls have changed. From their website page explaining the pipelines:

Please be aware that HMP1 funding ended in 2012, and therefore some of these resources may have changed, moved or been discontinued. This list is no longer regularly maintained.

The HMP has already experienced problems with users trying to re-use their data. Since their funding has ended, members of their consortium have wanted to write publications and still required support from the DCC for data wrangling, as the data essentially needs to be entirely reprocessed for each user. One solution might be to fund the HMP to update the pipeline and reprocess all the data, so that the data is more directly comparable to modern analyses.

Authentication/Authorization

Covered in [Appendix J](#).

Data Dashboards

The HMP currently has no funding to move their data to the cloud, and so cannot use a data dashboard.

CF Data Portal

The HMP is interested in having their data be searchable through a centralized portal, however the CFDE will likely face the same challenges they did with harmonizing metadata and the inaccessibility of protected metadata terms.

The current HMP portal offers a faceted query interface, and users can pull down a manifest file with standard fields. However, the HMP was skeptical that it could be re-used by other DCCs as it is a challenging piece of software, with many custom integrations. Porting the technology could probably be done, but the recipient would need assistance from the HMP. They also noted that many new technologies have been created since they built their portal framework, and newer technology might dramatically simplify shared data search.

Data Platform

Data Hosting

Ongoing data management is a problem at HMP. Processed data from the first phase of the HMP is stored locally, in the SRA, and in the cloud: several years ago, the HMP staff participated in a program where they received free, indefinite storage space from Amazon to host that data. However, phase 2 HMP processed data storage is only local, and currently on servers paid for by funds from initial HMP1 grant. The HMP does not store, or have access to the raw data or protected metadata from either phase of the project, which is stored at dbGaP.

The local servers that hold copies of both datasets were purchased on the original grant, so they are outdated and UMD is retiring them early next year. At the time of our meeting, the HMP

had no funding to move the HMP2 data to the cloud, and did not believe the data could be added to their current, free, Amazon allocation. They also do not have funding to purchase new local machines, and data cannot be moved to the SRA at NCBI, because they are no longer accepting large datasets. Due to NIH data retention policies, the HMP must keep copies of the data, however there is no requirement for the data to be easily accessible. In the absence of further funding to move the HMP2 data to the cloud, their plan is to shut down the HMP portal in early 2020, and archive all of their local data on tape drives for the remainder of their required data storage time.

Altogether, this means that the HMP1 data will likely remain available to the public for some time: it was created sufficiently long ago that it was uploaded to the SRA before their prohibition on consortia datasets, and it also benefits from an old Amazon initiative to host scientific datasets for free. However, the processed HMP2 data, which was publicly released only in the last few months, will be unavailable starting in early 2020.

If the HMP were to receive funding to migrate their data to the cloud, they indicated that they would need that funding very soon: the funding would need to be received by the end of December 2019, or January 2020 at the latest. This is because they need to move 10s of terabytes of data to the cloud or to tapes, and need to need to start that process before their servers are retired. The HMP estimated that they need to start data migration in January to finish in time, whether it be to the cloud or to tape archives. The HMP estimates it will take a month to transfer data to the cloud, but notes that depends on what we mean by 'data'. The data that are actively in use by the portal (the final calls in the form of FASTA files) is around 42Tb, but this is only about 20-25% of the overall data stored by HMP. The other 75% is intermediate files, and it's not obvious whether those need to be moved too. The current best practices for bioinformatics pipelines suggests that if a user has the raw files and a reproducible protocol, then the intermediate files are not necessary. However, given the age of the HMP project, those guidelines don't necessarily apply, as many of their pipelines cannot be replicated. Given that, perhaps 'data' should mean *everything*, including intermediate files. If that is true, uploading the data to the cloud or archive will take closer to five months.

Data Analysis

HMP does not have an analysis platform, and has no current plans or funding to support one. However, they indicated that given additional funding, they would be interested in having full stack access for their users.

Training

The HMP told us that there was little integrated outreach across HMP, however they offered several training opportunities. The finalized pipelines from HMP1, for instance, are available as walkthroughs - step by step tutorials - for users to help them perform the same analysis. In 2017, they hosted a one-off cloud workshop for fifty people, and they also host a standard

metagenomics workshop, and a regular microbiome analysis workshop. The HMP has a helpdesk support burden that continues today. Their portal has 652 registered users, and approximately 8000 active users per month, who need regular support.

Outcomes

The CFDE discussed several challenges, and potential solutions that could be implemented by the Common Fund Data Ecosystem:

CFDE Targets

- Cloud funding:
 - Even if the HMP is funded move their data to the cloud, it's not clear who will maintain it or pay continued storage and access costs.
 - The HMP group suggested that the Common fund or CFDE will have to pay for ALL cloud costs: transfer, download, upkeep.
 - The HMP *could* set up a bucket with user pay, but it would break the implicit system of taxpayer funded research: researchers expect NIH data to be free. It is technologically easy to make the users pay for data egress, but socially very difficult or impossible.
 - The HMP noted that when universities set up a cloud hosting contract, one thing they frequently negotiate is no egress charges. They suggest that the NIH should set up STRIDES to not have egress charges, or be prepared to pay those charges
- Data continuity/upkeep/maintenance
 - Among the challenges that the HMP noted concerning end of life cycle are the costs involved in keeping an online presence. Even a static website needs periodic maintenance and security updates for the site; this is significant additional effort for the portal, databases and other features.
 - They also noted that 'data integration' is an ongoing process, that has to be re-implemented every time a new DCC is added.
- Training/outreach continuity/upkeep/maintenance
 - For datasets to be reused, new researchers need opportunities for outreach and training on those datasets. This points to a need for funding well past the 10 year DCC lifetime to keep datasets from falling into disuse.
- Funding clarification
 - The HMP is past their 10 year funding limit, and it is unclear whether and how new funds might be allocated to them for moving data to the cloud or for participating in other CFDE activities. No one is sure what the nuances of the law are, and an official ruling from the NIH would make it easier for end of life cycle DCCs to determine if it is worth their remaining time to work with the CFDE.

Game Changers

During our discussions with the HMP group we heard about how meeting the requirements specified in the RFA for a DCC can sometimes have the unintended downstream consequence of making data less FAIR.

Shifting the mandate of DCCs The biggest problem at HMP, and the one that they are still working on even after the grant is over, is lack of harmonization, and the HMP group told us that the problem stemmed from their shift from an analysis center to data repository. As a repository in HMP2, they could not ask for more people or computational resources, nor could they force the sequencing centers to align. The HMP DCC had no way to incentivize people to use the same pipelines or adopt the same standards.

The HMP suggested that the single best investment that the NIH could make would be to routinely fund DCCs to also perform analysis roles. The analysis component is critical to keeping everyone on the same page, and would ensure data is collected and analyzed in a sustainable and interoperable way. Supplemental funding after a DCC is established is not as helpful as doing it right the first time, because enforcing changes on sequencing centers in the middle of a project is both technically and socially difficult.

Cross DCC Mentoring The HMP told us that they already are involved in some informal mentoring of other DCCs, and mentoring is an effective way for young programs to benefit from the experiences of older ones. This recommendation fits well with other suggestions for the CFDE to create a community of DCCs and increase opportunities for cross-pollination of ideas.

Agenda for Three Hour Session

10 Minutes Introduction from CFDE

Short introductions from engagement team members and attending HMP members. The overarching goal for the engagement team is to collect value and process data about HMP. Values data will include things like: mission, vision, goals, stakeholders, and challenges. Process data includes: data-types and formats maintained, tools and resources owned by HMP that they would like to have broader use, points of contact for follow up on technical resources, etc.

20-30 minutes HMP overview

Short overview of HMP. Can be formal or informal, choose 1-5 topics to cover. Suggested topics: What is your vision for your organization? What big problems are you trying to solve? What are your big goals for the next year? Who do you see as your most important users/stakeholders? What project(s) is currently taking up the bulk of your effort/time? What areas of your organization are you putting the most resources into? What is the rough

composition of your user base in terms of discipline? Do you have any challenges that are blocking implementation of your current goals? What skill set would you like to add to your project? How do you engage with your users? What kind of sustainability issues are you confronting? Can you currently do combined analyses with external datasets?

30 Review Goals Assessment

Please use the FunRetro board provided to sort this apriori set of goals that we expect DCCs might have **at least the day before** our meeting time. For each goal sticky in the ToSort column, drag it to the column that best describes the current state/thinking/goals of your organization. Then, leave a comment on each that specifics how desirable that goal is using these terms: “Critical”, “Nice to have”, “Neutral”, “Unnecessary”, and “NA to our org”. Please add as many other comments as you wish. If your organization has a goal that is not listed, please click the ‘+’ at the top of a column to add a new sticky.

1 - 1.5 hours Open discussion

Using the results of the goals assessment and a collaborative format, iteratively discuss goals, blockers, etc., such that the engagement team can accurately describe HMPs answers, motivations and goals. Topics don’t need to be covered in order, we’d just like to touch on these types of questions.

Topics:

Infrastructure:

- Do you intend to host data on a cloud service?
- Have you already started using cloud hosting? If yes:
 - Approximately how much of your data have you uploaded? How long did that take? How are you tracking progress?
 - What challenges have you faced?
 - How have you dealt with those challenges?
- What potential future problems with cloud hosting are you watching for?
- Does your org use eRA Commons IDs? Do the IDs meet your sign on needs?
 - If yes, did you have/are you having challenges implementing them?
 - If no, what do you use? What advantages does your system provide your org?

Use cases

- What is the rough composition of your user base in terms of discipline?
- What if any, use cases do you have documented? Undocumented?
- What things do people currently love to do with your data?
- What things would people love to do with your data, but currently can’t (or can’t easily)?
- What pipelines are best suited to your data types?
- What are the challenges associated with those desired uses?
- What other kinds of users would you want to attract to your data?

Review of metadata:

- What's metadata is important for your org? For your users?
- Do all of your datasets have approximately the same metadata? Or do you have many levels of completeness?

- Do you have any data already linked to outside resources?
 - Did you find the linking process easy? Challenging? Why?
- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- What automated systems do you currently have for obtaining metadata and raw data?

Training:

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?
- What challenges keep you from offering the training you'd like?

Policies:

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR:

- Has your org done any self assessments or outside assessments for FAIRness?
- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

Other:

- What search terms would make your data stand out in a shared DC search engine?
- Does your org have any dream initiatives that could be realized with extra resources?
What resources would you need?
- Any other topics/questions the HMP would like to cover

30 minutes Review of goals and CFC involvement

A quick review of what topics are priorities for the HMP with suggestions from engagement team on how we can help.

Appendix F - LINCS Interview

Wednesday July 10, 2019

2pm to 5pm

Meeting Logistics

We held a meeting with LINCS via Zoom on Wednesday July 10, 2019 for about three hours. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the meeting. Representatives in attendance from the CFDE were: Amanda Charbonneau (UCD), and Titus Brown (UCD). We spoke with three LINCS PIs: Mario Medvedovic from University of Cincinnati; Stephan Schürer from University of Miami; and Avi Ma'ayan from the Icahn School of Medicine at Mount Sinai (ISMMS). Also on the call were Jarek Meller from University of Cincinnati, who developed the LINCS proteomics portal; and three members from the ISMMS team: Alexandra Keenan is an MD/PhD student, who has been working with LINCS data throughout her PhD and presented on some of her work at our meeting, Daniel Clarke, a bioinformatician, and Sherry Jenkins, who is the project manager and the outreach lead of the center.

The engagement team began by reviewing their goals for the meeting. These goals include learning about the structure and goals of LINCS, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for their group. In turn, LINCS provided us with an overview of their data generation pipelines, datasets, and breadth of analysis tools. They also gave us a great deal of insight on how DCCs navigate their final year of funding.

Meeting Logistics

LINCS Overview

Harmonized data

Self-governed metadata standards

FAIR Assessments

Findability

Accessibility

Reusability

Authentication/Authorization

Data Dashboards

CF Data Portal

Data Platform

Data Hosting

DCIC

DSGCs

Data Analysis

Training

Outcomes

CFDE Targets

Game Changers

Agenda

LINCS Overview

The Library of Integrated Cellular Signals (LINCS) is a two stage CF program. In the LINCS Pilot Phase, LINCS Phase I, the program was comprised of two data generation centers, with no data coordination center. After four years of funding, the grants underwent a competitive renewal phase in 2014. Phase II of the project was to fund six data generation centers and one data coordination center for six years; this funding is due to complete on June 30th, 2020.

Conceptually, LINCS is based on a *Science* paper from 2006: 'The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease'. The idea is that by collecting gene expression profiles, and other multivariate data, for small molecules before and after the treatment of human cell lines, it is possible to create a reference library that could be used by the community to form hypotheses to accelerate drug and target discovery. The patterns of changes in response to the drug perturbation defines a 'signature'. The procedure is then repeated for various combinations of cell type, phenotypic assays, and perturbations to build a dense database of signatures that LINCS refers to as the LINCS cube. This database is useful for finding small molecules that have a specific effect on cell lines, and this can be used to enable predicting the effects of drugs on the human phenotype. By querying the LINCS database with signatures created by individual investigators, one can identify positive connections (i.e. drugs that mimic a disease signature) or negative connections (i.e. drugs that reverse disease signatures). Most of the LINCS data is from human cell lines, and includes a small set of primary cells.

The structure of the LINCS consortium is complex, and has many different, loosely integrated, pieces. In Phase I, there were 10 production centers: two that did data production and analysis, four that worked on biological technology development, and four that built computational tools. There was no overall coordination effort. In Phase II, the Data Coordination and Integration

Center (DCIC) was established. The DCIC has four components: DSR, which is focused on internal and external research projects; the IKE which maintains metadata and builds the portal, APIs, visualization, and integration tools; the CTO which manages training and outreach, and the CCA, which focuses on coordination and infrastructure. The DCIC is headed by three different PIs from three institutions, and they all contribute to each component.

Together, the DCIC works to coordinate the data coming from the six Data and Signature Generation Centers (DSGCs). Each of the DSGCs has a somewhat different focus for data collection and analysis. The Broad has two teams who are working independently, but sharing some conditions and cell lines. The first has generated the largest amount of data due to their L1000 assay for profiling human cells. The other Broad group focuses on high-throughput proteomics, however, their P100 and GCP assays are not as high-throughput as the L1000 team. The Oregon Health and Science University has a focus on RPPA proteomics and an assay. The DToxS DSGC also from ISMMS is examining how combination of drugs currently used in the clinic might be used to mitigate cardiotoxicity, using a variety of assays. The Harvard Medical School (HMS) DSGC LINCS center focuses on high throughput imaging, proteomics, and phenotypic assays (mostly in cancer) as well as cell signaling and modeling. NeuroLINCS, a multi-institute team, is focused on the systems biology of ALS. They create their own cell lines from patients with ALS and profile them using imaging, ATAC-Seq, proteomics, transcriptomics, and phenotypic assays to get a fuller picture of the molecular mechanisms of ALS. The MEP LINCS center at OHSU is using the microenvironment microarrays (MEMA) assay with microscopy to study how extracellular matrix proteins affect cell signaling in cancer cells. This breadth of studies resulted in the overall LINCS dataset hosting an unusually diverse set of data.

Harmonized data

LINCS has a data ingest pipeline that requires some basic metadata: reagents, assay type, dataset type, experimental metadata, processing pipelines, and data files. However, since all of the DSGCs tend towards having such specialized analyses, most of the metadata terms are non-overlapping. The LINCS DCIC reported that although their data is not generally harmonizable in the sense of having easily combined metadata, there are still ways in which their datasets are interoperable. The signatures created by their experimental methods can be compared independently of the pipeline used to create them. However, the DCIC explained that pipelines using different aligners and/or different expression analysis systems give similar output, resulting in what they refer to as gene lists. “At the gene set level, these pipelines produce approximately similar up/down lists.” Since these gene lists comprise the signature, and are fairly stable, they argue that one RNA-Seq signature can be compared with other RNA-Seq signatures, regardless of the pipeline used to process the data. On balance, the LINCS team also stated that comparability only applied to certain types of experiments.

Cross cutting metadata models

LINCS is committed to ensuring that their portal is compatible with the data model that will be used by the CFDE, referred to as the C2M2. Data in the portal is annotated with metadata, and there is already a version of JSON-LD that describe those datasets in our C2M2.

Self-governed metadata standards

The LINCS data is quite varied and their data is more like a loose confederation of related datasets. The common thread of LINCS datasets is that they can all be described by signatures; however, this does not mean that all of their metadata can be harmonized. The data LINCS hosts is very diverse in terms of data types, approaches, scientific questions, and scientific contexts; LINCS told us that they have no effective authority to enforce a unified processing pipeline, or to impose standards.

They noted that self-governed metadata standards are difficult to make and enforce, even within their own program. Most of the DSGCs have their own specialized pipelines and best practises, which they are not interested in changing. Similarly, although early on in the project everyone agreed to metadata standards for DCIC submissions, most of the DSGCs do not actually follow them. To ameliorate this, the DCIC built a comprehensive API that allows for easy data upload, and that allows the user to map any incongruent variables to the LINCS standard without having to edit their own copies. Still, the DCIC reported that most of the time, they need to go find the data and upload it manually, and this results in significant time following up with the DSGCs, and occasionally even a need to involve the program officers to convince the DSGCs to give them the data at all. Even then, most of the DSGCs do not use the API, opting to instead send files or links with little or no metadata. The DSGCs host their own data, or use data repositories like Synapse (<https://www.synapse.org/>), which mints DOIs and allows them be credited for data use, but also do not require metadata of any kind. So, while the DCIC staff were certain they could meet the minimal requirements suggested by the CFDE, they did not think it was practical to impose broader metadata standards on the data generators.

FAIR Assessments

Findability

The LINCS program as a whole has numerous portals and tools, for example, each DSGC has their own web address. The various portals host mostly non-overlapping information and there is no one portal where a user can access all of the LINCS data. However, a user can search though most of the processed data from the main portal at:

<http://lincsportal.ccs.miami.edu/dcic-portal/>.

Interoperability

Members of the DCIC have taken the initiative to build end-to-end data analysis pipelines, data registration, and building these resources according to the standards agreed on by the consortium. The current version of their data portal has dataset packages, where the data comes from DSGCs, packaged with standardized metadata so the data is intercompatible. All the signature elements, the entities that a user would try to quantify, are also all normalized. At the high level of processing that most of their users are interested in, the LINCS data is interoperable. By using signatures, LINCS has also made progress towards integrating all of their data with external datasets, and building LINCS tools to work with more datasets. They can also support users bringing their own data and instantly comparing it with the LINCS database. Researchers who want to compare their own data to the LINCS database can use the LINCS signature search tools. Since LINCS signatures are such high level overviews of the underlying data, they are, in theory, interoperable with many other Common Fund dataset, provided all Common Fund datasets are processed into signatures. At a more basic level of processing, the diverseness of LINCS data and metadata will make interoperating with other Common Fund assets a challenge.

Accessibility

As a user, building a new dataset from raw data generated at more than one DSGC would be difficult. The user would have to get access to each DSGC portal separately, and learn each system, as each portal has a different implementation, and each DSGC uses differing metadata. However, LINCS also told us that this is not a significant issue for their users because the LINCS projects are so diverse that combining data from multiple sources is often not a common requirement. Most users are interested in the more processed data that is readily available through the main portal, or one type of data which can be accessed from the DSGCs portals.

Reusability

LINCS has created more than thirty specialized analysis apps, and reports that they can be easily applied to other projects or datasets, provided those data are processed and saved in the format expected by the app. The tools themselves are generic and have APIs, however, they would need some additional documentation before they would be easily reusable by the community.

LINCS also told us that the signatures and the process to create them are very reusable. LINCS has done a lot of work on describing their processing pipelines and making them automated, so others should be able to re-use this work.

The data at LINCS, while reuseable in theory, is not uniformly reused. The LINCS datasets are very diverse, and some are much more specialized than others. LINCS told us that about 80% of the types of data they maintain, (about 10% of the overall data volume) are so highly specialized that it is unlikely to ever be re-used outside of the research project it was originally

generated for. However, a great deal of LINCS data is reused frequently. They estimate that by volume, about 90% of their data is reused.

Authentication/Authorization

Covered in [Appendix J](#).

Data Dashboards

The DCIC staff like the idea of a data dashboard, but are extremely leery of its implementation. They told us that early on in their program, they built a data dashboard very similar to the one the CFDE proposed. It monitored where data was physically stored, where it was in the process of being uploaded to the DCIC, how far along it was in the signature generation pipeline, and other similar metrics. However, the DCIC discontinued the use of the tool shortly after implementing it, because it was stressing their relationships with the DSGCs. As was discussed in the self-governed metadata standards section, the DSGCs maintain their own raw data, and it often takes significant effort on the part of DCIC staff to obtain a copy of their processed data. The data dashboard was widely disliked by the DSGCs, because it made the data handover process transparent in ways that disincentivized the DSGCs. For instance, the monitoring system made it simple to tell that a DSGC was holding data that had passed its embargo date, but not submitted to the DCIC. However, datasets that were delayed or withheld for transfer to ensure quality control or to fix other legitimate problems, could be construed by the NIH as simply being late. Keeping the data transfer process more opaque is less useful from a data tracking perspective, but protects the DSGCs from potential embarrassment over difficult to process and QC datasets. Discontinuing use of the data dashboard solved only the political issues: the DCIC is aware of several datasets that were generated by DSGCs more than two years before our interview, but that still have not been uploaded to the DCIC.

CF Data Portal

LINCS is interested in having their data be searchable through a centralized portal, but the CFDE will likely face the same challenges they did with finding overlapping metadata for search. The main data portal at LINCS is on its second version, and is an amalgamation of several portals. From the landing page, a user can choose to search by dataset, small molecule, cell or gene, each of which opens a more specialized portal implementation with varying search capabilities. As each section is highly adapted to the underlying dataset, the LINCS portal could not easily be reused by other DCCs.

Data Platform

Data Hosting

LINCS data as a whole is hosted on local storage clusters that are distributed across the country, and these different storage sites all host varying fractions of the overall data, at varying levels of processing. Here, it is useful to distinguish between the DCIC, which is the official coordinating center for the data, and the DSGCs, which generate the data, as all entities have some of the overall LINCS data, but no one has all of it.

The DCIC

The DCIC portal does not have a complete copy of the overall data created as part of LINCS, for two main reasons. First, the DCIC typically only receives partially processed data from the DSGCs, with varying levels of metadata, so the DCIC has almost never has access to the raw data. LINCS told us that they (and most users) would not have much use for the raw data aside from storing it. A great deal of the LINCS data is generated on highly specialized machines, and so the raw data can only be viewed and processed using equally specialized software that most people do not have access to. Second, the DSGCs are not always timely with their data deposits. The DCIC is aware of several datasets that were generated by DSGCs more than two years before our interview, but that have not been uploaded to the DCIC. The LINCS DCIC also told us that some of the DSGCs are still generating data, and it is unclear whether all of the processed data will be incorporated into the main portal before LINCS funding ends.

The DCIC hosts the main portal (<http://lincsportal.ccs.miami.edu/dcic-portal/>) as well as a tool marketplace (<http://www.lincsproject.org/LINCS/tools>) that contains links to tens of specialized apps created by the LINCS DCIC and the DSGCs. These apps generally do not access all of the data available at the main DCIC portal. A given app website is tailored to a specific audience or question, and only has access to the specific data, at a specific data processing level, that it needs to run. All of the data in the main portal as well as that used by the apps is unprotected.

The infrastructure that supports most of the tools and the data portal are running on local servers but are cloud ready. This means that the apps, including the portal are dockerized and can be launched in a cloud environment without much needed engineering work. For now, Miami, Cincinnati, and Mount Sinai groups each have their own local cluster, each with its own tools, methods, deployment and management solutions. This is done mainly to keep costs low since there is no direct charge to host these servers locally.

The LINCS DCIC suggested that moving to the cloud would not be extremely difficult, but that moving the data would be more difficult than moving the apps. They have tested transitioning their apps to a cloud based system, and currently have 'hybrid' apps that run both on the Amazon cloud, Google cloud, and local resources. However, several challenges remain. Their apps are containerized already for use on the cloud; but the apps are associated with many websites which would require careful coordination for their deployment. The apps are tied to

subsets of the data which also require careful deployment on the cloud. Some apps are owned by the DSGCs, so those groups would have to be involved with transitioning to a cloud platform. The LINCS DCIC team stated that it would also be of value to review which datasets would or would not be useful to the research community, for example, they own a relatively large dataset of imaging of nearly 1 PB, which would be costly to host at Amazon and likely not that useful for the broad research community.

The DSGCs

Most DSGCs store their own raw data and also have their own portals, however, they have varying levels of sophistication and useability. Generally, they do not distribute their raw data to the DCIC, however, some do additionally store their raw data in repositories such as dbGaP, NCBI GEO, Chorus/Skyline (for proteomics), and Synapse. In instances where there is protected data, it is also retained by the DSGCs and stored in dbGaP, but not distributed to the DCIC. LINCS has only a few datasets that require human subject protection.

Data Analysis

As with data storage, data analysis at LINCS is highly distributed. The portals associated with the DSGCs all have some data analysis tools, at varying levels of sophistication and access. The Broad's clue.io platform, for instance, allows users to do analysis of their L1000 data, however, it requires a log-in and has strict rules for commercial use. Other DSGC portals have fewer or no restrictions, and most offer less analysis power.

The DCIC itself also has tens of analysis apps, and these also vary in their data, and access restrictions. The DCIC does not have a single analysis platform or tool. Rather, the apps are all self-contained websites, linked to from the main LINCS portal. A given app website is tailored to a specific audience or question, and only has access to the specific data, at a specific data processing level, that it needs to run.

While none of the DCIC apps contain private data, some require a simple log-in before use. The DCIC staff suggested that the log-in requirement, or lack of one, in their own apps was largely due to the preference of the person who wrote the app rather than there being a particular need for some apps to track users. Internally, there has been some push for a universal log-in, however, it's unlikely to happen in the final year of the program due to other priorities. A universal login system was attempted but it presented both technical, and cultural challenges, and was discontinued. DCIC apps generally do not have access to all of the data available at the main DCIC portal.

The DCIC staff does not have access to usage statistics for the DSGC portals or analysis platforms, however, they know that their own apps are widely used. Their 31 tools, which between them have access to over a million signatures, have had over 700,000 unique users

since 2014, and have been cited by more than 2,500 papers. In 2019, the ninth year of their program, they still consistently see more than two thousand users a day.

LINCS told us that their large number of users results in many technological challenges. Most of their apps were not originally designed to run at this high volume, and maintaining or updating them takes time away from other projects. At the time of our visit, Enrichr, their most popular tool, was under a DDoS attack and had been down for several days, and fixing it was a difficult problem. For ongoing work, LINCS reports their software development practices are improving, and they now approach software in a more professional way. They credited their involvement with the DCPPEC with making their coding practices more deliberate and starting to use systems that help ensure reusability like GitHub and Docker. These practices should help mitigate future issues with new tools, but they could use resources to modernize their current batch of products.

Training

LINCS has a robust outreach and training program that uses a number of modalities: they host a summer research training program, deliver webinars, and offer a Massive Open Online Courses (MOOCs). They reach an astounding number of people at various levels of experience. The summer research program is a data-intensive fellowship that supports ~10-15 undergraduate students per year. This year the program had over 500 applicants. The LINCS webinar series is currently inactive, after running consistently for the first four years of the program. For the series, LINCS invited researchers who were active in the community, but not a part of the consortium, to give seminars, which the DCIC would record and post. These webinars typically have hundreds of viewers, and can be found on YouTube by searching for 'dcic lincs'. LINCS has two MOOCs that have been available for five years through Coursera. One MOOC is about LINCS specifically, while the other teaches systems biology. Between them, the videos within these MOOCs have been viewed more than half a million times. About 100,000 people have officially signed up for the MOOCs, and more than 10,000 of those people completed a series. They also have a fellowship for mid-stage researchers, that falls somewhere between being a postdoc and an assistant professor. The fellows are physically located at a DSGCs and work with both the coordination center and the generation center to do a data intensive project.

LINCS told us that they value their training program for a wide variety of reasons. Their training and outreach programs function as advertising for the LINCS project, and increase its use. They also directly increase the sophistication of their users and encourage proper data use. LINCS credited their fellowship programs especially with swaying talented scientists to do public biomedical research rather than moving on to higher paying corporate jobs. By giving people the freedom and resources to explore interesting projects, the LINCS DCIC members said they can compete with industry for the best candidates. The DCIC also told us that their training program drives research and innovation. When students come to learn about a tool or attend a workshop, their reaction to subject matter can help inform how to build future tools, databases,

and interfaces to be more user friendly, and to accomplish tasks that the development team might not have thought of. Students and fellows that come for longer periods to work on a project for a few months to a year are an important force that keeps projects going. 'Students drive the whole operation forward.' LINCS told us that projects done by students, even high school students, benefit the overall mission of the center, and that their projects seed more and bigger research. The students benefit as well, they gain useful experience in a complex field.

Outcomes

Most of the new outcomes that LINCS would like to see from the CFDE involve building a robust community and assisting periods of transition at DCCs.

CFDE Targets

- Institutional memory:
 - CFDE could serve as a resource for "here are how other DCCs have done things". This would be particularly useful to newly opening DCCs that might be struggling with ramping up operations
- NIH K grantee program
 - If there was a program for people coming out of DCCs, that would help recruit people into them in the first place. The challenge is recruiting good career-focused people to an end-dated project.
- Structure and Organization
 - "At LINCS there was a lot of building the airplane as it flew." LINCS suggested that there is a middle ground between a set rulebook for DCCs and complete freedom. As a young DCC, they would have appreciated a framework to build from rather than starting from scratch. This is closely related to the idea of maintaining an institutional memory.

Game Changers

During our discussions with LINCS we heard about the need for a framework for DCCs, the need for institutional memory, and the feeling of abruptness at both the beginning and end of a DCCs lifecycle. This target is labeled a "game-changer" because it would positively impact the entire ecosystem.

Creating a lifecycle support program for DCCs. LINCS told us about several different challenges they faced as both now and as a young DCC, however these issues all seemed to stem from inadequate support for specific stages. They also told us that their Program officers are very concerned about data continuity. This applies not only to the DCIC data, but for the original data producing centers. They want to sustain resources used to generate data and all of the portals. Conclusions from our discussion with LINCS was the CFDE could create a lifecycle support program to:

- Interview and engage with DCCs throughout their lifecycle to learn about how they are operating, and how they have addressed issues. The CFDE would maintain the institutional memory of lessons learned at various DCCs and pass that knowledge on through DCC engagement.
- Provide DCC best practices and other lightweight structure for young DCCs, to help reduce administrative burden. This could include institutional memory, “here are how other DCCs have done things”, as well as generic tools for common tasks like onboarding.
- Create long-term outlook guides for young DCCs to help them better plan for any asset transitions that are needed at the end of lifecycle.
- Work with DCCs in their last year or two to identify Common Fund assets that should be maintained after the DCC funding ends, and help to transition those assets to a longer term owner, or recommend those assets for longer term funding by the Common Fund.
- Potentially play a role in recruiting and retaining well trained specialists by relocating them to new DCCs.

Agenda for Three Hour Session

10 Minutes Introduction from CFDE

Short introductions from engagement team members and attending LINCS members. The overarching goal for the engagement team is to collect value and process data about LINCS. Values data will include things like: mission, vision, goals, stakeholders, and challenges. Process data includes: data-types and formats maintained, tools and resources owned by LINCS that they would like to have broader use, points of contact for follow up on technical resources, etc.

20-30 minutes LINCS overview

Short overview of LINCS. Can be formal or informal, choose 1-5 topics to cover. Suggested topics: What is your vision for LINCS once the program ends next year? What big problems are you trying to solve? What are your big goals for the next year? Who do you see as your most important users/stakeholders? What project(s) is currently taking up the bulk of your effort/time? What areas of development the LINCS DCIC is putting the most resources into? What is the rough composition of your user base in terms of discipline? Do you have any challenges that are blocking implementation of your current goals? What skill set would you like to add to your project? How do you engage with your users? What kind of sustainability issues are you confronting? Can you currently do combined analyses with external datasets?

30 Review Goals Assessment

Please use the FunRetro board provided to sort this apriori set of goals that we expect DCCs might have **at least the day before** our meeting time. For each goal sticky in the ToSort column, drag it to the column that best describes the current state/thinking/goals of your

organization. Then, leave a comment on each that specifies how desirable that goal is using these terms: “Critical”, “Nice to have”, “Neutral”, “Unnecessary”, and “NA to our org”. Please add as many other comments as you wish. If your organization has a goal that is not listed, please click the ‘+’ at the top of a column to add a new sticky.

1 - 1.5 hours Open discussion

Using the results of the goals assessment and a collaborative format, iteratively discuss goals, blockers, etc., such that the engagement team can accurately describe LINCSS answers, motivations and goals. Topics don’t need to be covered in order, we’d just like to touch on these types of questions.

Topics:

Infrastructure:

- Do you intend to host data on a cloud service?
- Have you already started using cloud hosting? If yes:
 - Approximately how much of your data have you uploaded? How long did that take? How are you tracking progress?
 - What challenges have you faced?
 - How have you dealt with those challenges?
- What potential future problems with cloud hosting are you watching for?
- Does your org use eRA Commons IDs? Do the IDs meet your sign on needs?
 - If yes, did you have/are you having challenges implementing them?
 - If no, what do you use? What advantages does your system provide your org?

Use cases

- What is the rough composition of your user base in terms of discipline?
- What if any, use cases do you have documented? Undocumented?
- What things do people currently love to do with your data?
- What things would people love to do with your data, but currently can’t (or can’t easily)?
- What pipelines are best suited to your data types?
- What are the challenges associated with those desired uses?
- What other kinds of users would you want to attract to your data?

Review of metadata:

- What’s metadata is important for your org? For your users?
- Do all of your datasets have approximately the same metadata? Or do you have many levels of completeness?
- Do you have any data already linked to outside resources?
 - Did you find the linking process easy? Challenging? Why?
- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- What automated systems do you currently have for obtaining metadata and raw data?

Training:

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?

- What challenges keep you from offering the training you'd like?

Policies:

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR:

- Has your org done any self assessments or outside assessments for FAIRness?
- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

Other:

- What search terms would make your data stand out in a shared DC search engine?
- Does your org have any dream initiatives that could be realized with extra resources?
What resources would you need?
- Any other topics/questions the LINCS would like to cover

30 minutes Review of goals and CFC involvement

A quick review of what topics are priorities for the LINCS with suggestions from engagement team on how we can help.

Appendix G - CFDE tech team deliverables to December 2019

<i>Item</i>	<i>Team</i>	<i>Description</i>	<i>Delivery Date</i>	<i>Delierable Type</i>
July				
Financial Report and State of Project	White			
Manual FAIR data gathering	Ma'ayan	5a will be Excel and JSON-LD files and 8a will be collections of files that are ready to be consumed by a database; 5a is at the dataset metadata-level.	31-Jul	
July Report	White	Feed OP Plan; CFDE website for hosting CFDE info from CFDE Awardees; Plan for (C2M2, Portal, Dashboard, Data staging s/w); use cases; high-level	31-Jul	
August				
Financial Report and State of Project	White		15-Aug	
Presentation to NIH	White	Aug 9 presentation on the revised operating plan and proposals within the July 31 report		
CFDE common metadata format	Foster/Kesselman	Document the agreed upon common metadata format	30-Aug	Document
CFDE metadata ingest flow POC - I	Foster/Kesselman	CFDE plans to use DERIVA for generating user views of DCC data manifests. attributes to provide FAIRness metrics.	30-Aug	GitHub Repo
MOTRPAC visit	Brown		27-Aug	
September				
Financial Report and State of Project	White		15-Sep	
Use case update	Brown / RTI	Use cases reviewed monthly at Monday meetings	23-Sep	
FAIR Assessment outline	Ma'ayan/Sansone		27-Sep	
Training Plan	Brown	Feed OP Plan; include dashboard onboarding	27-Sep	
FAIR Stuff (update)	Ma'ayan		30-Sep	
website wireframe	Brown	General over view description of portal	27-Sep	
Automate generation of GTEx metadata	Foster/Kesselman	Capture all manual steps of producing metadata for GTEx data assets in the CFDE common format	30-Sep	GitHub repo
October				
Financial Report and State of Project	Bob/Owen		15-Oct	
Deep Dive Summary Update	Brown		15-Oct	
Complete GTEx metadata ingest flow	Foster/Kesselman	Extend the ingest flow to make the initial step of the flow transform the GTEx metadata into the CFDE common format	23-Oct	Github repo
Training materials (initial)	Brown / RTI		25-Oct	
Operating Plan	NIH			
November				
Financial Report and State of Project	White		15-Nov	
Second DCC for metadata ingest	Foster/Kesselman	Capture the steps to producing metadata a second DCC's data assets in the CFDE common format.	19-Nov	Github repo (code update)
December				
Financial Report and State of Project	Bob/Owen		15-Dec	
website update	Brown / RTI		18-Dec	
CFDE metadata ingest flow POC - II	Foster/Kesselman	Automatically populate a DERIVA catalog using data drawn from at least two DCCs.	18-Dec	Github repo; demo
Training materials (update)	Brown / RTI		18-Dec	
FINAL REPORT	White		23-Dec	
FAIR report (FINAL)	Ma'ayan		23-Dec	

Appendix H - Cost estimators

These two tables are used as baseline numbers to build the approximate costs of events and trainings.

Resource Costs

Salary estimates in this table are based on nationwide ranges from www.glassdoor.com. Travel costs are based on extrapolations from the UC Davis travel policy. Meeting room costs are based on estimates from www.peerspace.com. Catering estimates are from www.thumbtack.com. All salary rates listed here DO NOT contain institutional facilities and administrative rates, and DO NOT contain fringe benefit costs.

Resource	Notes	Low Estimate	High Estimate
FT Software Engineer	Salary only, per year	\$70,000	\$130,000
FT Bioinformatics Analyst	Salary only, per year	\$60,000	\$110,000
FT Bioinformatics Engineer	Salary only, per year	\$60,000	\$120,000
Event Admin	Salary only, per year	\$40,000	\$90,000
Short Travel	1 person, 2-3 days	\$1500	\$2500
Extended Travel	1 person, 1 week	\$3000	\$5000
Meeting room	~30 people per day	\$300	\$500
Multi room meeting space	~100 people per day	\$500	\$1000
AV rentals	per room, per day	\$100	\$500
Assessment	per day, per ~30 people	\$100	\$350
Catering, meals	per person, per meal	\$15	\$40
Catering, snacks	per person, per day	\$5	\$10
Miscellaneous Supplies	per event, per day	\$50	\$100

Time Costs

The following table is estimated based on the extensive workshop development experience at UC Davis.

Outcome	FTEs	Time	Effort
New 2 day workshop materials	1	6 months	100%
New 1 week workshop materials	5	6 months	100%
Updated 2 day workshop materials	1	1 month	100%
Updated 1 week workshop materials	2	1 month	100%
Planning a 2 day event	1	1 month	25%
Planning a 1 week event	1	1 month	75%

Example Specification Workups

The following are the expected costs associated with a two day conference for approximately 30 attendees. The expected total ranges from \$48,580 - \$83,975

Per person Resource	Low estimate	High estimate
Short Travel	\$1500	\$2500
Catering, meals	\$45	\$120
Catering, snacks	\$10	\$20
30 people	\$46,650	\$79,200

Per Event Resource	Low estimate	High estimate
Meeting room	\$600	\$1000
AV rentals	\$200	\$1000

Assessment	\$200	\$700
Miscellaneous Supplies	\$100	\$200
Event Administration	\$830	\$1875
Totals, 2 days	\$1930	\$4775

The following are the expected costs associated with a two day workshop for approximately 30 attendees. For an all-inclusive workshop, where the CF pays for the cost of attendees, we expect the cost to be \$57,025 - \$99,300. A workshop where attendees pay for their own travel and accommodations would be approximately: \$10,105 - \$19,075

All Inclusive:

Per person Resource	Low estimate	High estimate
Short Travel	\$1500	\$2500
Catering, meals	\$45	\$120
Catering, snacks	\$10	\$20
Total:30 attendees plus 5 staff	\$54,425	\$92,400

Per Event Resource	Low estimate	High estimate
Meeting room	\$600	\$1000
AV rentals	\$200	\$1000
Assessment	\$200	\$700
Miscellaneous Supplies	\$100	\$200
Event Administration*	\$1500	\$4000
2 days	\$2600	\$6900

*Paying attendee costs will likely more than double the typical administrative burden

Attendee pays

Per person Resource	Low estimate	High estimate
---------------------	--------------	---------------

Short Travel, staff only	\$1500	\$2500
Staff Catering, all meals	\$45	\$120
Attendee Catering, lunch only	\$15	\$40
Catering, snacks, all	\$10	\$20
30 attendees, 5 staff	\$8,175	\$14,300

Per Event Resource	Low estimate	High estimate
Meeting room	\$600	\$1000
AV rentals	\$200	\$1000
Assessment	\$200	\$700
Miscellaneous Supplies	\$100	\$200
Event Administration	\$830	\$1875
2 days	\$1930	\$4775

Appendix I - GTEx and Kids First joint exercise

AIM 1 - An assessment of potential linkage of DCC-to-DCC assets.

One premise of the CFDE is that it will enable linking between data assets hosted by two different DCCs, and that jointly analyzing those data will result in meaningful results that ideally are of publication quality. This Aim will test this premise, and determine the extent of work that is required to enable linking between two large-scale data generation projects.

GTEx and Kids First have similar types of data (genotypes and vcf files, RNA-Seq expression data, variant calls, sample types, etc). Both DCCs have utilized separate, and possibly incompatible, best practices approaches for grooming these data, so compatibility of the assets at each site must be evaluated before any analysis is done. To determine if it is feasible to co-analyze assets hosted at each DCC, we will perform the following activities:

- determine if the identifiers used by each site to reference genes in the human genome are the same, compatible, or mutually incompatible;
- determine if identifiers that describe gene variants are the same, compatible, or mutually incompatible;
- identify tissues and cell types that are shared between the two DCCs;
- review the formats of VCF files, and establish whether the pipelines that were used for generation of those files are compatible;
- review whether the quality control and gene count pipelines for RNA-Seq data are the same (which is quite unlikely), and if not, whether they are compatible;
- for each of the above assessments, determine what steps, if any, will need to be taken in order to make the data compatible

Note, this Aim refers to the data assets host at each DCC. It does not refer to their metadata.

AIM 2 - Compare pediatric 'normal' to GTEx adult normal gene expression data.

To gain insights into changes in gene expression by age (child vs adult), which will be useful to better understand drug targeting in patients, we will evaluate gene expression data of normal tissue from adult at GTEx versus pediatric samples at Kids First. We assume human protection issues would not allow Kids First data to be displayed at GTEx. However, we will also generate mock ups of how these results could be displayed using visualization tools at GTEx, in a scenario where sharing of human protected data was possible.

As described in Aim 1 above, Aim 2 is dependent on the premise that genes can referenced at each DCC can be mapped, and that the expression levels described at both DCCs are appropriate for comparison, or can be made so.

AIM 3 - Integrate structural, and other, variants with gene expression data.

We will evaluate structural variants derived from studies at Kids First, and review the consequences of these variants in the context of expression. Using variant data we will define the genes that are knocked out, and use GTEx data to ask which of those genes are associated with tissue specific expression. We will evaluate if that tissue expression corresponds with disease phenotypes for any of the genes, if there eQTLs for those genes, and if effect size or direction inform our understanding of what occurs in a heterozygote.

Results will be evaluated, and we will also generate mock ups of examples for how results could be displayed using GTEx visualization tools. For example, we expect that at GTEx, we can query a gene, display all the known disease mutations from Kids First for that gene on the genome browser, and determine whether those genes can be mapped to eQTLs, splice-QTLs, chromatin marks or methylation sites. We will also test for tissue specificity of the effects, and if they match phenotype specificity.

As described in Aim 1 above, Aim 3 is dependent on the premise that genes can referenced at each DCC can be mapped, and that the expression levels described at both DCCs are appropriate for comparison.

COST ESTIMATION

This project will require 1 Bioinformatics Analyst for data analysis and quality control, and 2 Software/Bioinformatics Engineers, 40% effort from a senior level supervisor and 6 months of computational support. Based on using the high range for salary estimates:

Administrative / senior supervisor: \$170,000

Software or Bioinformatics Engineer: \$130,000

Bioinformatics Analyst: \$110,000

Computing costs: \$1800/month

Total costs are approximately \$448,000 for each site, totaling: \$897,600

Appendix J - Single sign on and authorization assessment

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

Authors and Contributors	4
Versions	4
Document Link	4
Introduction	4
Research on the Cloud	5
Report Goals	8
Report Anti-Goals	9
Profile of Common Fund Data Coordination Centers	10
Overview	10
Genotype-Tissue Expression (GTEx)	10
Overview	10
Current DCC Functionality	10
Datasets	11
V7	11
V8	12
Current SSO, AuthN/Z, and Clouds Storage Strategies	13
Single Sign On	13
GTEx Portal via Google OIDC	13
dbGaP via eRA Commons	14
Authentication and Authorization	16
Cloud Storage	17
What Works Well?	18
Single Sign On	18
Authentication and Authorization	19
Cloud Storage	19
What Does Not Work Well?	19
Single Sign On	19
Authentication and Authorization	20
Cloud Storage	20
What Would a Shorter-Term Improvement Look Like?	21
What Would a Longer-Term Solution Look Like?	22
Gabiella Miller Kids First Pediatric Data Resource	23

Overview	23
Current DCC Functionality	23
Functionality	23
Infrastructure	27
Datasets	27
Current SSO, AuthN/Z, and Cloud Storage Strategies	28
Single Sign On	28
Authentication and Authorization	29
Cloud Storage	31
What Works Well?	32
Single Sign On	32
Authentication and Authorization	32
Cloud Storage	33
What Does Not Work Well?	33
Single Sign On	33
Authentication and Authorization	33
Cloud Storage	34
What Would a Shorter-Term Improvement Look Like?	34
What Would a Longer-Term Solution Look Like?	35
DCC Requirements and Preferences Summary	36
Available Solutions	38
SSO	38
Google Sign-in and Facebook Login	39
eRA Commons/NIH Login	39
Auth0, Keycloak, and other SSO Implementations	39
Authentication and Authorization	41
Gen3 - Fence	41
NCBI JWT based on Virtual Directory Service from CIT	42
Cloud Storage	44
Gen3 - Indexd + Fence	44
SRA Cloud Storage	44
Emerging Themes	46
Shorter Term	46
Longer Term	46
Shorter Term Proposal	47
Overview	47
Goals	47
Anti-Goals	48

Requirements	48
Proposed Work	49
SSO	49
Authentication and Authorization	50
Authentication: OIDC for eRA Commons login	50
Authorization: JWT system from NCBI	50
Data Access on Cloud	50
Sharing Knowledge	50
Longer Term Proposal	50
Passport	50
DUOS	50
Conclusions	52
Next Steps	52
Appendix - Technology Primer	52
Core Concepts	52
Authentication	53
SAML	54
Benefits:	54
Drawbacks:	55
Very Simplified Auth Process:	55
OIDC (OpenID Connect)	55
Authentication and AWS/Google	56
Authorization	57
SAML	57
OAuth 2	57
Authorization and AWS/Google	58
Storage on the Cloud	58
S3	58
Google Storage (GCP)	59

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.

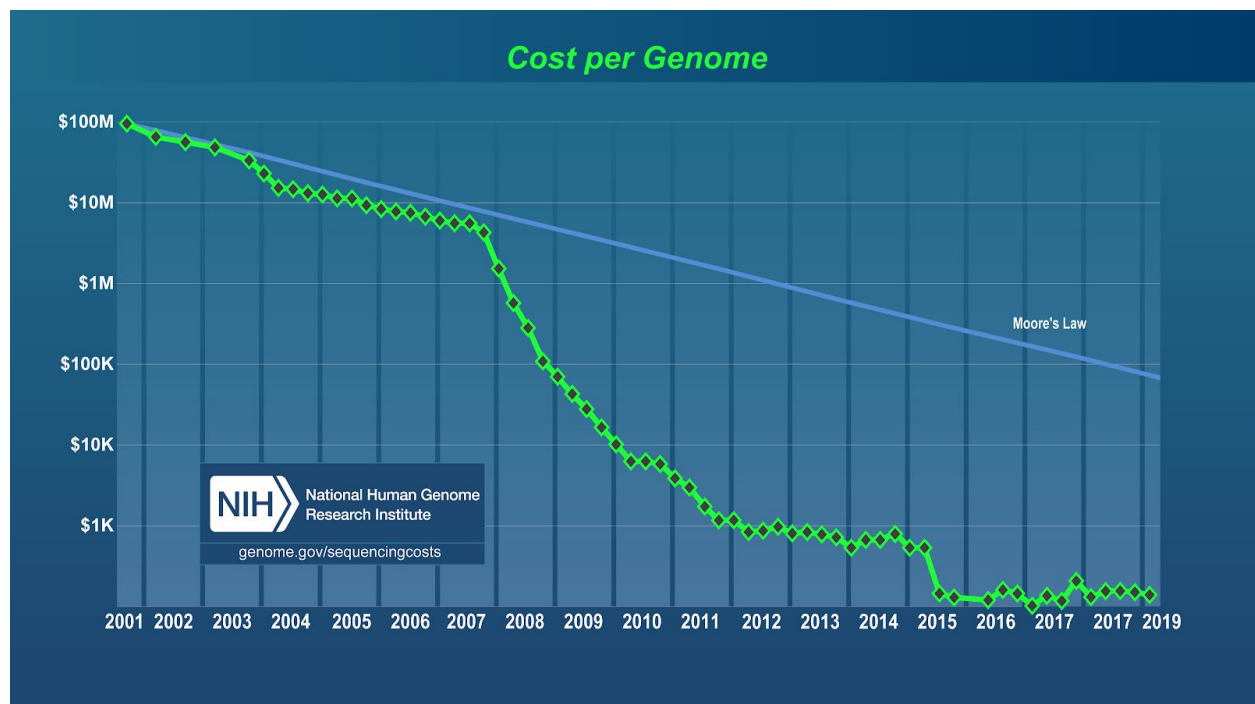


Figure 1: The dramatic decline in the cost per genome over time, source NHGRI (<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>)

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).

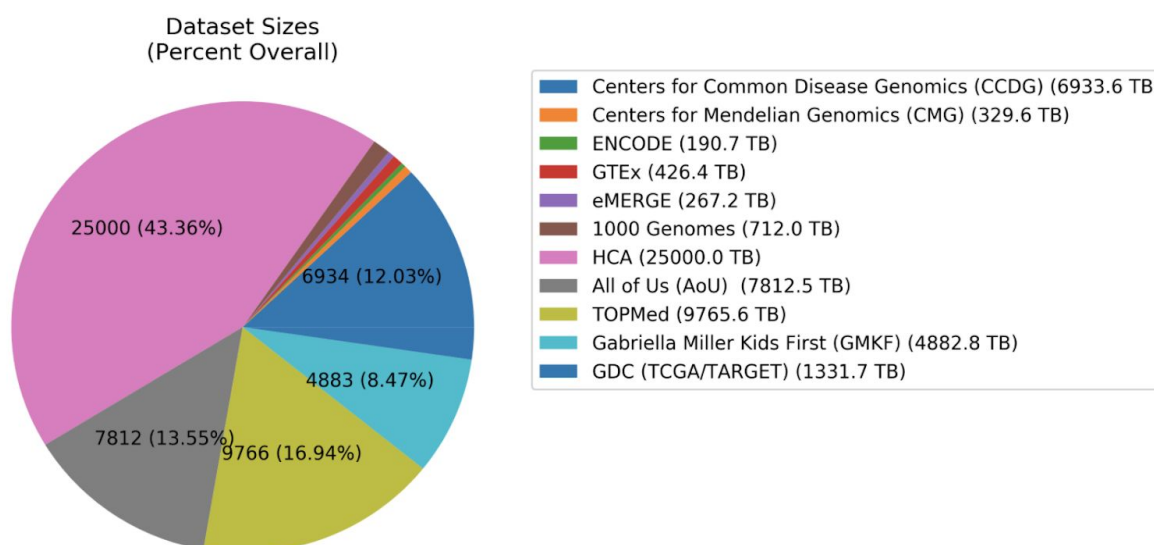


Figure 2: An approximate projection of cloud-based, genomic dataset sizes over the next 5 years based on stated data generation goals from project websites and FOAs for several notable projects.

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 billion¹.

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.

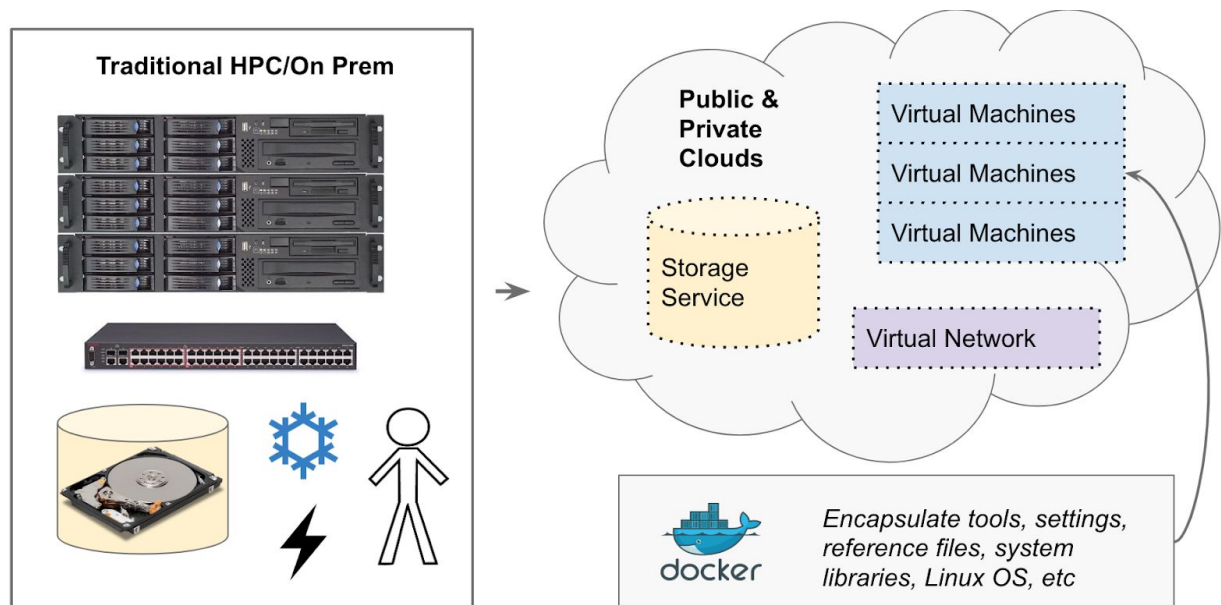


Figure 3: The cloud transition in biomedical research. Researchers are moving from local, self-hosted infrastructure that requires expensive upfront costs and maintenance to cloud systems offering compute and storage for rent. Coupled with datasets being stored and distributed on the cloud, researchers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

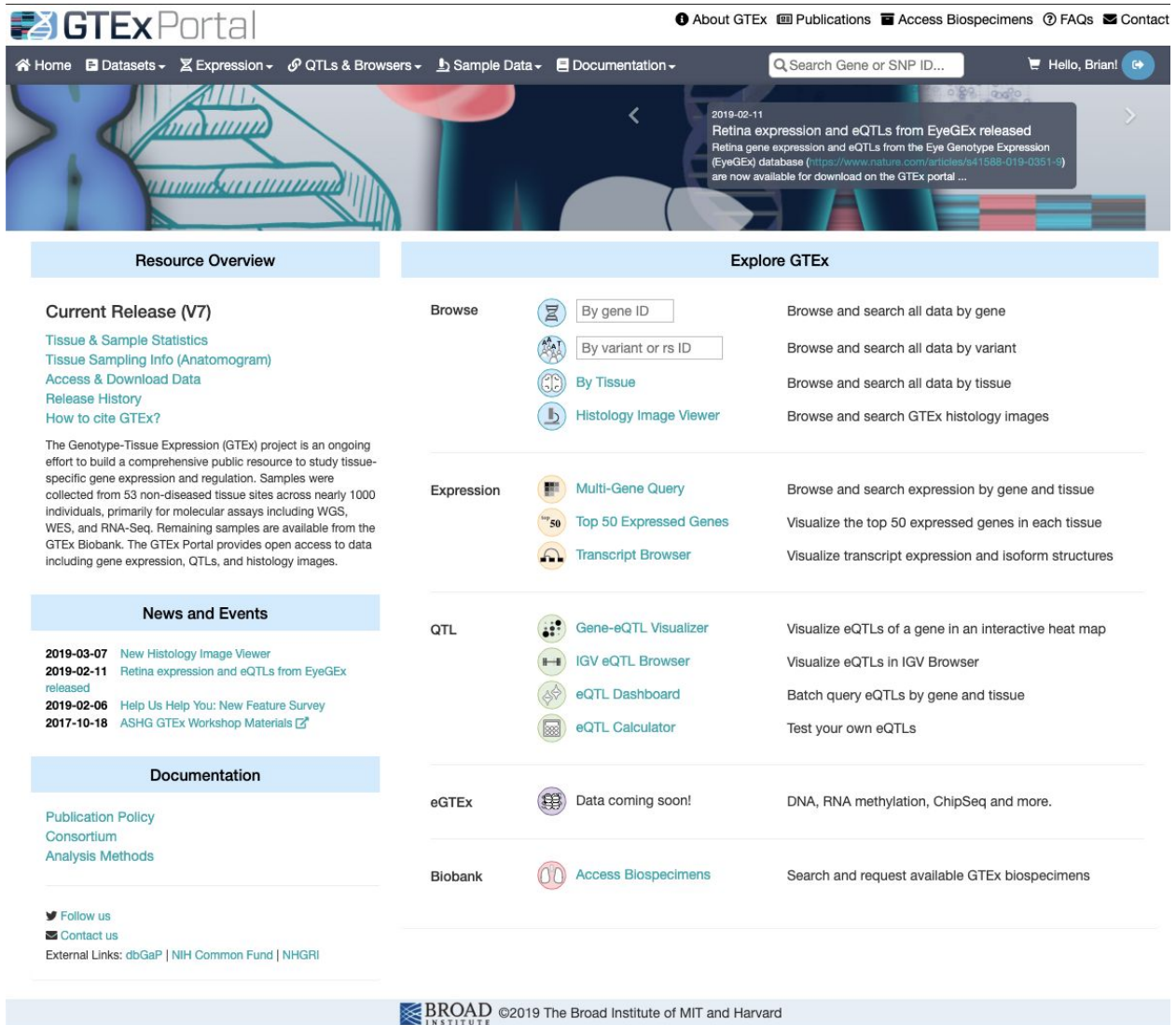


Figure 4: GTExPortal provides access instructions for data and biosample ordering in addition to several easy to use analysis and visualization components.

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.

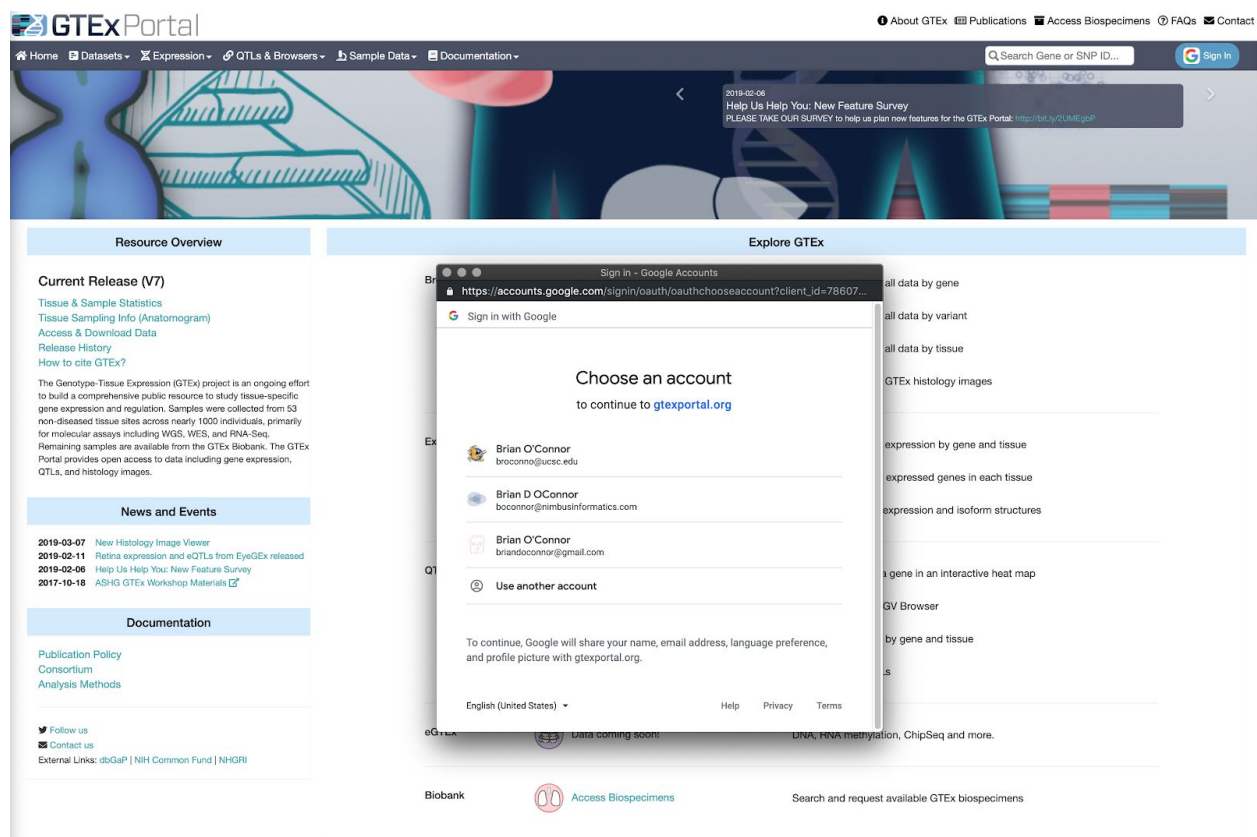
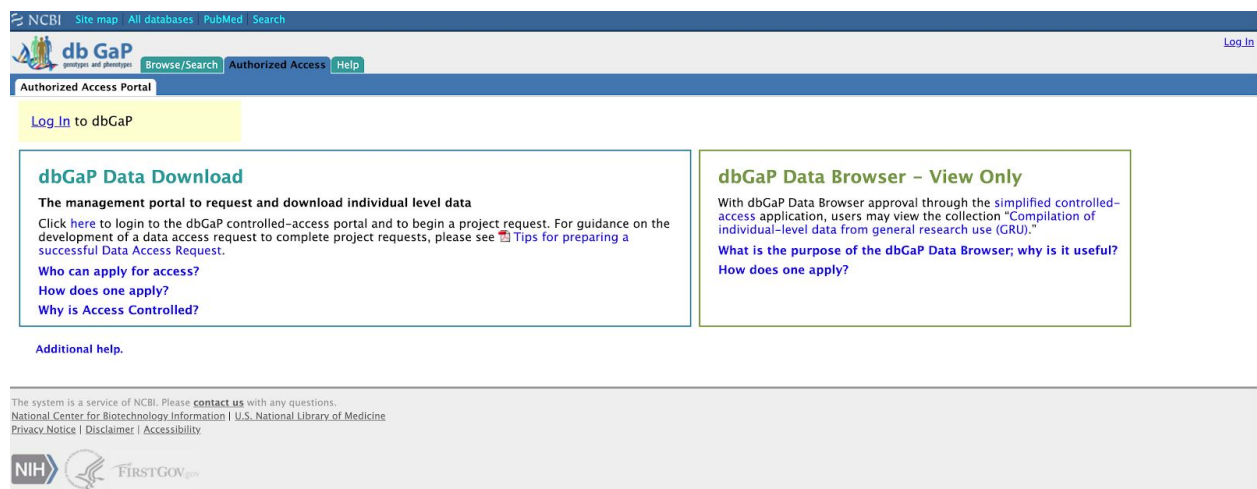



Figure 5: Clicking login takes a user to a Google to perform a login using the OIDC flow.


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.




eTrust
NIH SECURE IDENTITY SOLUTIONS

User Name: 

Password:  [Change Password](#)

OR

 Insert your PIV card into your smart card reader before attempting to login.
For more information visit <http://smartcard.nih.gov>

Log in

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](#)


NIH Center for Information Technology 

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.

NCBI dbGaP SRA Run Selector Help Permalink

Organism: Homo sapiens
Platform: ILLUMINA

Common Fields

	Runs	Bytes	Bases	
Total:	42,961	628.39 Tb	2.46 P	

Download

RunInfo Table Accession List

Selected: RunInfo Table

24,455 Runs found Page: 1 2 3 4 5 6 7 8 ... 490

dbGaP study	dbGaP study name	Consent	BioProject	SRA Study	Subject is affected	Sex	Analyte type	Assay type	Center	Run	BioSample	Sample name
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR598484	SAMN00992416	GTEx-PW2O-0526-SM-213DX
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8228298	SAMN00992416	GTEx-PW2O-0526-SM-213DX
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR598124	SAMN00992418	GTEx-NPJ8-0011-R4a-SM-2HML3
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8227730	SAMN00992418	GTEx-NPJ8-0011-R4a-SM-2HML3
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR599192	SAMN00992420	GTEx-N7MT-0011-R5a-SM-213G6
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8227642	SAMN00992420	GTEx-N7MT-0011-R5a-SM-213G6
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR801925	SAMN00992423	GTEx-OHPK-0526-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8227810	SAMN00992423	GTEx-OHPK-0526-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR801068	SAMN00992424	GTEx-Q2AG-0126-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8228416	SAMN00992424	GTEx-Q2AG-0126-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR802598	SAMN00992427	GTEx-Q2AG-0011-R9a-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8228415	SAMN00992427	GTEx-Q2AG-0011-R9a-SM-2HMLB
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR807586	SAMN00992428	GTEx-OXRL-0526-SM-213EZ
<input type="checkbox"/> phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8227984	SAMN00992428	GTEx-OXRL-0526-SM-213EZ

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.

Terra BETA WORKSPACES

Workspaces > anvil-datastorage/AnVIL_GTEx_V8_hg38 >
Data (read only)

Dashboard DATA NOTEBOOKS WORKFLOWS JOB HISTORY

TABLES +

participant (979)

sample (17382)

REFERENCE DATA +

OTHER DATA

Workspace Data

Files

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 identity, 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_GTEEx_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_GTEEx_V8_hg38/data) for use by the NHGRI AnVIL project.

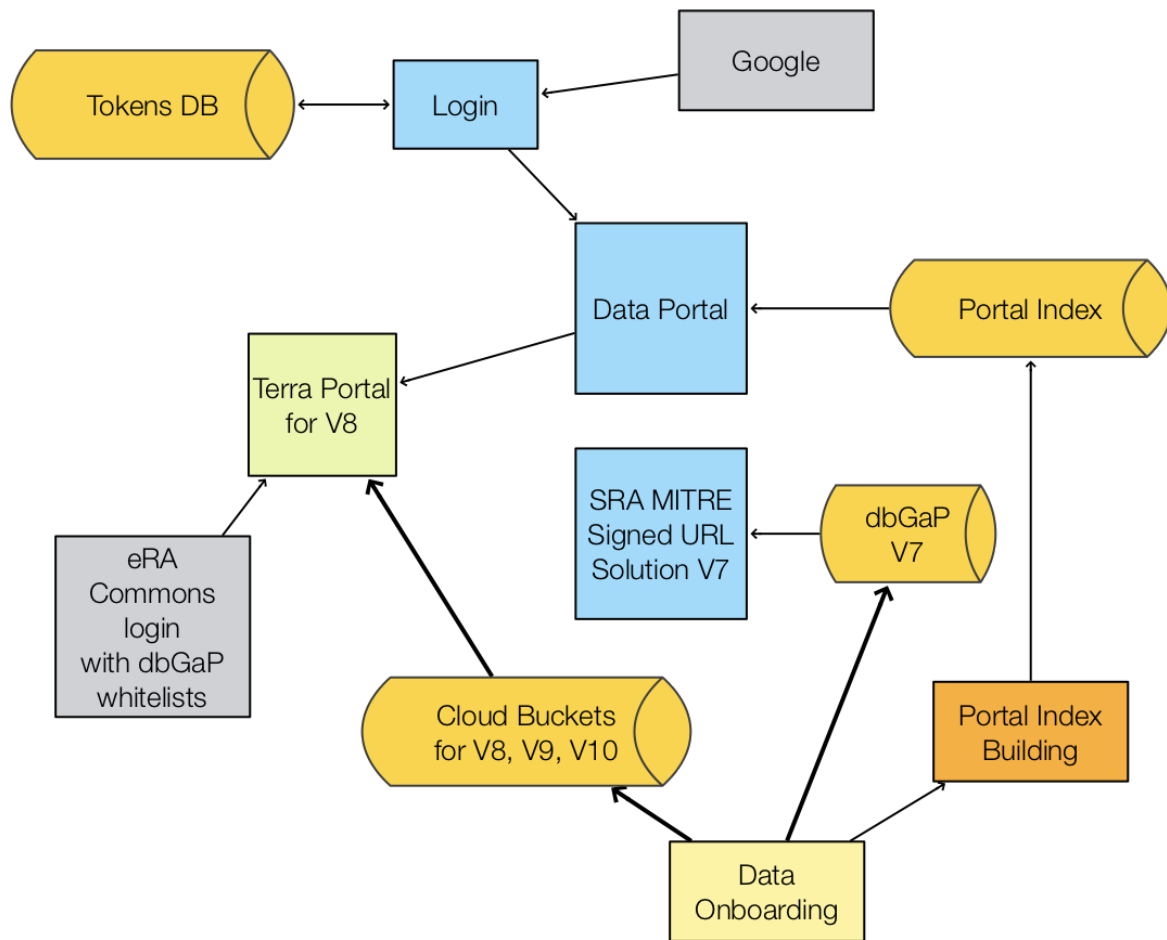


Figure 9: The current GTEx portal architecture. V8 data is supported on Terra directly through access on Google Storage buckets and linked eRA Commons/dbGaP whitelists. V7 is present in dbGaP and cloud copies are shared via signed URLs from the SRA/MITRE solution.

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 safety net 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.

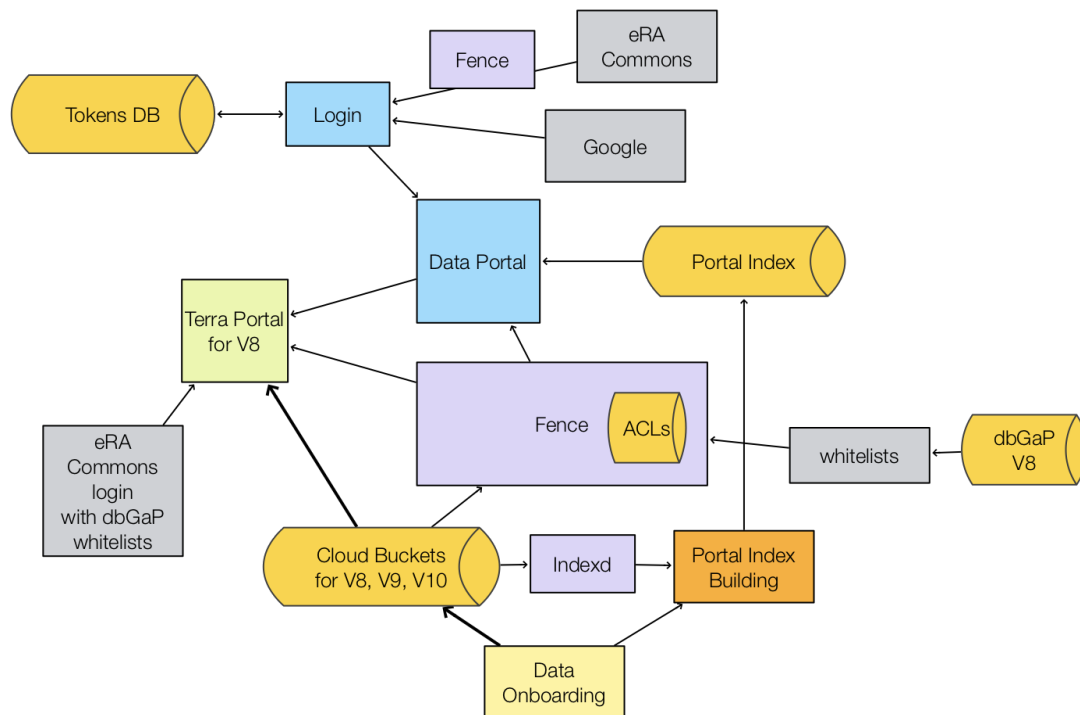


Figure 10: The DCC could use a system like Fence to provide both signed URL and native URI access to data in a cloud bucket. This would allow Terra to work with V8 data (as it does now) and also allow the portal to offer controlled access data via signed URLs for direct download for user logged in with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

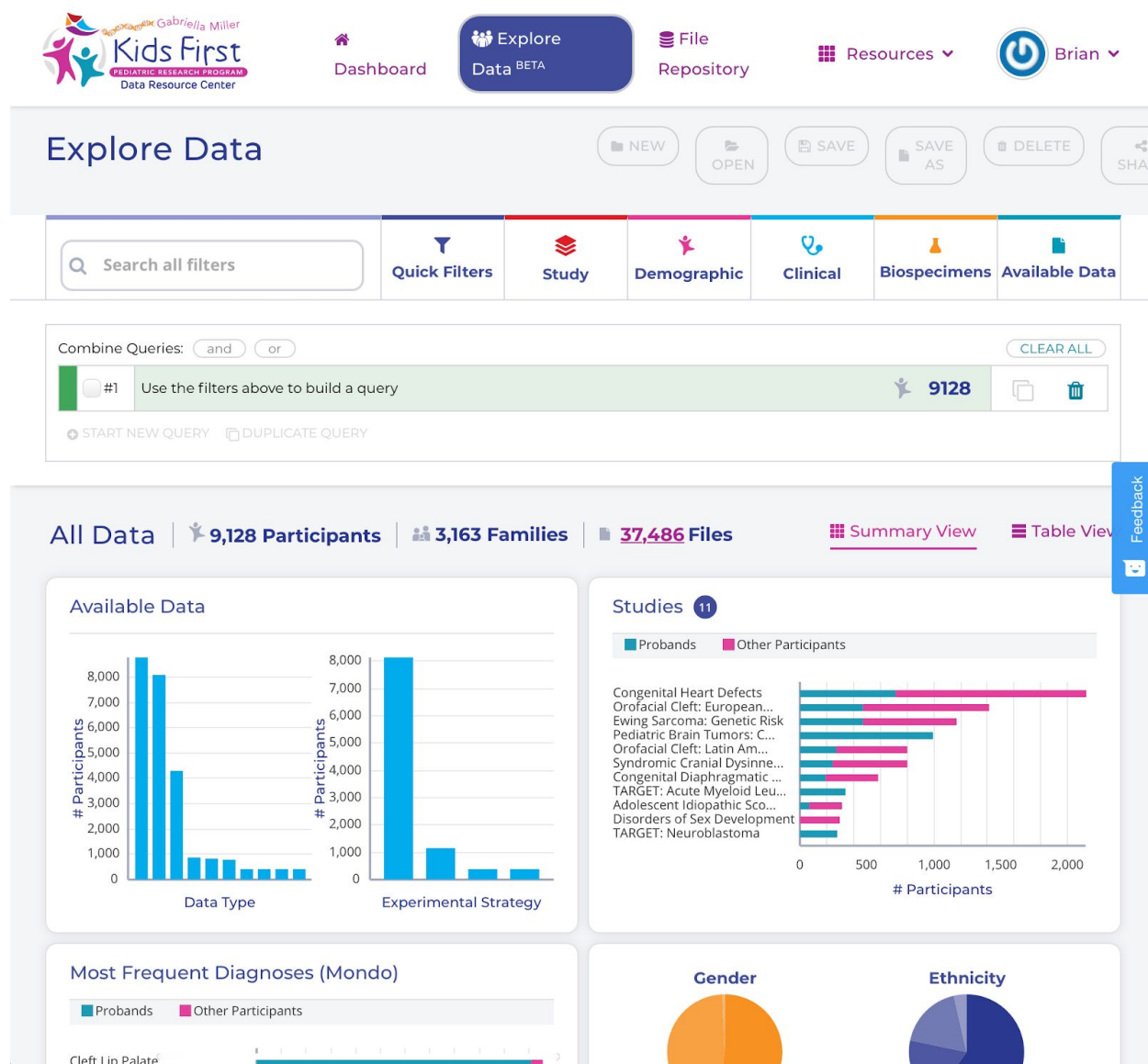


Figure 11: The Explore Data interface.

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.

The screenshot displays the Kids First File Repository interface. The top navigation bar includes the Kids First logo, a home icon, 'Dashboard', 'Explore Data BETA', 'File Repository' (active), 'Resources', and a user profile 'Brian'. The left sidebar contains filter sections: 'Experiment Strategy' (WGS: 1,608 files), 'Harmonized Data' (Any: 1,346, No: 262), 'Data Type' (Variant Calls: 1,608), 'File Format' (vcf: 1,608), and 'Family Shared Data Types' (Variant Calls: 1,604). The main content area shows a summary: 1,608 Files, 4,288 Participants, 1,414 Families, and 1.56 TB Size. It includes buttons for 'ANALYZE IN CAVATICA', 'DOWNLOAD', 'Columns', and 'Export TSV'. A table lists files with columns: File ID, Participant..., Study Name, Proband, Family Id, and Data Type. The table contains five rows of data, all with 'Variant Calls' as the data type.

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_KHMXXKH...	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.

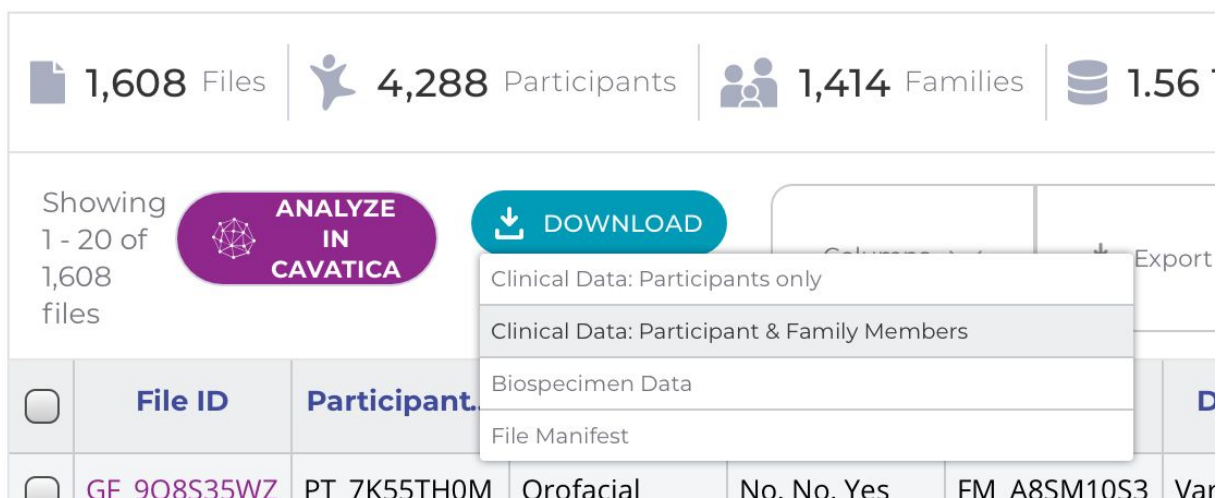


Figure 13: Download options available in the portal.

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.

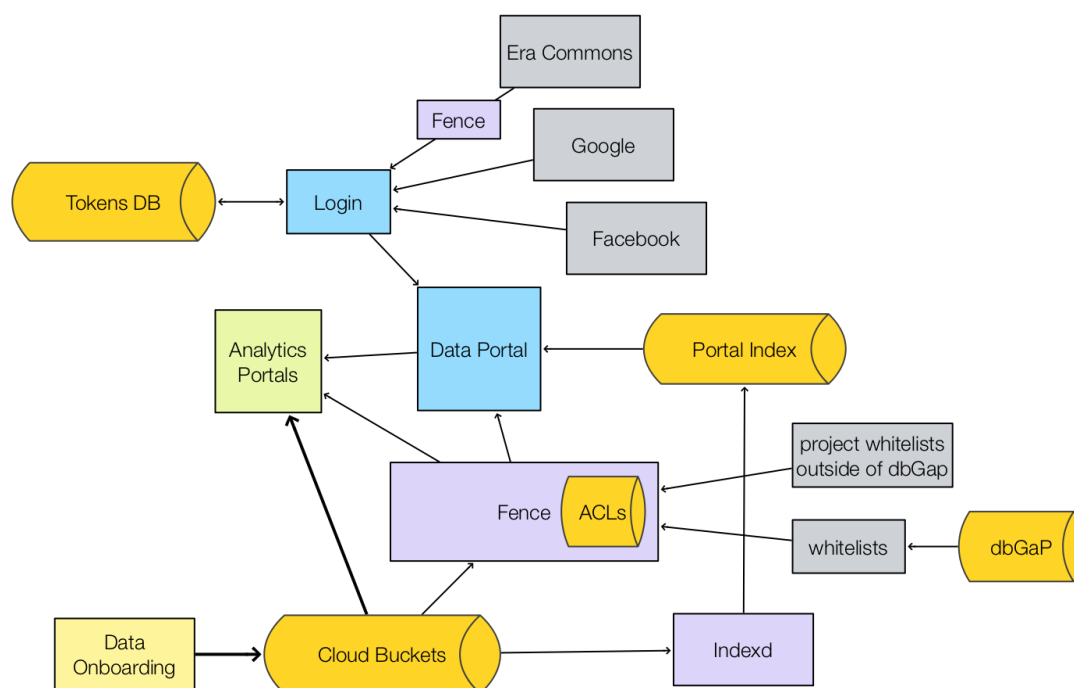


Figure 14: The current architecture of the GMKF Data Resource Center. Arrows represent information/data flow with darker lines indicating large data file access.

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.

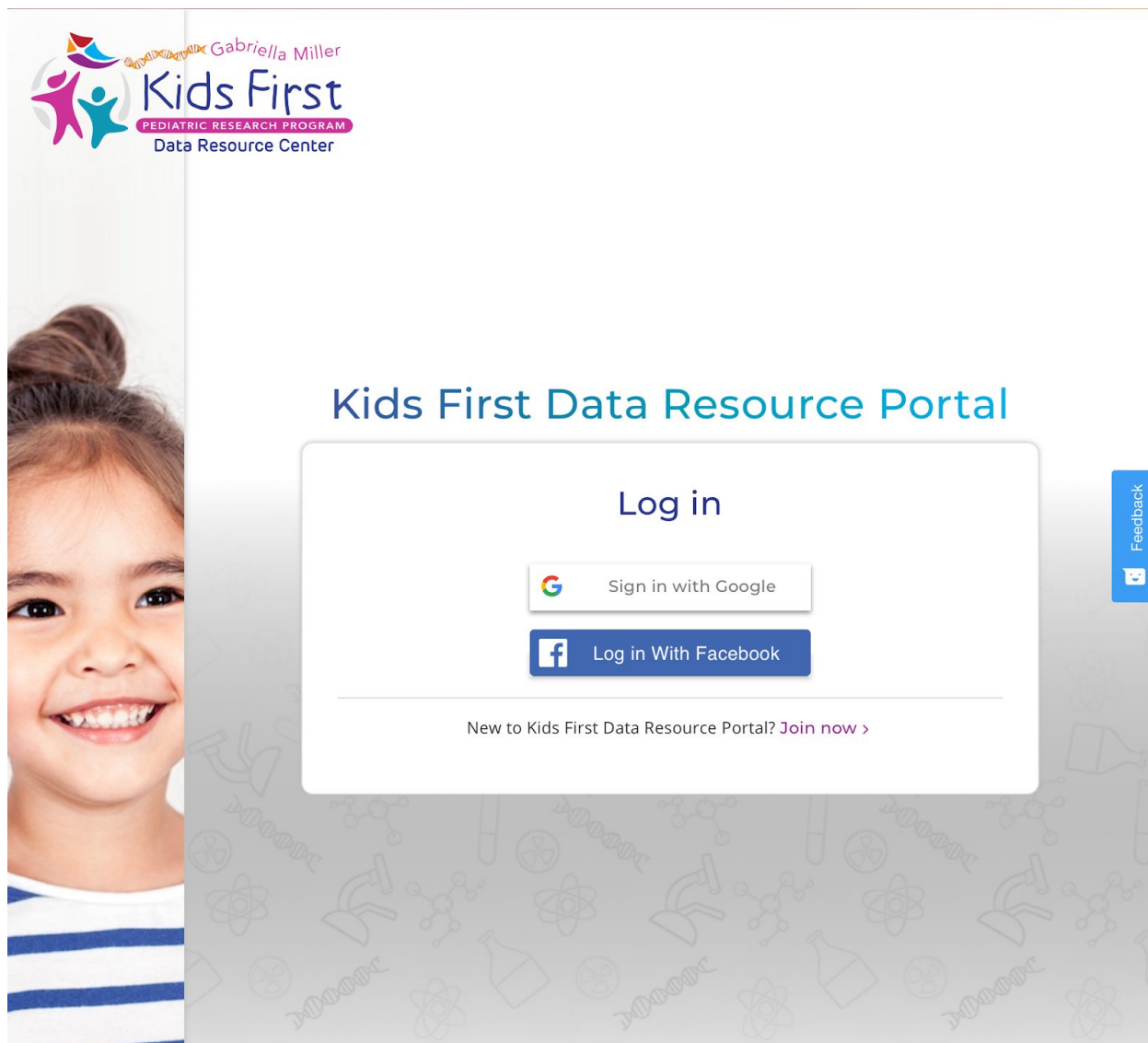



Figure 15: The GMKF Data Resource Portal uses either Google or Facebook login SSO solutions for establishing the identity of users.

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.




[Dashboard](#)[Explore Data BETA](#)[File Repository](#)[Resources](#)

 **Brian** ▾









Email Address:

broconno@ucsc.edu



 Connected with Google

Data Repository Integrations

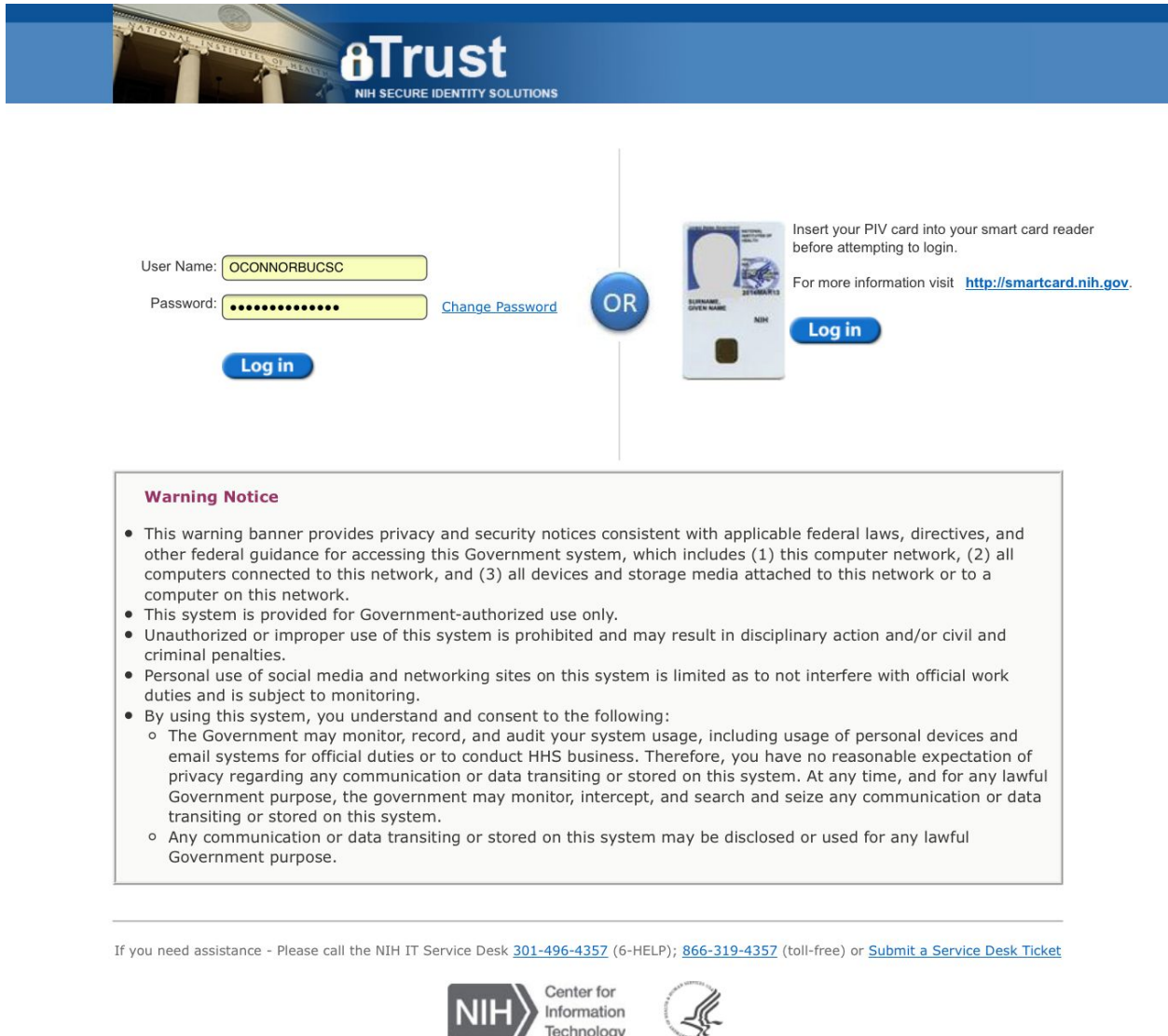
The Kids First DRC provides the ability to integrate across different data repositories for pediatric research. By connecting to each integration (powered by [Gen3](#)), you will immediately have the ability to analyze the data available from these repositories in Cavatica through this portal. Please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

 	Access all released Kids First controlled access data by connecting your account using your NIH login credentials.	<div><div> AUTHORIZED STUDIES</div><div> DISCONNECT</div></div>
 	Access select NCI controlled access data by connecting your account using your NIH login credentials.	<div><div> AUTHORIZED STUDIES</div><div> DISCONNECT</div></div>

Application Integrations

	Analyze data quickly by connecting your account to the cloud compute environment, Cavatica .	<div> CONNECT ></div>
---	--	---

Feedback



eTrust
NIH SECURE IDENTITY SOLUTIONS

User Name:

Password: [Change Password](#)

Log in

OR

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

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

Log in

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](tel:301-496-4357) (6-HELP); [866-319-4357](tel:866-319-4357) (toll-free) or [Submit a Service Desk Ticket](#)

NIH Center for Information Technology

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.

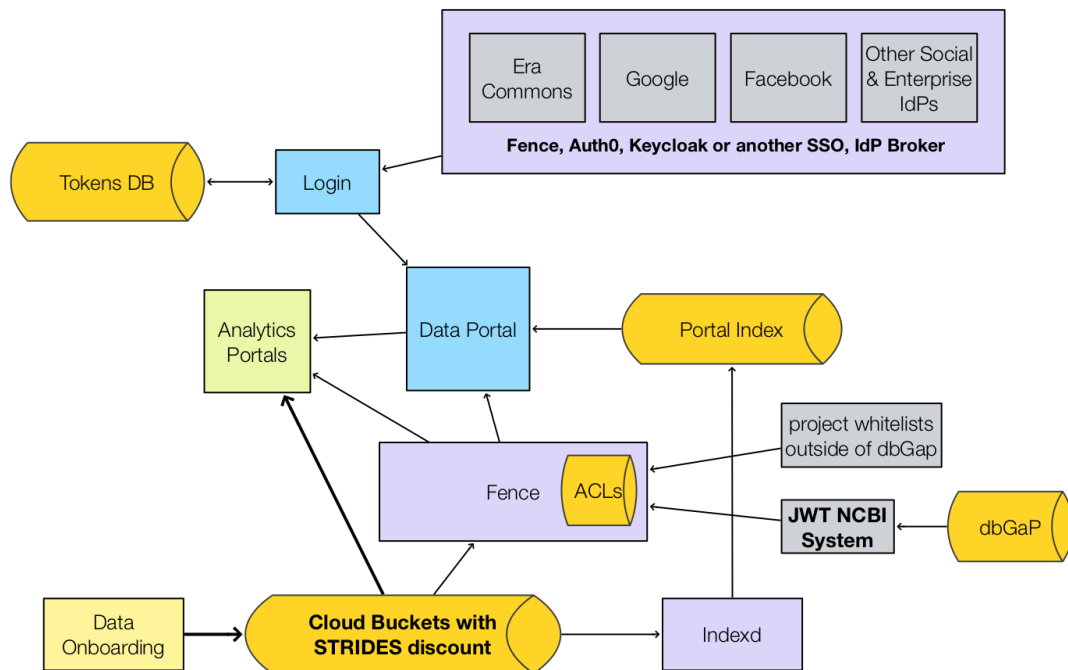


Figure 17: An overview of how a prototypical DCC might work in the next year based on our interview and suggestions by the GMKF Data Resource Center. Text in bold are items that would be fruitful components to improve, prototypes, and test in the short term.

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 researcher 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 broker 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 GTEEx 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 GTEEx 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 as Data GUIDs (<https://dataguides.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.

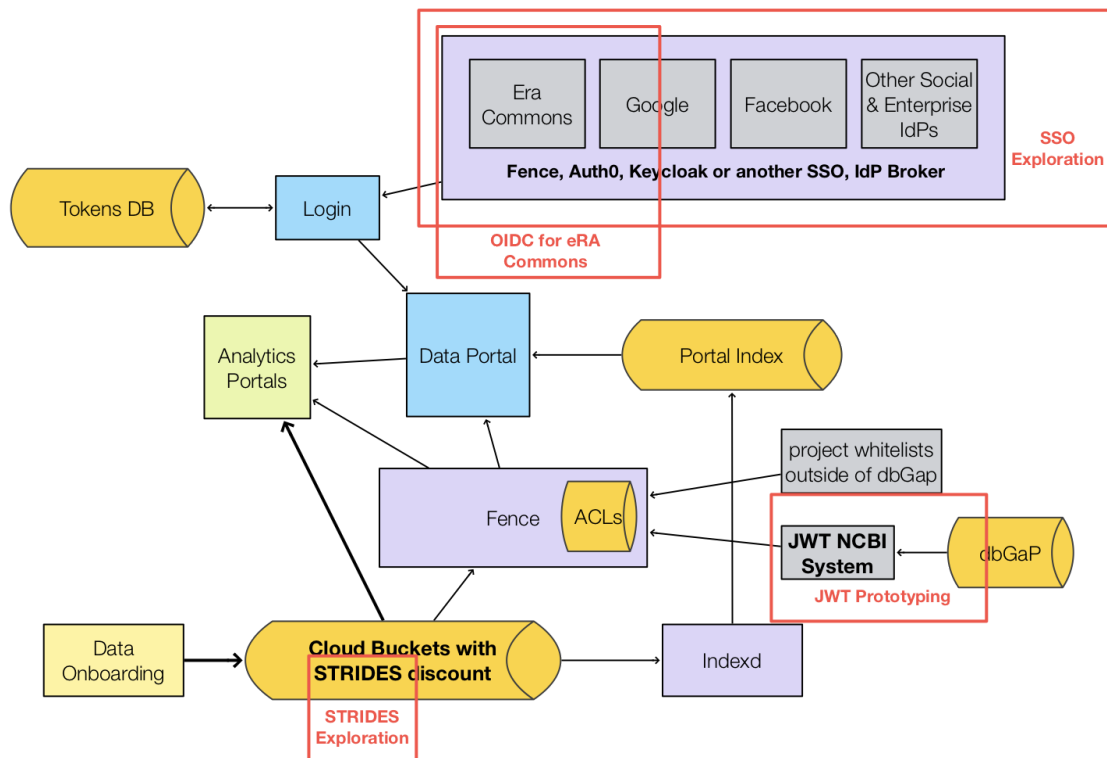


Figure 18: Areas (highlighted with red boxes/text) where a short term effort could both improve the functionality of the GMKF DCC but also provide a template and working example for other DCCs to leverage.

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.

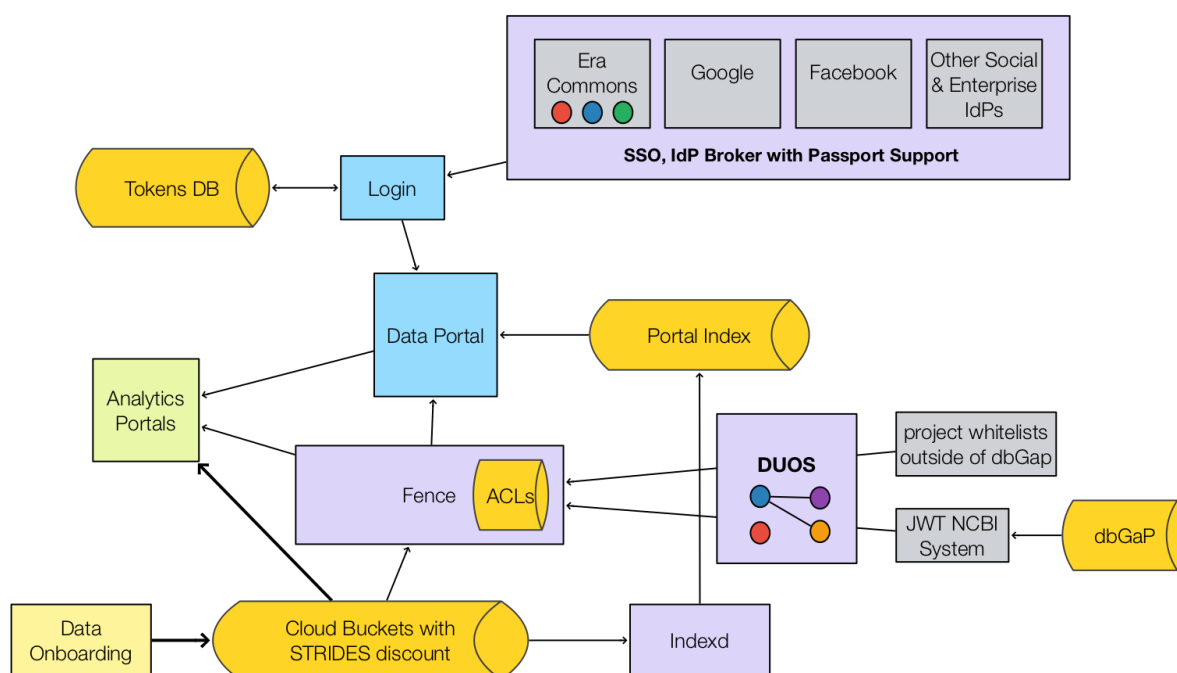


Figure 19: An example of how DUOS and researcher identity Passports can be overlaid into a DCC similar to the GMKF architecture. User identity tokens include claims about the researcher while DUOS performs matching between the user and dataset access policies, proving the result to a system like Fence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

²SAML 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 academic³ and research facilities⁴. It's also the older implementation that has primarily been replaced by Open Connect / OpenID⁵.

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:

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.

⁷OAuth 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.