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

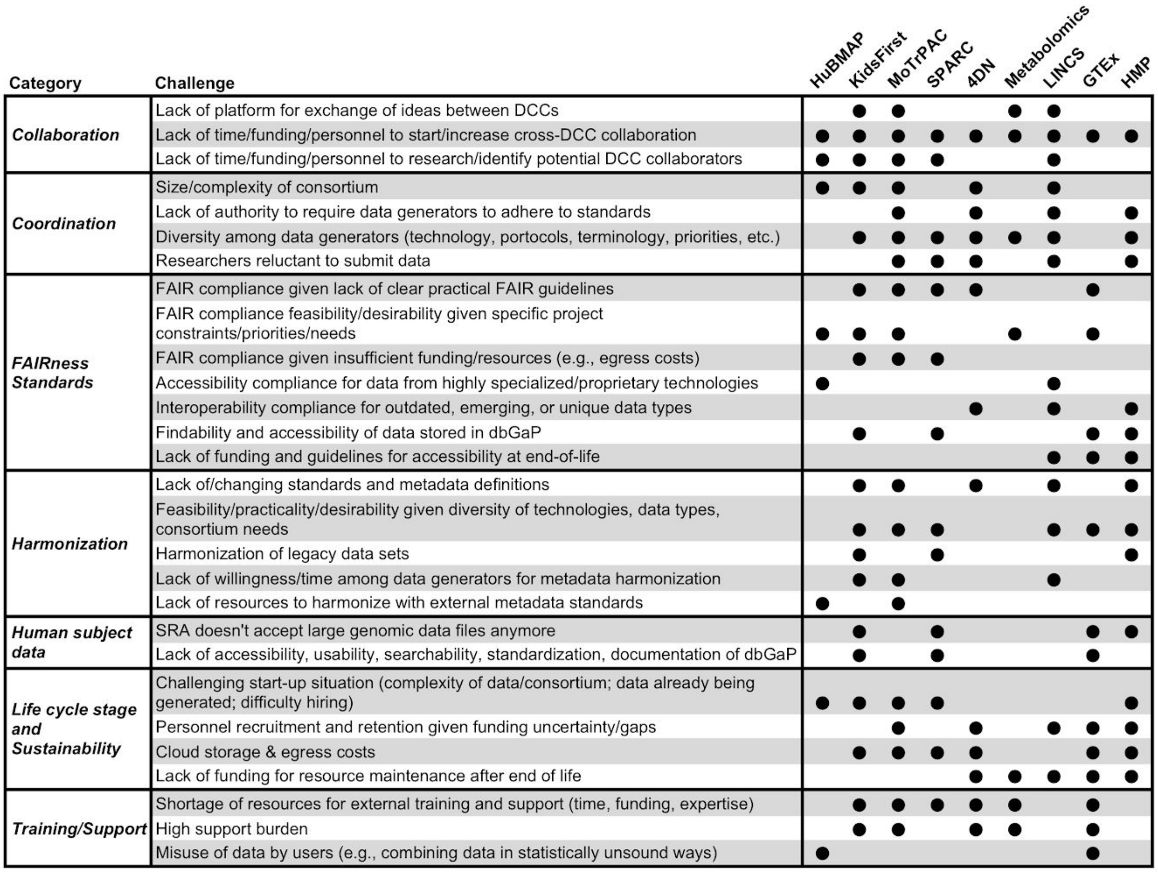
December, 2019

Metabolomics Workbench
MoTrPAC
HuBMAP
SPARC
GTEX
CFDE
4D Nucleome
HMP
LINCS
Gabriella Miller Kids First Pediatric Research Program Data Resource Center

Introduction

Common Fund datasets are highly diverse -- representing collections of genomic, expression, proteomic, metagenomic, and imaging assets. The CF data are also incredibly deep, derived from hundreds of studies, with samples collected from thousands of human subjects. The sheer volume, richness and complexity of data challenges clinical and biomedical researchers to use the data effectively. There are tremendous opportunities to organize this inherently complex data to better support researchers.

During 2019 we engaged in a listening tour of Common Fund Data Coordinating/Resource Centers (DCC)s to better understand the obstacles DCCs face in making Common Fund datasets more accessible to researchers. The effort resulted in two previous interim reports [July, October] that describe a series of institutional burdens faced by the CF DCCs that impede interoperability across the Common Fund. Overcoming these challenges will lead to a more vibrant Common Fund digital ecosystem of interoperating datasets.



Category	Challenge	HuBMAP	KidsFirst	MoTrPAC	SPARC	4DN	Metabolomics	LINCS	GTEx	HMP
Collaboration	Lack of platform for exchange of ideas between DCCs		●	●			●	●		
	Lack of time/funding/personnel to start/increase cross-DCC collaboration	●	●	●	●	●	●	●	●	●
	Lack of time/funding/personnel to research/identify potential DCC collaborators	●	●	●	●			●		
Coordination	Size/complexity of consortium	●	●	●		●		●		
	Lack of authority to require data generators to adhere to standards			●		●		●		●
	Diversity among data generators (technology, portocols, terminology, priorities, etc.)		●	●	●	●	●	●		●
	Researchers reluctant to submit data			●	●	●		●		●
FAIRness Standards	FAIR compliance given lack of clear practical FAIR guidelines		●	●	●	●			●	
	FAIR compliance feasibility/desirability given specific project constraints/priorities/needs	●	●	●			●		●	
	FAIR compliance given insufficient funding/resources (e.g., egress costs)		●	●	●					
	Accessibility compliance for data from highly specialized/proprietary technologies	●						●		
	Interoperability compliance for outdated, emerging, or unique data types					●		●		●
	Findability and accessibility of data stored in dbGaP		●		●				●	●
	Lack of funding and guidelines for accessibility at end-of-life							●	●	●
Harmonization	Lack of/changing standards and metadata definitions		●	●		●		●		●
	Feasibility/practicality/desirability given diversity of technologies, data types, consortium needs		●	●	●			●	●	●
	Harmonization of legacy data sets		●		●					●
	Lack of willingness/time among data generators for metadata harmonization		●	●				●		
	Lack of resources to harmonize with external metadata standards	●		●						
Human subject data	SRA doesn't accept large genomic data files anymore		●		●				●	●
	Lack of accessibility, usability, searchability, standardization, documentation of dbGaP		●		●				●	
Life cycle stage and Sustainability	Challenging start-up situation (complexity of data/consortium; data already being generated; difficulty hiring)	●	●	●	●					●
	Personnel recruitment and retention given funding uncertainty/gaps			●		●		●	●	●
	Cloud storage & egress costs		●	●	●	●			●	●
	Lack of funding for resource maintenance after end of life					●	●	●	●	●
Training/Support	Shortage of resources for external training and support (time, funding, expertise)		●	●	●	●	●		●	
	High support burden		●	●		●	●		●	
	Misuse of data by users (e.g., combining data in statistically unsound ways)	●							●	

Table 1: A summary of site visits to 9 independent DCCs conducted over 2019.

TABLE 1 lists the challenges faced by the DCCs. There are a number of pain points for DCCs associated with their lifecycle stage, the absence of standardized FAIR metrics, the need for data harmonization, the reusability of infrastructure tools, Single-Sign On and user access to protected data, and a strong need for training and user support.

Challenges Faced By DCCs

Lifecycle Stage

There are common sets of life stage challenges among DCCs as they are initially ramping up, as they are expanding to meet the needs of their research consortium, and as they come towards completion of their funding.

Early lifecycle issues revolve around the number of important decisions that must be made in a short time frame, often with little guidance or institutional knowledge. Beginning a project typically requires hiring qualified experts as quickly as possible. Choosing infrastructure technology and establishing protected data policies within the consortium and with NIH requires a great deal of subject matter expertise and forward planning. However, at most institutions, awardees often need to wait until their awards are processed to begin advertising open positions, taking on infrastructure contracts, or devoting their own time and effort to the project. This means that most DCCs have a 6-12 month lag time between their award start and having a full, on-boarded, work-force, ready to begin the ‘real’ work. Since award money is typically split evenly over the grant period, most DCCs reported having far more money than they could spend in the first year. Unfortunately, this money is often lost, as DCCs are not typically allowed to carry over funds.

Middle-aged DCCs face both social and technical challenges. On the technical side, DCCs must create, test, and iterate on infrastructure such as data submission, validation, and processing pipelines. Socially, CF Programs often operate without any strong community. Data generators, DCCs, tool creators and others in the program all have seperate grants, and little or no incentive to work together. DCCs often find that although their mandate is to collate and curate data, they instead spend the vast majority of their time trying to persuade data generators to regularly submit data and metadata, which alters both the type of data that the DCCs end up with and the rate at which they receive it. Many DCCs also reported that they were unexpectedly mandated to collaborate with other programs, and need to meet new technical challenges in their middle years. By this stage, the personnel and infrastructure actually needed to run the DCC may bear only passing resemblance to what was originally proposed and planned for. These lifecycle challenges can result in slow ramp-up, suboptimal infrastructure choices, delays in getting robust data pipelines in place, fragile software, and lost opportunities for data reuse.

Mature DCCs have considerable expertise, infrastructure, and other important assets that could potentially be used by other sites, but sharing their experience and assets is limited by both time and financial concerns. Although their budget has remained static, mature DCCs are now at their peak funding requirements: they have a full workforce, more data than ever, and a large number of users. There is also a great deal of uncertainty about what happens to a DCCs data once CF funding ends. For instance, detailed knowledge about data processing procedures is vitally important for maintaining the value and utility of the datasets. Unfortunately, thorough internal documentation is time consuming and difficult to provide and thus may not be a priority, especially if the middle years involved a great deal of pipeline fluidity and infrastructure change. As there are currently no standardized mechanisms for stable identifiers for data, or custodians to track those assets, there is no easy way to merge a DCCs data into another repository and end users are far less able to find those resources after the DCC is deactivated. The problem of storage costs for maintaining the data of deactivated DCCs have also not been solved by the CF programs, nor has the need for user help, software maintenance, web hosting considerations, or many other long-term costs associated with keeping a community resource alive.

Fairness Standards

While each Common Fund program has different mandates, communities, data types and pipelines, their data coordinating centers are all working towards the same basic goal: to coalesce that programs data into something broadly useful. In short, making data Findable, Accessible, Interoperable, and Reusable is the express purpose of the DCCs. So, it is unsurprising that every DCC staff person we spoke to was dedicated to the meeting the FAIR principles for their data. But while everyone agrees with FAIRness in theory, there is no one unified set of practices that DCCs can use to make their data more FAIR. Our analysis of the DCCs assets confirmed each individual group has reached a high level of FAIRness for their data; yet, there are no specific criteria that DCCs can use to converge on a common approach to FAIR across the entire CF portfolio, and how each DCC actually implements FAIR principles varies widely. This makes creating a single objective metric to measure or compare FAIR across programs nearly impossible. There are also few incentives for any given DCC to take on the monumental task of trying to create a unified set of standards, and to socialize them across the Common Fund; nor do any DCCs have the time or resources to do so.

Sustainability

A recurring theme from the CFDE's site visits in 2019 (see Appendices from this report, as well as the July and October reports) is the challenge of paying for data storage and compute. Issues raised include basic data storage, serving raw and processed data to internal and external users, and supporting the cost of small-scale and large-scale analyses with unpredictable compute volumes. These costs are fundamental to the basic utility of the datasets being generated by Common Fund programs and the inability of programs like the HMP and LINCS DCC to pay for continued data hosting represents a major threat to the continued utility of these datasets.

Cloud-based technologies do not eliminate costs associated with data download, storage, transfer and analysis: as one DCC PI noted, “Open Data is not Free Data”. If a goal of the NIH is to increase the number of users accessing data and analysis capability on cloud systems, possibly The biggest threat to the CF goal of increasing the number of users accessing data in the cloud is the absence of a unified, streamlined, and cost-effective mechanism to host and manage data across all the Common Fund centers. The nascent STRIDES program may help *reduce* storage and analysis costs for the NIH, but that does not necessarily lend itself to increased use in the community. End users are accustomed to downloading data from free resources such as dbGAP, and using it locally. Moving these resources to the cloud requires end users to either pay egress costs, which they seldom have funding for, or to work in a cloud environment, which few have the computational background to use, and which can be quite costly.

The exact model for how the STRIDES program applies to the CF DCCs also remains unclear, and the challenge of managing these costs in the long term is compounded by the lack of clear policies and metrics at the Common Fund for supporting data resources beyond the lifetime of the project. While the STRIDES program provides one avenue for consolidated billing that could be independent of an award, it is only being used by one Common Fund program (Metabolomics) of the nine we have interviewed. Regardless of the funding source, or the reductions made available through STRIDES, costs for data hosting should be managed across all sites using a comprehensive approach that includes considerations for data reuse by their end users.

Some DCCs noted that the use of on-premise infrastructure still remains an attractive model in comparison to cloud-based systems. As the data ecosystem continues to expand, the balance of cloud-based capability versus local (on-premises) computing, will continue to be complex decision that relates not only to the needs of each individual program, but also to the long-term sustainability of the data resources, and NIH plans.

User Training Needs

Most DCCs reported that user support takes up a substantial fraction of their time and effort. For some, like GTEx and Kids First, their support burden primarily comes from end users. Both programs told us that they have constant questions and help requests from their community. For other programs, the support burden is internal to the consortium. 4D Nucleome and Metabolomics, for example, devote most of their resources to helping their data providers format and submit data to the DCC. Nearly every DCC told us they would welcome both extra resources, and help with building more effective training.

Many DCCs were interested in expanding biomedical user support and training for users inside and outside the project. A range of specific motivations emerged from our site visit; these included increasing access to data, educating users in appropriate analyses, supporting more sophisticated data analysis, driving more integrative analysis, and identifying opportunities for enhancing DCC infrastructure and functionality. Despite this interest, relatively few DCCs have a significant training mandate or an active training program. The support burden that additional users would add to projects was also a significant concern. A related concern raised by particularly widely used resources such as GTEx was that many biomedical scientists need introductory bioinformatics training in order to use the DCC resources, but this kind of training is too broad in scope for any one project to tackle. Finally, many DCCs are interested in offering more training but even with funding would lack the expertise to develop materials, offer workshops, and do assessments.

Findability And Interoperability

The CF DCCs host a rich set of file types (e.g., genomic sequence, metagenomic, RNA-Seq, physiological and metabolic data) but it currently is not possible to discover these files across DCCs. Metadata, which is crucial to making use of these assets, is not readily accessible to users in a uniform manner. At present there are several interoperation efforts, including projects within the Common Fund (for e.g. GTEx/Kids First, MoTrPAC/Metabolomics) and outside the Common Fund (for e.g. GTEx/Kids First/ANViL, HuBMAP/HCA). Several DCCs, as well as some end users, have told us that the ability to combine dataset cohorts across DCCs is highly desirable as it would directly address important biomedical and clinical questions. Unfortunately, there are no unified practices for electronic formatting and transport of Common Fund data, and no standardized interoperable mechanism for transporting datasets between DCC portals, or to platforms such as Terra or Cavatica for data analysis.

Combining just two datasets is a huge challenge: a researcher needs to be able to first Find both of the datasets through a query system, and then gain access to both. She then needs to ensure that the two datasets have a compatible study design and metadata. Once she gains access to the raw data, she needs to manually check each metadata field and figure out how to make the studies match. This might include doing transformations such as changing pounds to kilograms in one study, and re-coding cancer/healthy to case/control. If done at the level of the DCCs themselves, this process would need to happen for every dataset pair across the Programs. A given DCC knows its own assets, but like any end user, would have to go searching through all the other individual portals, all of which use different terms, to even begin finding data to interoperate with. Currently there is no common portal for CF data, so it is virtually impossible to easily across all Common Fund assets and determine whether any given dataset exists.

It is important to note that most of these challenges are scientific and social challenges rather than technical ones. Technical solutions like transitioning to cloud storage can reduce the time and costs associated with moving and analysing data, and building a centralized search portal would greatly improve Findability, but these address only a tiny fraction of the interoperability challenge. The true difficulties lie in making scientifically sound decisions about topics such as whether two datasets should be combined, what metadata are compatible, and what metadata terms are ‘important’. Staff at DCCs frequently emphasized that the substantial cost of increasing interoperability should be justified with well-defined use cases, and conversely some data custodians were concerned that facile integration of datasets could lead to mis-interpretation of that information.

Human Subject Data

In addition, DCCs have reported their users are faced with several barriers to accessing dbGaP data. The practices and infrastructure of dbGaP are fairly confusing. For instance, dbGaP renames files when they are deposited, but doesn’t inform the DCC of the new names. Similarly, obtaining access across several clinical studies is cumbersome, for e.g. accessing the data for phase two of the Human Microbiome Project would require approval for twelve different consent groups. Another significant concern is that each DCC, has its own signon and authorization system, so end users need to sign up separately for every portal.

Engagement/Mobilization

During our listening tour nearly all of the PIs expressed enthusiasm for an additional level of coordination and interaction across the Common Fund DCCs. Several sites also expressed an interest in greater opportunity to collaborate with other groups in the areas of data integration, harmonization, resource and expertise sharing, as well as development of common training activities. These examples reflect not only a need to perform some type of common technical development across these groups, but also will require increased levels of engagement, social interaction, and increasing communication opportunities across all the DCCs. While there is some official interaction between DCCs - for example, Dr. Subramian, the PI of the Metabolomics DCC, serves on the SAB for LINCS and MoTrPAC - by and large the DCCs operate independently and do not communicate regularly at a technical or scientific level. Interestingly, both MoTrPac and 4DN have established informal relationships with ENCODE in order to adopt and adapt their workflows for transcriptome and epigenome studies. At present, 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 others. The DCCs have mostly developed tools and strategies to support their mission in isolation, and don’t have the time or resources to build new collaborations.

The Value Proposition for a Common Fund Data Ecosystem

A data ecosystem is a collection of data silos or commons joined together by a set of standards and services that facilitate findability, accessibility, reuse, and interoperability of datasets between silos/commons. A data ecosystem is focused on enabling multi-way connectivity between datasets, in a horizontal fashion, rather than deeper “vertical” analysis within each dataset. The goal of an ecosystem is to enable use cases between data silos, not within.

A key feature of an ecosystem is that improvements to one member of the ecosystem provides benefits to other members of the ecosystem, which implies a kind of interdependence. This is something that we can measure. It also implies that there is some kind of governance that provides for standards evolution and formal adoption of standards between members of an ecosystem; many of these standards are all concerned with FAIR.

Levels of maturity for an ecosystem progress through findability and accessibility of data, data reuse, and data interoperability. Findability requires a search or catalog mechanism, e.g. an inventory mechanism. Accessibility requires convergence on auth/auth so that datasets can be accessed across the ecosystem. Reuse requires workflows and compute. And interoperability requires mechanisms for syntactic and semantic exchange. Nothing about an ecosystem requires that all datasets must interoperate, and indeed most won't be able to. However an ecosystem should contain enough information on individual datasets so that the compatibility of those datasets can be determined, ideally in an automated way.

Common Fund Programs are creating and curating transformative, cross-cutting datasets, and are staffed by domain experts. Operating independently allows them to quickly adopt new standards, create new tools and protocols, and develop new workflows to meet the needs of their users. Adding interdependence and interoperability requirements to CF programs risks slowing down innovation and burdening already over-taxed data resource centers with additional work. However, interdependence does not require uniformity, and we see a number of ways that participation in an ecosystem would benefit all Common Fund Data Coordinating Centers, with minimal costs:

Improving Cross-common Fund Findability And Accessibility Of Datasets

There is substantial opportunity to increase the findability of Common Fund datasets by standardizing on a minimal asset specification, implementing practical FAIR metrics across the Common Fund, and engaging with NIH-wide access and authentication protocol development efforts such as RAS.

Managing Data Storage And Compute Costs

At the technical level, there are an increasing number of open source platforms that support both NIH- and non-NIH payments for analysis of cloud datasets without incurring egress charges, e.g. Broad's Terra (in use by GTEx and ANViL) and Seven Bridges Genomics' Cavatica (in use by Kids First). Evaluation and/or adoption of these platforms represents a specific technical opportunity to solve problems across current Common Fund projects.

Opportunities for Cross-Pollination Between Common Fund DCCs

There are many opportunities to connect between scientific questions and datasets as well as share infrastructure and coordination solutions. Many of the datasets may be complementary, if the metadata and analysis pipelines can be harmonized. Most of the DCCs use similar open source data analysis software systems (R and Python/Jupyter) and are using GitHub to distribute workflows. Almost every DCC is using Google or Amazon cloud hosting for data and compute.

Regular connections between DCCs could result in the sharing of solutions and enhancement of expertise in all of these areas. These connections could be facilitated by a centralized team who can field specific requests for information, as well by informal cross-pollination events to discuss shared technical and scientific goals. Increased interactions between all groups would lead to increased specific DCC-to-DCC projects involving data integration, or harmonization, as well as creating a network for sharing resources that can be used by other DCCs.

Opportunities for User Support and Training

There is a significant opportunity to increase usage of CF resources by supporting training across the CF as well as within individual DCCs. Training options could include material development for workshops, executing in-person and online workshops, offering and recording webinars, providing MOOCs, and doing hackathons for advanced users. Data reuse fellows and summer programs for undergraduates could further increase the usage and accessibility of CF data. Paired assessment programs could be developed to evaluate, iterate, and improve upon training offerings.

Separately, if resources become easier to find across the CF, a centralized help resource and tier 1 helpdesk could provide a first line of engagement to biomedical scientists seeking to find and integrate CF data assets.

Preparation For The Future

Several projects are underway to establish cloud-based data platforms across NIH (e.g., NHGRI-Anvil, NHLBI Bio Data Catalyst, and NCI-TCGA), and the Common Fund recognizes the need to prepare for a future of federation and interoperation. Adoption of standards for federation across all of NIH have yet to emerge but we can be certain that integration with data hosted by other ICs is of critical importance. Sequence based technologies for variant detection, whole genome/exome analysis, single cells and human cell atlas, as well as epigenomic analysis will increasingly be used by future Common Fund programs, as well as programs at other ICs. If properly linked to CFDE assets, the data available from these many projects, will create a synergistic data network for the research community. To realize this network, it will be important to collaboratively develop national and international interoperability efforts, and to encourage the use of these new standards throughout the Common Fund.

The Charter for the Common Fund Data Ecosystem Coordinating Center

The Common Fund DCCs have primarily operated in isolation from each other. There is a valuable opportunity to leverage the expertise found at the DCCs, increase engagement among the DCCs, and to develop and share best practices across this network. The CFDE Coordination Center has been established to help organize CFDE activities, engage with participating Common Fund programs, connect with the DCC user communities, support training, develop tools and standards, and provide technical expertise.

Over the coming year, the Common Fund Data Ecosystem Coordinating Center (CFDE-CC) will improve growth in use and reuse of CF program data by supporting DCCs that participate in the Common Fund Data Ecosystem. Building a new Common Fund Data Ecosystem would begin to address many of the challenges currently faced by the Common Fund DCCs, however it will require a great deal of technical and social work. The technical solutions needed to integrate data will depend on intelligent standards that are broadly agreed on by the ecosystem community, and communities themselves require a surprising amount of management and effort to sustain. DCC staff are the experts on their data, and well equipped to make these decisions, but lack the time and resources to take on these additional burdens. By taking on the much of the administrative burden of community management, while simultaneously providing resources for DCCs, the CFDE-CC hopes to reduce the barrier to entry for DCCs joining the ecosystem.

The CFDE-CC will work to connect external users to CF data, identify unmet needs within CF programs, connect infrastructure experts across CF programs, and lower the user support burden for individual programs. The CFDE-CC also plans to support CF datasets across their entire lifecycle, from startup to spindown.

The CFDE-CC is also defining and measuring FAIR, to guide systematic improvement of Common Fund asset FAIRness. One of the CFDE-CC's missions is to guide improvement of Common Fund asset FAIRness by providing consistent definitions, metrics, and reports across the CFDE. By applying the same objective measurements to each Program, we will establish an even playing field across all of the sites. This will incentivise sites to improve individually and learn from each other, and at the same time will lead to a more specific, consistent, and sophisticated set of FAIRness metrics for the CFDE. More importantly, the improvements to each site and across the ecosystem will enhance user abilities to find and make use of Common Fund data.

The charter of the CFDE-CC is to realize a vision of improved understanding and more rapid translation to cures from the Common Fund investment, by building an ecosystem that connects datasets and people across the Common Fund and beyond. We will start by making the existing data more findable, accessible, and reusable. Ultimately we hope to drive linkage *across* datasets by enabling researchers to interoperate between existing datasets, and do integrative analyses. We will tackle both the technical challenges involved in increasing data use and lowering collaboration barriers, and the social challenges around incentivizing cooperation and collaboration. Part of our mission will be to increase the sustainability of CFDE data over time by developing metrics and creating end-of-lifecycle approaches.

What Is The CFDE -- CFDE-CC Implementation

Federation

Each of the DCCs host many assets (files) (e.g., genomic sequence, metagenomic, RNA-Seq, physiological and metabolic data) and it is hard to discover these assets 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. To improve federation, the Common Fund Data Ecosystem (CFDE) will be based on a collection of inventories derived from data that are being hosted on cloud based systems by a number of DCCs. The inventories will describe all the assets at each Program. As part of the ecosystem the Coordinating Center will make these inventories available via a central catalog to enable discovery of the assets across the DCCs. The advantage of this approach is that formation of the ecosystem does not require the data assets themselves be moved to a central repository, only the inventories describing those assets. Cataloging all of the Common Fund assets is a simple and effective means of liberating data from what would be several siloed repositories, and therefore greatly increases Findability, Accessibility, Interoperability and Reusability of *all* Common Fund data. This form of data federation can also be extended to programs funded by other institutes, and easily linked to other NIH ecosystems: once an inventory system is available, it can be used by anyone.

Common Fund Data Asset Specification

We will simplify discovery of assets hosted at the DCCs by creating a specification of a minimal set of descriptors for each of these files, and electronically encoding this information into a common format. 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-CC team. The types of files (e.g., genomic sequence, metagenomic, RNA-Seq, physiological and metabolic data) that are referenced with the Common Fund Data Asset Specification will be flexible, and our current specification contains a small number of essential elements such as: a Global Unique IDentifier (GUID), Originating institution (e.g., "Broad Institute"), Assay type (e.g. “whole genome/exome”, “transcriptome”, “epigenome”) File type (e.g. "fastq", "alignment", “vcf”, “counts”), and tissue source and species name for the sample. The data asset specification also enables us to use readily available internet technologies to get additional information for each asset, such as metadata (e.g. patient variables, project name), and to resolve access issues such as files being hosted on the cloud or local servers.

Common Fund Data Asset Manifest

The ecosystem will support the concept of a “Manifest” that describes a collection of files. The manifests enable bundling lists of CFDE data assets into a machine-readable file using a common format. Manifests will also be used to publish the complete inventories of data from each DCC, and will enable uniform collection of asset metadata, and to support indexing of the assets in the 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 may not be adopted by the broader data resource community, the CFDE project represents an excellent opportunity to drive creation of a standard for all of the Common Fund DCCs, and we expect this approach to be compatible with other federated systems (e.g. GA4GH) as they emerge.

Common Fund Data Asset API

Another important element of the ecosystem will be the standardization and publishing of an application programming interface (API) that can be used by data consumers to retrieve the inventories, the data asset specification, and additional metadata associated with the assets. This will allow for consumers of these inventories to be able to programatically interrogate the federated system for information that may be relevant to a consuming service.

The CFDE Portal

We will provide a portal enabling users as we as administrators to search all of the federated data assets at each Common Fund Program. The CFDE portal will increase a user's ability to find these important resources, as well as mix and match sets of data from each site to use in subsequent analysis. We refer to lists of assets as manifests, which are similar in function to a shopping cart on a commercial web site. Generation of user-specified manifests will enable users move information off the portal for use in the analysis tool of their choice. Other important functions will be developed in the portal over the next year. End users will be able to answer the question: "where are all the RNA-Seq datasets associated with all Common Fund programs?". Similarly, Administrators and Program Officers at Common Fund will be able to go to a single website and view the growth of data from their program over time, to review objective FAIR metrics for these assets, to understand download statistics and geographic distribution, and view the degree of harmonization of these data in comparison to other sites.

Functionality of the portal will be expanded to include additional usage information from each of the programs. For example, we plan to request and display portal metrics such as the number of users that register at each of their sites, and how often their data is downloaded or analyzed. Once this capability is established, an important outcome of the CFDE will be to give Common Fund leadership the new ability to objectively review the overall use of resources at each data center, and to easily perform that review in comparison to all other Common Fund data centers. We anticipate this type of information will assist in making better informed decisions with respect to maintaining and prioritizing which Common Fund datasets over time.

Fair Metrics

Under the CFDE, each data center's inventory will be evaluated consistently based on FAIRshake, and the Coordinating Center will work with the individual Common Fund programs to adjust FAIR measures to meet the needs of the Common Fund. This approach overcomes a major obstacle for the Programs, because the Programs can not easily work with other groups to align around a common set of metrics.

In a pilot study of data of 7 CF DCCs we employed FAIRshake (https://fairshake.cloud) to evaluate the FAIRness of digital objects including datasets, tools, and repositories. This required mapping metadata elements from each DCC to FAIR metrics according to a customized rubric that was created from the list of case studies developed by CFDE members. The conversion process was essentially a manual process requiring customized scripts and a subjective selection of metadata elements that were mapped to pre-existing ontologies and controlled vocabularies. Data transformation was performed using the Frictionless Tabular Data Package, a simple electronic format used to describe a collection of data (https://frictionlessdata.io/data-packages).

Retrospective evaluation of FAIRness for Common Fund data is of limited value. What will be far more effective is for the CFDE community to create an agreed upon set of metrics across all of the DCCs, and to use these metrics to drive up overall levels of FAIR data. The Coordinating Centers goal is to create consensus based machine-readable standards that will provide quantitative and verifiable FAIRness measures of all Common Fund data. We will achieve this by working with the DCCs to create inventories of DCC data, prioritize which metadata elements should be evaluated, and applying openly accessible rubrics to data hosted at each site. Each participating site will be able to review their inventories prior to public release at the CFDE portal, and portal web pages will that summarize the CFDE FAIR metrics will be published on an on-going basis.

Develop CFDE Best Practices

In order to achieve its goals the CFDE-CC 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.

Policy Development And Technical Implementation

There are multiple examples of technical solutions that could favorably impact all of the ecosystem members of the CFDE, that also require elements of administrative coordination with these efforts. For example, Common Fund leadership has partnered with the STRIDES initiative, which provides lower-cost cloud services to NIH projects. The CFDE-CC will help ensure the technical implementation of the Common Fund Data Ecosystem resources are tailored to enable each DCC site to be able to take advantage of the STRIDES cost-reduction program. The Researcher Auth Service (RAS) is a service under development by the NIH's Center for Information Technology that will facilitate access to controlled data assets and repositories. Members of the CFDE have also partnered with Globus to enable researchers with an eRA-Commons account to use their credentials to simplify access to controlled data assets. The CFDE-CC will continue to advance this initiative in the coming year, and provide guidance to the CF DCCs in order to make use of the RAS system.

Training Program Development

The CFDE Coordination center will also host a Training Coordination Center (TCC), staffed by experts in bioinformatics curriculum development, teaching, and community building. This center will provide support and resources for the development of DCC specific training programs as well as end-user training on CFDE products and general topics of interest to the Common Fund research community. DCCs will be able to request personalized assistance with all aspects of designing, piloting and refining bioinformatics workshops or webinars. The TCC can also help with logistical support for hosting workshops, as well as providing guidance on how to grow and build a sustainable training program. As part of this effort, the TCC will provide instructor training for the DCCs, and assist with creating useful qualitative and quantitative feedback and assessment tools. In addition to site-specific training, the TCC will also offer training on CFDE products as they become available, and will pilot a ‘general bioinformatics’ workshop curriculum on topics of broad interest within the Common Fund.

DCC Cross-Pollination Events

Individual DCCs have significant expertise in complementary areas, and the CFDE-CC will facilitate conferences to bring DCC personnel together in person to discuss their technological challenges, approaches, and solutions. Annual conferences would serve as an avenue for building cross-DCC collaborations and discussions, and identifying complementary expertise and technologies across the DCCs.

Increased Software Reuse

There are community-based open source software and data projects where engagement could increase the software and data analysis capacity provided by Common Fund projects. Software projects include JupyterHub and BinderHub for data analysis, and the R and Python data science ecosystems; all of these are used by current CF programs. Working with these projects to facilitate broader use of the software within the CF and broader distribution of the data outside the CF is a straightforward opportunity where minimal investment could reap many benefits.

Platform For Innovation

The DCCs are brimming with ideas that could revolutionize the Common Fund. For instance, Metabolomics told us that we underestimate the breadth of diversity of the CF data, and warned that we have become complacent about how exciting it would be to link between datasets to better understand biology. They have a collection of cross-program use cases that span Common Fund and that could open new avenues of research. Similarly, SPARC suggested a more universal way to share across a social network that could be transformative. If implemented, the CFDE could serve as an incentive to increase cooperation, be easier to share with a colleague, easier to get analysis services, add value is added to the data, and lower collaboration barriers underestimate the breadth of diversity of the CF data. The CFDE-CC hopes to provide a forum for discussing, vetting and securing resources for these game changing ideas.

Summary

During 2019, members of the CFDE-CC built an understanding of the concerns faced by the Common Fund DCCs. The DCCs act as custodians of biomedical data and possess a rich amount of expertise and dedication to their research community. They face a related set of challenges, but despite this have largely been operating in isolation of each other. By instituting a team with experience in facilitation, engagement, and technology to mobilize across the Common Fund DCCs, we can grow a comprehensive data ecosystem. We will also be better prepared to work with efforts like the NIH Interoperability working group, that represents four ICs (NHGRI, NHLBI, NCI, and Common Fund) as well as the NIH Office of Data Science Strategy to create a common authorization system. Collectively we will form a network across the Common Fund DCCs, CFDE-CC and Common Fund leadership will work more cohesively and rapidly to provide increased capability to NIH biomedical researchers.

Appendices

Appendix A: 4D Nucleome Deep Dive

4D Nucleome Site Visit

Location: Harvard Medical School. 302 Countway Library

Date: Thursday, October 17, 2019

Attendees: Representatives in attendance from the CFDE were Amanda Charbonneau (UCD), Alex Waldrop (RTI), Titus Brown (UCD), Owen White (UMB), Brian Osbourne, and Anup Mahurkar (UMB). The representatives from 4D Nucleome included Peter Park (PI), Burak Alver (Scientific Project Manager), Andy Schroeder (Senior Data Curator), Koray Kirli (Data Curator), Sarah Reiff (Data Curator), and Luisa Mercado (Data Curator).

Meeting Logistics

We held a meeting with the 4D Nucleome (4DN) Network Data Coordination and Integration Center infrastructure team at the Countway Medical Library at Harvard Medical School on Thursday October 17, 2019 to discuss their ongoing program. During the meeting, we used the agenda at the end of this document as an informal guide to structure our conversation and address key issues.

The engagement team began by introducing themselves and their goals for the meeting. These goals included learning more about:

- Structure, vision, and goals of 4DN
- Platform stakeholders, important users, and common data types
- Information about training and organization
- Ongoing and upcoming organizational challenges
- Overall set of priorities

After reviewing the day's agenda and objectives, the 4DN team provided an extensive overview of the overall 4DN program goals as well as the 4DN DCIC goals and products. The overview touched briefly on the scientific motivation behind 4DN and how the DCIC's current data platform has evolved to serve the needs of program stakeholders. Follow-up conversations led by the engagement team focused on understanding the organizational structure, future goals, ongoing challenges, and potential productivity bottlenecks experienced by the 4DN team.

The engagement team finished by presenting their high-level preliminary vision for how the CFDE might support member DCICs like 4DN in the future. The day concluded with a more specifically tailored brainstorming session of concrete ways the CFDE might support and add value to the 4DN team's ongoing and future work.

4D Nucleome Overview

The broad goal of the 4D Nucleome program is to study the three-dimensional organization of the nucleus through space and time (the fourth dimension) in order to better understand how changes to this architecture impact biological function. The program grows from an increasing realization that structural variation in the nucleus (e.g. chromatin structure) plays a critical albeit poorly and likely misunderstood role in overall biological function. The high-level goals of the 4D Nucleome program are to

1. Investigate the functional role of various structural features and nuclear processes,
2. Develop, benchmark, validate, and standardize next-generation technologies for investigating nuclear organization in 4 dimensions, and
3. Develop an open data platform and set of data and data analysis standards for housing, harmonizing, processing, and distributing the diverse array of imaging (e.g. FISH, ChromEMT) and molecular data (e.g. HiC, CHiP, ATAC) pioneered by and generated through 4DN program centers.

Operationally, 4DN is a technology and research development network comprising 29 partner centers organized into 7 larger initiatives addressing one or more of these core goals (shown below).

Center	Member Organization(s)	Responsibilities
Nuclear Organization and Function Interdisciplinary Consortium (NOFIC)	6 centers	Develop, benchmark, standardize, and validate high-throughput technologies that can produce three dimensional physical and functional maps of mammalian genomes
4D Nucleome Imaging Tools	9 centers	Develop, benchmark, standardize, and validate imaging technologies for visualizing structural and functional organization of the mammalian genome
Network Data Coordination and Integration Center	1. Harvard Medical School 2. Washington University (visualization tools)	Collect, store, curate, and display all data, metadata, and analysis tools generated by the 4DN Network. Develop data, metadata, and analysis standards
Network Organizational Hub	University of California San Diego	Coordinates activities across 4DN centers and teams

Study of Nuclear bodies and Compartments	6 centers	Investigate 1) topography of nuclear bodies and transcriptional machineries, 2) structure and function of poorly characterized nuclear structures, 3) role of specialized proteins and RNAs in the assembly, organization, and function of nuclear bodies
Nucleomics Tools	1. The Babraham Institute 2. California Institute of Technology 3. Baylor College of Medicine 4. Cornell University 5. University of Pennsylvania	Develop and validate physical, chemical and biochemical approaches for measuring properties and dynamics of the 3-D organization of the genome

As described in the table above, the majority of 4DN network centers, including those within the NOFIC, Nucleomics Tools, Study of Nuclear Bodies, and 4D Nucleome Imaging Tools consortia, are responsible for generating high-quality datasets with standardized metadata annotations using state-of-the-art technologies to characterize, quantify, and visualize various spatial features of genomic architecture. As part of this work, 4DN centers are also responsible for developing, benchmarking, validating, and standardizing the next-generation technologies they employ as part of their data generation efforts. Example technologies encompassed by 4DN include emerging molecular methods for quantifying chromatin spatial structure and genomic interactions like Chromatin Conformation Capture (3C) and HiC, as well as cutting edge imaging techniques like ChromEMT and Electron Tomography for providing high-resolution 3-D visualizations of chromatin structure and sub-nuclear cellular components. In total, 4DN network centers collectively generate nearly 30 distinct data modalities characterizing various aspects of the 4-D genomic architecture across a range of experimental conditions (e.g. heat shock, various CRISPR modifications) in human, mouse, and drosophila cell lines.

Data generated by 4DN partner institutions are integrated, curated, analyzed, and disseminated by the 4DN Data Coordination and Integration Center (DCIC). As part of this work, the 4DN DCIC develops and maintains a web-based portal (https://data.4dnucleome.org/) to support data submission from 4DN network centers, and provides tools and support for data access, visualization, and analysis to the broader community. The 4DN DCIC is also tasked with leading the ongoing data integration and harmonization efforts across submissions from the program's 29 partner centers. These efforts include working directly with data generators to define and develop standards for 1) routine data analysis/processing pipelines, 2) experimental metadata and definitions, as well as 3) file format standards for raw/processed genomic and image data files.

Program Lifestage

The 4DN program started in 2015 and is currently in its final year of Phase 1 funding, which is set to end in 2020. After receiving initial funding in September 2015, the 4DN DCIC spent much of its first year conducting a somewhat difficult hiring process. They had hired a core team of 3 developers and 2 data curators by June 2016, and the remainder of 4DN's first year was spent engaging the broader 4DN network through working groups tasked with establishing the program's data sharing policies and standards for cell line work and metadata terms. The DCIC released an alpha version of the 4DN portal in October 2016. The 4DN DCIC team also began an ongoing collaboration with the more established ENCODE DCC to learn how their group had addressed many of the same issues 4DN was facing. As a result, much of the current 4DN platform is built on tools first developed by or in tandem with ENCODE teams. Over the next year, 4DN DCIC team members led working groups tasked with developing omics data analysis and imaging standards that continued to guide platform design and development. The DCIC released a beta version of the platform in 2017. A year later, a production version of the 4DN web portal was officially released in 2018.

In that time, 4DN network centers have generated a tremendous amount of data resulting in the publication of nearly 200 peer-reviewed publications. Since the platform's official release, the 4DN web portal has grown to house data from over 700 studies, encompassing 27 data modalities across more than 2,300 separate experiments. The 4DN DCIC infrastructure team attributes this growth to the active and often interactive role they play in soliciting and facilitating data submissions from across the 4DN network.

Most of the day-to-day work at the 4DN DCIC is in direct service of the 4DN network. This includes processing user submissions through standardized pipelines, expanding the data platform to integrate evolving protocols and metadata terms from collaborators, assisting 4DN network partners through the data submission process, and engaging the 4DN network to develop imaging and omics data analysis standards. Interestingly, the 4DN DCIC team says much of this work focuses on simply trying to convince 4DN network partners to submit their data to the platform. Continued work on the 4DN platform focuses on adding features and services to incentivize 4DN network partners to host data on the platform, such as data processing services through the 4DN analysis platform, and data visualization tools for emerging data types like HiC.

Ongoing challenges faced by the 4DN DCIC stem largely from the program's size and the diversity and complexity of the techniques under development. Data generated by 4DN network centers are hypothesis driven, which has led to the exponential growth of an increasingly sparse matrix of experimental conditions that need to be defined by 4DN data curators and incorporated into the data platform. Complicating this matter, there is an increasing lack of consensus among 29 institutions, who continue to develop their own internal protocols for shared technologies. The submission process is becoming increasingly complicated by the growth and complexity of the underlying metadata model data to the point that contributors almost always require direct assistance from the data curation team; to date exactly one user has been able to submit data on their own.

Despite these challenges, the data portal continues to grow and 4DN is beginning to navigate the transition from Phase 1 to Phase 2 funding. At present, the 4DN DCIC team is currently preparing their RFA submission for Phase 2 funding that would last until 2025. The team expressed concerns over navigating the impending funding uncertainty for Phase 2 and a potential funding discontinuity between Phases 1 and 2. Even if the current 4DN DCIC team receives Phase 2 funding, they don't know when those funds would become available, or how they would cover cloud storage and personnel costs in the interim. In spite of this uncertainty, the 4DN DCIC team continues to develop their state-of-the-art platform and provide a high level of support to their user-base as they navigate this period.

Data Platform

Infrastructure

The 4DN DCIC web portal and underlying data platform are hosted entirely on the Amazon Web Services (AWS) cloud platform. The 4DN infrastructure primarily utilizes cloud storage (e.g S3) for datasets hosted through the web portal, and on-demand computing resources for executing data processing pipelines.

The 4DN web portal and larger data platform are undergirded by several modularized services integrated through the platform's data API. At the core of this architecture, 4DN uses SnoVault--an object-storage system developed by the ENCODE DCC that combines ElasticSearch with a PostgreSQL backend--to manage metadata and data file objects on cloud storage, display metadata statistics on the web portal, and power the web portal's search functionality. 4DN leverages SnoVault's RESTful API across it's platform to provide a single, standard interface for both authenticated external users and internal software components (e.g. web portal, analysis platform, data ingestion pipelines) to access 4DN data and metadata.

As an added benefit of the collaboration between 4DN and ENCODE and their shared use of SnoVault, many services are automatically cross-compatible between the two platforms. For example, visualization tools on 4DNs web portal can be augmented with metadata tracks directly from ENCODE. It's unclear whether this interoperability would easily extend to any Data Coordinating Center (DCC) using SnoVault, or whether this is merely a by-product of the similar underlying data types and metadata terms shared across 4DN and ENCODE. At the very least, this successful cross-pollination should merit closer inspection as the CFDE looks to facilitate greater interoperability across Common Fund Programs.

To ensure the stability of its platform, 4DN developed a tool called FourSight to manage, monitor, and maintain the network of persistent AWS resources that support the 4DN web portal and data platform. FourSight provides the 4DN DCIC with automated monitoring of the 4DN web portal's underlying network of web, database, API, and data ingestion servers on AWS that run the 4DN web portal.

4DN has also developed a robust infrastructure for reproducible data analysis on AWS. The core of this infrastructure is built on Tibanna, a stand-alone open source tool developed and released by the 4DN DCIC for automated workflow execution on AWS. Though Tibanna supports a number of workflow languages, a motivating factor behind the tool's development was the lack of open-source tools supporting CWL execution on AWS. Tibanna automates workflows by orchestrating the provisioning of cloud computing resources, transferring inputs from cloud storage, executing workflow steps in dockerized environments, and saving processed output on cloud storage. With Tibanna and AWS, the 4DN DCIC can reproducibly automate complex data processing pipelines on-demand at virtually any scale.

Analysis

The 4DN DCIC team develops and maintains a range of vetted data analysis pipelines for use by 4DN network members. Currently these include pipelines for Hi-C, CHiP-seq, ATAC-seq, and Repli-seq data processing. The 4DN web portal does not provide any tools for pipeline execution, but 4DN network members can have their data processed by the DCIC team upon request after data submission. 4DN's data analysis pipelines are implemented in CWL and fully Dockerized to ensure 100% analysis reproducibility and portability. They opted for CWL over similar workflow specification languages (e.g. WDL, Snakemake) based on their preference for the stronger, more explicit I/O typing it provides. They use Tibanna to execute CWL workflows on AWS, and CWL workflows are version controlled through GitHub and integrated with SnoVault. Upon successful execution, processed data files become available through the web portal, where users can also view the data provenance of processed outputs.

4DN provides a few smaller tools for data analysis and visualization through the web portal. Among these, the 4DN DCIC created an open-source tool called HiGlass for visualizing very large contact matrices from Hi-C experiments. HiGlass is fully integrated into the web portal, and is also available as a stand-alone web-page (https://higlass.io). 4DN also provides access to a beta version of its 4D JupyterHub service through the web portal. The tool provides users a workspace integrated with 4DN data that currently only supports very small analyses. A future goal for this service is to provide a fully functional analysis environment to allow users to work closely with 4DN data without having to download anything.

Access

4DN users can download data manually through the web portal or programmatically via their API service. Users with data submission privileges can submit data through both the web portal and a stand-alone python application they provide. Users can also access data from their 4DN centers which is not yet for public release using their credentials via the web portal, the API, or JupyterHub.

Users can register and login with GitHub or Google accounts for additional services like the 4DN JupyterHub. They use an OAuth-based user-permission system to define access to platform resources, but most services are available to everyone. This is mainly because the 4DN web portal currently does not provide tools for more extensive analysis which would incur larger compute costs and require more controlled access to limit spending.

Harmonization and Metadata

In tandem with 4DN network working groups, the DCIC has defined its metadata structure to describe biological samples, experimental methods, data files, analysis steps, and other pertinent data. The 4DN data model is based largely on the framework developed by the ENCODE DCC. They use established ontologies to define metadata terms where possible, including Uberon for anatomy and tissue types, and EFO for cell lines and experimental methods. 4DN also uses NCBI taxon IDs and EntrezGene IDs to support further interoperability. The DCIC curates an internal 4DN controlled vocabulary to provide definitions for emerging technologies like HiC and some cell lines used by 4DN network partners. Where applicable, the DCIC submits controlled vocabulary terms to EFO for future inclusion.

The DCIC maintains a team of 4 data curation specialists that work closely with network partners to guide them through the submission process. This typically includes helping users select appropriate metadata terms, and increasingly has included working with partners to develop new controlled vocabulary for emerging techniques.

4DN uses version-controlled CWL workflows to define standardized data processing pipelines for common analysis. Most of these are developed through 4DN network working groups or collaboration with the ENCODE DCC. Though the DCIC has the infrastructure to process user submissions through these pipelines, many labs opt to use their own internal pipelines instead. There does not appear to be an obvious solution to this problem, but the DCIC does publish its pipelines through GitHub and the 4DN web portal to indirectly facilitate standardized data processing across the network.

The DCIC continues to support working groups developing imaging standards and metadata definitions, but these areas have proved more challenging. While the imaging working group has been led by community members who support detailed metadata collection, the data curation team has received pushback from network members over conforming to such standards. The DCIC says most labs have different microscopes and lab techs typically do not record the kinds of details that might help develop metadata standards for uploaded images. In some cases, 4DN partners refuse to even submit image data to the web portal because they don't want their results to be mis-interpreted. The DCIC has started focusing on working with a subset of the more collaborative labs to gain some consensus on microscopy standards and metadata terms, but reproducibility issues even among this small group raise the question of whether standards are even appropriate this early in the development of these techniques.

Sustainability

Major concerns over the sustainability of the 4DN DCIC stem from cloud computing costs. As it nears the end of Phase 1 funding, the DCIC has been increasingly impacted by rising cloud storage costs resulting from exponential data growth as their platform continues to mature. The team was excited to learn about discounts available through STRIDES--which they had not yet heard about--and thinks the more targeted support would help relieve the strain of mounting infrastructure costs to some degree. They also think these issues could have been solved even without additional funding or discounts if they had been allowed to carry over unused infrastructure funds from earlier years when storage costs were low to cover the higher storage costs of more recent years.

Long-term cloud computing costs present a more existential threat to the sustainability of the 4DN DCIC. The most important question that remains to be addressed by the Common Fund is what will happen to data hosted on cloud storage at the end of the program's funding, and how these costs will be covered in the interim between Phases 1 and 2. Without substantive changes to the way Common Fund Programs are supported after the 10-year limit on Common Fund grants, there's a chance the 4DN portal will simply cease to exist 6 years from now.

4DN also faces long-term questions over data stewardship. As with other DCCs, it's unclear what personnel support will be available to maintain the DCIC's infrastructure after the 5 or 10-year funding is up. Without continued support for the AWS infrastructure that hosts the web portal and metadata database, 4DN data will become effectively inaccessible regardless of whether there is a long-term solution for cloud storage.

Training

Internal

The 4DN DCIC team engages the broader 4DN network primarily through working groups addressing key program areas. Since the program's inception, the DCIC team has led 5 working monthly groups (below) to discuss program standards and disseminate information across 4DN network partners.

Working Group	**Task**
Policy	Develop data sharing policies
Samples	Pick cell lines, develop protocol standards, define metadata terms
Omic and Data Analysis	Define standard analysis pipelines
Joint Analysis	Continuation of Omic and Data Analysis Work
Imaging	Develop imaging protocols and standards

The 4DN DCIC stressed several times during our meeting the difficulty of coordinating information and building consensus across the program's 29 centers. To supplement the efforts of working groups, the DCIC data curation team provides one-on-one support to network collaborators through the data submission process. The DCIC also provides additional channels for 4DN network partners through a helpdesk email service, protocols.io, and a feedback service on the 4DN data portal. Internal training within the DCIC is centered on weekly meetings for the data curation and development teams.

External

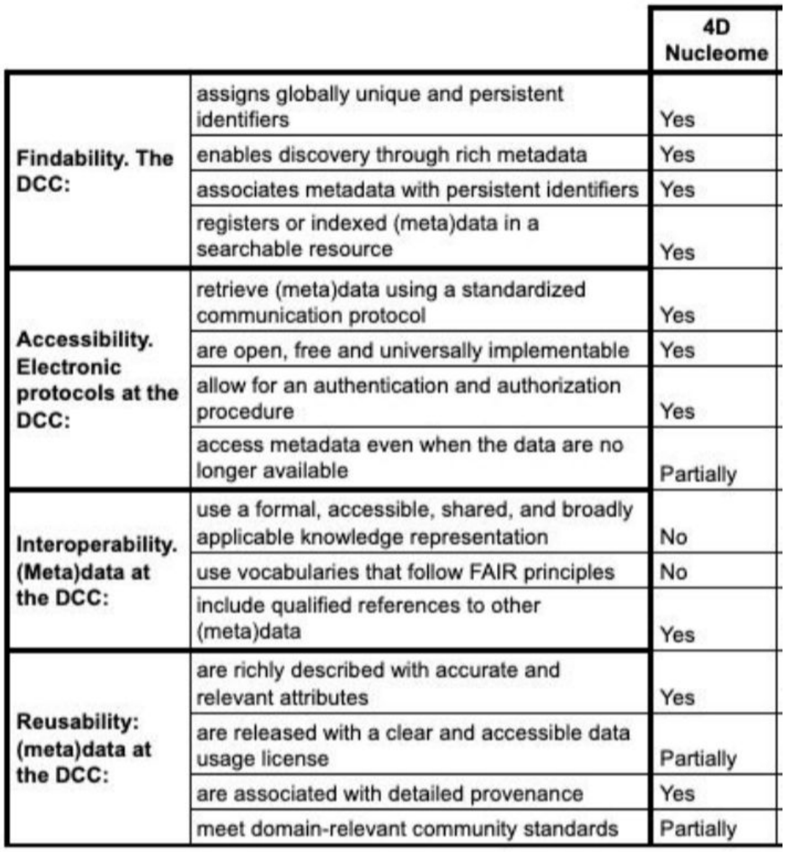
The 4DN DCIC primarily supports external users by providing high-quality documentation and self-guided resources on its web portal. The DCIC currently provides extensive tutorials to familiarize users with the platform's API, metadata model, and data submission pipeline. The DCIC also makes 4DN pipelines and metadata terms available to external users through the web portal.

The 4DN DCIC would like to provide more interactive training resources for its external users, but this has largely taken a backseat to the increasing demands of the 4DN network. Burak Alver (program manager) acknowledged this dynamic during our conversation: "We should be doing more bootcamps, videos, webinars, but do not have bandwidth to do this." The team has been able to lead occasional training bootcamps for HiC data analysis at conferences despite these resource limitations. Given their ongoing obligations to the 4DN network, the DCIC team doesn't see an easy solution for increased training beyond additional funding for dedicated personnel. Not surprisingly, additional support for user training was the top thing mentioned when asked what they could do with additional support through CFDE.

FAIR

The 4DN DCIC is dedicated to FAIR principles and continues to develop its platform in support of these standards. 4DN currently supports FAIR data access by providing a robust search interface enabling discovery through metadata, an API for accessing portal data, a largely ontology-driven metadata database, and data provenance though CWL.

Although the 4DN platform scored highly on the Common Fund’s FAIR assessment (below), they were surprised they didn’t have a perfect score. The DCIC expressed some confusion/concerns over the assessment itself, citing a disconnect between their interpretation of FAIR and the criteria on which they were being assessed. In particular, they said “Findability” was the most ambiguous/confusing component and has been difficult to interpret and operationalize. Interestingly, 4DN scored perfectly on Findability, which the team says has been the primary focus of ongoing FAIR-related platform development. The DCIC team also expressed similar confusion over the criteria used to assess platform interoperability. The assessment docked 4DN for potential interoperability issues stemming from the internal controlled vocabulary they maintain to define new metadata terms that don’t yet appear in existing ontologies. Contrary to the assessment’s findings, the 4DN DCIC team feel they do in fact use a formal knowledge representation, and say it still isn’t clear to them why their current setup was deemed insufficient.



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

The DCIC team was aware of some of these issues: “We do a good job with the F and the A and the R.” The team has experienced occasional interoperability issues due to changing metadata definitions. Issues with the 4DN internal vocabulary highlight the difficulty of both developing metadata standards for emerging methods and developing useful measures of data FAIRness. After all, it’s unclear how metadata being defined for the first time could be made more interoperable.

Overall, the 4DN team is in favor of greater interoperability among DCCs, but isn’t sure how useful this will be in practice. They were mixed on the idea of data re-use post-docs and generally wary of larger initiatives they feel often superficially combine data for the sole purpose of combining data: “the best science seems to come from researchers who have a specific question.”

Cross-pollination

The 4DN DCIC already collaborates heavily with the ENCODE DCC, but is open to future pairwise collaborations. The ENCODE DCC has effectively served in a “DCC mentor” capacity to the 4DN group, and 4DN continues to hold monthly calls with Ben Hitz who runs the ENCODE DCC. The relationship between 4DN and ENCODE could serve as a model for how to better support early-stage DCCs and foster greater interoperability. In addition to ENCODE, 4DN thinks it might be interesting for their platform to interface with GTeX to link changes in chromatin structure to gene expression. Beyond these pairwise interactions, the 4DN DCIC team doesn't see other immediate opportunities to integrate data from other DCCs on their platform.

The DCIC team was also generally enthusiastic about attending annual Common Fund cross-pollination events. They said they would be more interested if there was a specific focus or purpose being addressed. As an example, they suggested hosting CFDE mini-conferences where Program representatives could get together and present their solutions to commonly faced problems.

SSO (Single Sign-on)

4DN uses the OAuth authentication system for all security, including identify management, data movement, sign on for their portal, query API permissions, and the 4DN JupyterHub. OAuth uses Google and GitHub as authentication sources, and would not be considered meeting NIH requirements for SSO as they do not support ERA Commons or ORCID.

Outcomes

Infrastructure and Resource Reuse

4DN's infrastructure provides a good model for Common Fund DCCs opting for a cloud-based approach. Because 4DN is built largely on modular, open-source tools developed for stand-alone use (e.g. SnoVault, Tibanna, and FourSight), the platform's core services can already be reused by other platforms without any modification. For example, DCCs that want to execute CWL workflows on AWS can simply download and use Tibanna independent of the 4DN platform. The same can be said for SnoVault and FourSight. As a caveat, 4DN's software components are closely tied to AWS and in most cases not reusable across cloud platforms. In any case, 4DN's well-designed, reusable software infrastructure demonstrates the long-term value Programs can provide each other and the broader community through high-quality tool development. In particular, Tibanna is already being heavily used by the ubiquitous Snakemake open source workflow project, which is completely independent from 4DN.

Although 4DN's cloud-based infrastructure has worked well for them, it may not be a sustainable option for larger Programs with higher storage needs. 4DN's increasing difficulties with rising storage costs also highlight the funding issues cloud-based infrastructures can create. Unlike locally hosted DCCs where infrastructure spending is typically heavily concentrated in the early years of the program, cloud-based infrastructures incur most of their costs once the platform matures. This new dynamic may create more issues with the Common Fund's flat annual funding structure, as data growth and the resulting storage costs can make budget forecasts and future spending decisions difficult. Despite these challenges, 4DN largely views their infrastructure as a strength and think the benefits of cloud computing (e.g. on-demand scaling of computing and storage, streamlined development, no administrative overhead or start-up costs) have so far outweighed cost concerns. As the 4DN DCIC continues to mature, its platform will provide a test case for concerns over the long-term benefits of an entirely cloud-based infrastructure that should be used to inform future decisions by new Programs, the CFDE, and the Common Fund more broadly.

Challenges

The DCIC team highlighted a number of challenges it has faced over its lifetime in support of the large 4DN network, and we had a productive discussion regarding how the CFDE could provide future assistance.

Hiring and Retention. Since its inception, the 4DN DCIC has experienced ongoing issues with personnel recruitment and retention. Many of the challenges the DCIC faced during its ramp-up stemmed from a protracted hiring process that took more than 9 months to recruit enough quality developers and data curation specialists to begin platform development. 4DN says this problem was exacerbated by competition for highly-skilled technical personnel from places like Google and Facebook. Given the numerous other demands the DCIC faced as it began coordinating efforts across 29 centers, the limited administrative overhead and support available through Common Fund for hiring placed 4DN at a distinct disadvantage to industry competitors. Their experience highlights the need for more concentrated hiring support during the first year and beyond. They also highlight an ongoing need to make DCC positions more attractive to the talented personnel they require.

Training and Support. The amount of time and effort invested by the 4DN DCIC in support of its internal network leaves little time to develop external training resources and outreach programs. Because of this dynamic, the 4DN DCIC feels they operate less to provide a general public resource and more in direct support of a large research consortium generating a tremendous volume of data. The DCIC team feels the only solution at this point would be hiring more dedicated personnel to assist in these efforts. Because of this, 4DN highlights a potential role for CFDE in providing additional funding and/or personnel to support external training and outreach at Programs like 4DN that simply don’t have the additional bandwidth to support these efforts.

Hosting. The DCIC team has had issues getting network collaborators to host their data on the 4DN platform. They attribute this to some degree to the lack of incentives for PIs to release data before publication. In lieu of top-down enforcement mechanisms that likely wouldn’t be received well, additional outreach and assistance through CFDE for partner engagement might help PIs and their labs better understand the submission process and the larger value of the resource they’re helping to build. There is also the possibility that further incentives, like automated pipeline support through their web portal, might make hosting data on the platform more attractive for PIs. The CFDE could potentially play a role in helping cover increased computing costs to support data analysis.

FAIR Definitions. 4DN wants to use the FAIR principles but feels some of the definitions are confusing. In particular they think the Common Fund's definition of "Findability" could be more concrete and has been difficult to interpret and operationalize. The DCIC team was actually surprised they didn't score 100% on the Common Fund FAIR assessment, which highlights the need for community input on FAIR assessment criteria, providing greater access to FAIRness audits and check-ins throughout the development life-cycle of DCCs.

Flexibility of FAIR Principles. Also on the subject of FAIR principles, there is concern that if the CFDE attempts to normalize the implementation across different programs, it will create an unneeded hardening of the requirements, which will reduce the flexibility for the individual programs to implement FAIRness. 4DN's controlled vocabulary provides a good use case where FAIR principles may need to be relaxed or expanded upon in order to allow DCCs the flexibility to develop new metadata standards they think best describe their data. The CFDE could potentially support these efforts by facilitating more collaborative development of metadata standards across similar Programs that could lead to both richer data annotations and greater interoperability. The ongoing collaboration between 4DN and ENCODE serves as a good example, and highlights the potential value CFDE could provide through a more formalized "DCC mentorship" program to promote these collaborations.

STRIDES. While 4DN thinks discounts available through STRIDES will help ease the strain of rising cloud storage costs, they feel the current funding structure through the Common Fund is at odds with the exponential data and storage growth they've experienced over Phase 1. Until the web platform went live in 2018, 4DN storage costs were insignificant, and much of their infrastructure budget went unspent. As the portal has continued to grow, cloud storage costs have recently become a significant issue for 4DN, but as funds don't roll over, they can't now use those unspent funds from earlier years to cover current costs. 4DN thinks a more flexible/dynamic funding structure that acknowledges the reality of exponentially increasing cloud infrastructure costs over ramp-up may have prevented this issue even without additional funding. The CFDE could be instrumental in helping develop programs for more dynamic assistance to help cloud-based Programs deal with this problem.

Cloud Storage Costs. 4DN is also faced with more existential challenges stemming from long-term cloud-storage costs that the Common Fund is perhaps unprepared to address. In the immediate, 4DN has no idea how cloud storage costs with be paid in the interim between Phase 1 and 2. In the long-term, it's unclear what will happen to the entire 4DN infrastructure when the program sunsets. This hasn't been as big an issue for locally-hosted DCCs, where the majority of infrastructure costs are up-front, and where the worst possible scenario of walking away and never thinking about the data again is they become inaccessible on some forgotten server. With cloud-hosted DCCs, storage costs continue at peak levels indefinitely whether or not the data are being used or the web portal is being maintained. In short, 4DN highlights the need for the Common Fund to substantively address a question that it has largely been able to ignore until now: What happens to a DCC when its funding ends?

Financial. 4DN is also heading towards similar challenges related to personnel funding gaps and stoppages. As with problems presented by cloud storage, 4DN is similarly unsure of how to cover personnel costs in the interim between Phases 1 and 2. There is also the longer-term question of whether there will be funding to retain some of the current staff to maintain the system in the long-run. Much like cloud-storage costs, personnel costs continue whether or not funding exists to cover them. These issues ultimately exacerbate the challenges 4DN has faced with hiring and retention, as funding uncertainty makes these positions less desirable. 4DN also mentioned the possibility of “talent flight” during these funding lapses or at the end of the program that could set progress back months or years. The CFDE could potentially play a role in helping provide more definite structure around the mid-life and end-of-life challenges Common Fund Programs experience.

Potential Solutions

Hiring and Retention. The CFDE could provide administrative overhead or/and temporary personnel to assist with hiring during ramp-up; fund postdocs, fellowships, or more prestigious opportunities through CFDE that would help DCCs retain and recruit personnel; and provide further career opportunities to help cover funding gaps and/or provide assurances that would make DCC work more stable and attractive.

Training and Support. Create a playbook to help Programs navigate challenges faced over various life-stages: start-up, end of Phase 1, end of Phase 2 and beyond. Provide CFDE resources, information, and possibly personnel to help early programs deal with start-up, and mature programs deal with middle- and end-of-life issues. Provide personnel to work closely with developers to help them create engaging documentation and training resources. Help organize and facilitate external training opportunities like Hackathons, webinars, MOOCs, etc. Provide consultation services through CFDE for FAIRness questions and audits. The CFDE could also provide guidance on how DCCs can operationalize FAIRness through better infrastructure and design principles. Host forums or periodic meetings to showcase exemplar Common Fund programs, allowing developers to share their platform design with other teams.

Interoperability. The CFDE could create and maintain a metadata asset store for programs to share their metadata definitions that don’t yet fit into existing ontologies. The CFDE could also provide dedicated personnel to work with these groups to help them get their metadata definitions added to existing ontologies.

Promote Pairwise Interactions. Provide “matchmaking service” to help Common Fund Programs find partner groups with shared challenges who could benefit from collaboration. Provide “DCC Mentor” service to pair new DCCs with more established groups who work with similar data or who have faced similar challenges. The partnership between ENCODE and 4DN could provide a model.

Promote Sustainable Solutions for Cloud Infrastructure. To address issues with back-loaded cloud storage costs, the CFDE can work with the larger Common Fund Program to create more flexible/dynamic funding structures and provide temporary funds through the CFDE for cloud costs between Phases 1 and 2. The Common Fund programs need real solutions for long-term data stewardship and financial support for programs after their 10-year funding limit is reached. Mainly the CFDE could work together with the larger NIH and Common Fund programs to answer outstanding questions over what happens to cloud-hosted data and the supporting infrastructure after programs end? The CFDE could create a data stewardship program to address sustainability concerns of both cloud-based and locally-hosted DCCs. This could include dedicated CFDE personnel for ongoing support and maintenance upon program sunsets, or provide additional funding to allow DCCs to retain some of their personnel after the program ends.

Potential Projects

The CFDE could develop a metadata search framework/service that uses natural language processing to find semantically similar metadata terms in addition to textually similar ones. For example, a text-based search for “heart attack” wouldn’t find datasets annotated as “myocardial infarction,” but a semantic search would. This model could be provided as a low-level API service or modular tool that could be easily adapted to a variety of purposes. Potential applications of a semantic-search service include powering DCC data portal searches, as well as automated metadata mapping across DCCs. This functionality could ultimately make the CFDE search portal significantly more useful considering the variety of ontologies employed for different purposes across Common Fund programs that may not always use textually similar terms for the same things. Beyond a CFDE search portal, a semantic search tool could help harmonize metadata and promote interoperability among programs without the need to enforce overly-restrictive metadata standards that place an undue burden on DCCs and their programs.

Game Changers

Creating a Common Fund Metadata Asset Store. Existing ontologies provide powerful tools for data interoperability, but Common Fund programs like 4DN often need to define new metadata terms as they work to describe emerging data and experimental types. The controlled vocabulary used by 4DN captures important data features not defined by previous ontologies, but also creates interoperability issues with other DCCs. The CFDE could foster greater interoperability and more collaborative development of new metadata through a central repository allowing programs to define new terms in a common format and share them with other Common Fund programs. As part of this, the CFDE could also work with groups to help them get new metadata terms incorporated into existing ontologies.

Creating a Common Fund Mentorship Program. The ongoing collaboration between 4DN and ENCODE could serve as a model going forward that could help upstart DCCs navigate their first years and foster great interoperability across Common Fund programs. From its inception, ENCODE has played a vital role in helping 4DN develop their infrastructure, define metadata terms, and learn strategies for coordinating across the 4DN network. Beyond making 4DN's first years significantly easier, the relationship has also led to increased levels of interoperability between the two platforms. By formalizing future collaboration through a Common Fund Mentorship Program, the CFDE could both help upstart DCCs gain their footing with guidance from a more experienced team and foster greater interoperability through fruitful pairwise collaborations.

Enabling Data Analysis on the Cloud. While 4DN has the infrastructure to provide robust cloud-based analysis through its web portal, the DCIC does not receive extensive funding to support this capability. Allowing users to compute alongside their data would incentivize program members to host data on their DCC's platform, and increase overall reproducibility by enforcing data analysis standards through shared pipelines and workspaces. If the CFDE could provide additional funding or computing resources to help pay for/support/develop/refine analysis tools hosted on program portals, this could increase both the scale, quality, and reusability of data hosted across Common Fund Programs.

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. Suggested topics to cover: 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?

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
- Move data to cloud
- Metadata harmonized within DCC
- Metadata harmonized with _________
- Metadata harmonized across Common Fund
- Implement new service/pipeline ____________
- Increase number of eyeballs at 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: as a group, or seperate, whatever is convenient for 4D staff

1 - 2pm Open discussion (with breaks)
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:

- 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?

NIH Cloud Guidebook:

- What would you like to see included?
- What would be better left to individual DCCs to decide?
- Would you be interested in contributing to it?

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

Day 2

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, Thoroughness checking

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

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

- datatypes they maintain
- formats etc of same
- tools / resources they think might be useful for the 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 B: Metabolomics Deep Dive

Metabolomics Site Visit

Location: University of San Diego, Bioengineering Building, 9500 Gilman Drive, La Jolla, CA

Date: Wednesday-Thursday, November 13-14, 2019

Attendees: Representatives in attendance from the CFDE were Amanda Charbonneau (UCD), Meisha Mandal (RTI), Titus Brown (UCD), and Owen White (UMB). The representatives from Metabolomics were Shankar Subramaniam (PI), Eoin Fahy (Senior Scientist), Kenan Azam (Scientific Programmer), and Andrew Caldwell (Research Scientist). Manish Sud (Lead Research Analyst) was not able to attend the meeting. Two representatives from the San Diego Supercomputer Center (SDSC): Christine Kirkpatrick (Executive Director) and Kevin Coakley (Senior Systems and Cloud Integration Engineer).

Meeting Logistics

The CFDE engagement team met with representatives of the Common Fund's Metabolomics program on Wednesday, November 13, 2019 and Thursday, November 14, 2019 in the lab of Dr. Shankar Subramaniam at the University of California, San Diego (UCSD) to discuss the Common Fund's ongoing Metabolomics program headed by Dr. Subramaniam. The agenda at the end of this document was used as an informal guide for structuring the day.

After brief introductions, the engagement team and Metabolomics representatives, each reviewed their goals for the meeting. For the engagement team, these goals include learning about the structure and goals of Metabolomics, including technical specifications about the data they host, as well as information about training, organization, and the overall set of priorities for their group. For Metabolomics, these goals included discussing data integration as well as metadata, ontologies, and harmonization. In turn, our hosts provided us with an overview of their Metabolomics Workbench site, including descriptions of the studies involved and types of data generated, online data submission process, data sharing/harmonization, data analysis, and a live demonstration of their portal.

Metabolomics Overview

The Common Fund's Metabolomics program aims to stimulate metabolomics research and to expand the capacity of researchers to study small molecules and their relationship to human disease. While the field of metabolomics research is several decades old, it has lagged behind other biomedical fields in terms of popularity and ease of use. Processing metabolomic data requires the use of specialized equipment (NMR, mass spectrometry, liquid chromatography) and requires expert users for both the technical machine aspects and the first pass data analysis. The Metabolomics program represents a dedicated effort to bring metabolite screens into more widespread use by working to standardize methods, compound names, and analysis tools, thereby lowering the barrier to entry for researchers.

The Metabolomics Program was funded in two unequal stages. They are currently halfway through their second phase and the consortium consists of six Regional Comprehensive Metabolomics Resource Cores (RCMRCs), sometimes also called the Compound Identification Cores (CIDCs), and seven Data and Tools Cores (DTCs) that are overseen by the Metabolomics Consortium Coordinating Center at the University of Florida. See below for a table summarizing the roles and member organizations for the Metabolomics centers.

The creators of the National Metabolomics Data Repository (NMDR), who hosted our visit, are responsible for collating, analyzing, and distributing the data gathered by the RCMRCs and hosting the tools and methods created by the DTCs. For the first six years of Metabolomics funding, there were also three training centers, however dedicated training centers were dropped for the second phase of funding.

Data, including associated metadata, for the Metabolomics program are hosted by the NMDR. The NMDR contains data from 11 major taxonomic categories that are generated through both targeted (focused on a set of defined metabolites) and non-targeted (covering the entire metabolome, including unknown metabolites) metabolomic assays. Currently, the NMDR contains ~1200 studies from ~200 different institutions worldwide. It is dominated by human data, but a considerable portion is from mice and other mammalian species.

Center	Member Organization(s)	Responsibilities
Metabolomics Workbench/NMDR	1. UC San Diego	Collate, analyze, and distribute data generated by RCMRCs. Host tools and methods created by DTCs.
Metabolomics Consortium Coordinating Center (M3C)	1. U. of Florida	Coordinate activities across Metabolomics centers and promote the work of the Consortium.
Regional Comprehensive Metabolomics Resource Cores (RCMRCs)	1. U. of Georgia 2. U. of Michigan 3. UC Davis 4. Emory University 5. Pacific Northwest Nat. Lab.	Develop processes and resources to facilitate and improve the accuracy of metabolite identification.
Data and Tools Cores (DTCs)	1. MD Anderson Cancer C. 2. Vanderbilt University 3. U. of North Carolina Charlotte 4. Emory University 5. U. of Michigan 6. U. of Colorado, Denver 7. U. at St. Louis	Develop methods and computational tools to facilitate the collection, processing, and analysis of metabolomics data.

Program Lifestage

The Metabolomics program was initially funded in 2012 and has been receiving data submissions since March 2013. The program wrapped up the first stage (six years) of funding in

FY17 and is currently in the second stage, which spans FY18-FY21. As the program has matured, its database has come to include diverse data sources and the NMDR has become a hub of their community. Early in the program, starting in June 2015, the NMDR additionally began accepting data from a variety of sources outside of their consortium as well as aggregating data from their RCMRCs. Currently, of the ~1200 studies in their database, about 150 were deposited by NIH-funded programs or consortia. The rest were submitted by non-NIH-funded institutions—over 900 came from within the US, and a further 100 were international submissions. Similarly, their analytical chemistry-centric effort to standardize molecule names, RefMet, the "Reference list of Metabolite names" is an international effort, and its governing board has eight international experts. They also developed mwTab, a standardized format for sharing metabolite data and metadata in a single file. The mwTab format is now widely used, and is the backbone of the MetabolomeXchange, an index of metabolomics data used by five countries.

Data Platform

The Common Fund Metabolomics program developed and maintains the Metabolomics Workbench (MW), the Metabolomics Workbench Metabolite Database, the Human Metabolome Gene/Protein Database (MGP), and the Reference list of Metabolite names (RefMet).

The Metabolomics Workbench (MW), https://www.metabolomicsworkbench.org, is an online interface to the NMDR developed at UCSD. It allows users to manage and upload studies as well as browse and search available studies. Using the MW interface, submitters upload data and results, including metadata, targeted data measurements, protocols/methods files, untargeted data measurements, and raw data (MS/NMR files, etc.). Other researchers can then use the MW website to browse, search, analyze, and download data as well as view summary figures of key study search parameters (e.g., bubble chart showing studies by sample source). For example, studies can be filtered by study metadata (disease, sample source, species, instrumentation) or metabolite information (metabolite classification, biochemical pathways, retention time, etc.) to identify data relevant to the user's needs. Additionally, it provides analysis tools and access to metabolite standards, protocols, tutorials, training, and other resources to support metabolomic researchers.

The MW analysis offerings include normalization/averaging, clustering/correlation, univariate analysis, multivariate analysis, feature analysis, classification, and comparative analysis across studies. Users can also programmatically query and access data (metabolite structures, metadata, experimental results) through the MW REST API (https://www.metabolomicsworkbench.org/tools/mw_rest.php) using HTTP requests in a browser, script or third-party application. The REST API was released last April and was developed in collaboration with the Scripps Research Institute.

The Metabolomics Workbench Metabolite Database is a Postgres database containing over 65,000 structures and annotations of biologically relevant metabolites collected from public repositories (e.g., LIPID MAPS, ChEBI, HMDB, BMRB, PubChem, KEGG). Users can search for metabolites in the database by substructure, text, or mass (m/z ratio). Each entry contains key information about the metabolite, including structure, molecular weight, common and systematic names, PubChem compound ID, and classification. Entries also contain cross references to external databases and repositories (e.g., HMDB, ChEBI, LIPID MAPS, METLIN, ChemSpider, KEGG, etc.) as well as links to the MoNA MS spectra and human metabolic pathways containing the metabolite. Additionally, the open-source chemistry cartridge enables substructure searching, generation of chemistry-centric attributes (formula, exact mass), and interconversion of molecular formats.

The Human Metabolome Gene/Protein Database (MGP) is a database of metabolome-related genes and proteins containing over 7,300 genes and over 15,500 proteins. Users can search by gene (name, symbol, entrez ID, etc.), HMDB Pathway, or Reactome Pathway. MGP displays genes/proteins and metabolites associated with a pathway of interest. Searching by gene displays information about the gene's associated proteins and pathways, including a summary of the function and metabolites involved in the pathway.

The Reference list of Metabolite names, RefMet, is effectively a large spreadsheet that provides a standard nomenclature for over 95,500 chemical species. Users can interact with RefMet in a number of ways. From the Metabolomics Workbench website, it can be browsed and searched directly or a user can input a list of metabolite names and have them automatically converted to RefMet nomenclature. A user can also directly download the data, either in whole or after filtering as one would with a simple Excel sheet. Or the entire dataset can be downloaded as part of a Shiny R app and queried locally.

Infrastructure

The NIH Common Fund Metabolomics program runs entirely on cloud platforms, with some caveats. They have taken advantage of STRIDES and have worked to integrate that billing system into their project to pay for data storage; however, they are also part of a cloud computing collaboration with the San Diego Supercomputer Center. Their Workbench platform is hosted on an internal cloud system at UCSD. Their cloud-based workbench and cloud-hosted notebooks are on Google Cloud Platform (GCP). Analysis tools are in Docker containers running Jupyter notebooks and utilize both R and Python modules. The public can access the notebooks through binder, a free service that hosts Jupyter notebooks (mybinder.org). Jupyter files can be copied to a local installation for private access.

Analysis

The MW offers a number of data analysis tools, including a study-specific analysis toolbox accessible from the main study page. The analysis workflows use common R statistics packages that have been integrated into embedded R Shiny modules. Users can run a wide range of statistical analyses on the workbench using data hosted in the NMDR. As the pipelines are built entirely of standard R packages, a user could technically re-implement the pipeline locally, however this would be difficult for a novice user. In creating the toolchains encoded by the Shiny modules the Metabolomics group has made a number of decisions about parameters and defaults that are not transparent to the user. This makes the Shiny apps extremely easy to use, but not straightforward to recreate. A user wishing to run the same pipeline locally would need to create the code that strings the various tools together, as well as set a number of parameter options. However, users can analyze their own data on the workbench site using the pre-standardized pipelines by uploading a tab or comma delimited file. Once uploaded, the user is presented with a Shiny interface where they can set several useful options such as defining sample groupings, choosing which experimental factors to include, and what normalization method should be used to ensure that the data is analyzed correctly. For most users, who have only one or two datasets to analyze at a time, uploading their data to the site and using the point and click interface is likely preferable to local reimplementation.

Many of the individual analysis tools as well as the ability to search RefMet have also been ported to Python, and can be run in Jupyter notebooks using binder or could be downloaded and run in a local Jupyter notebook. These are all available at the Metabolomics Workbench GitHub page: https://github.com/metabolomicsworkbench. These tools are still actively being ported, and are not yet as user friendly as the Shiny R apps embedded in the MW. As of this writing, the Jupyter notebooks are designed as examples of the tool, and do not have instructions for how to edit the code to use other data. As such, an end user would need to be familiar with Python to use them.

Access

Uploaded data are publicly accessible via the MW's web portal or API. Unlike any other DCC we've visited, an end user at Metabolomics can access both raw and processed mass spectrometry (MS), nuclear magnetic resonance (NMR) and related files. There is also open access to all available metadata, protocols, and methods for both targeted and untargeted studies. Users also have free access to a variety of browsing, searching, and statistical analysis tools, can upload their own data for processing, and can download any of the files in a variety of formats. All of these data, features and tools can be accessed by any user, without a log-in.

Data creators are similarly unrestricted. The NMDR welcomes data from any source as long as it is a metabolite analysis file, is submitted with the required metadata, and passes their review process. To facilitate the review process, data creators are required to create an account to obtain authorization. Authorization requests are reviewed by the MW staff and are typically approved within five business days. Once authorization is obtained, researchers need to register their study prior to uploading data. A step-by-step tutorial is available to assist researchers in the data submission process. Once the study is submitted, NMDR reviews the uploaded dataset, which typically takes less than 30 days. After the NMDR review process is completed the submitter is notified and may review their uploaded data. After the entire review process is completed, the study is made available on the public MW website. During the submission process, submitters can set an embargo if the study has not yet been published and can lift the embargo after publication to allow access to the data. At the time of our meeting, approximately 180 submitted studies were under embargo.

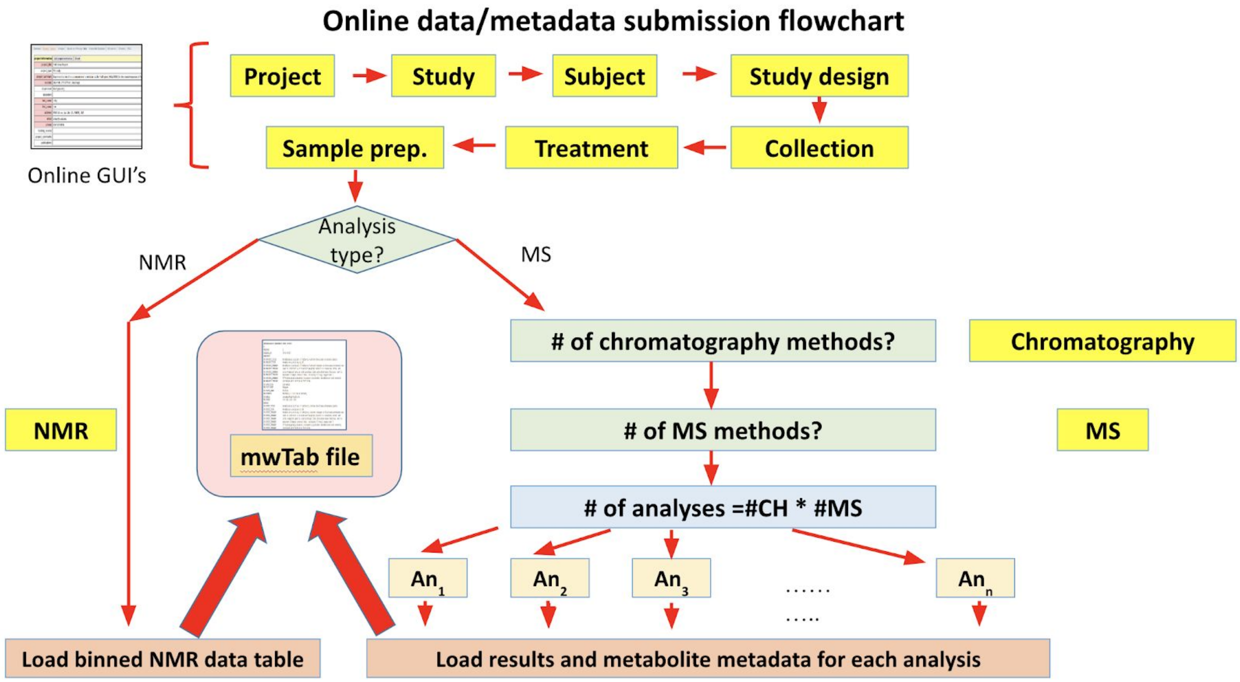
Harmonization and Metadata

The Metabolomics program has invested heavily in data/metadata harmonization efforts, most notably with the development and maintenance of a reference list of metabolite names and a new file format (mwTab). However, they warned that metabolomic data will always be difficult to combine across studies as results can vary widely due to differences in machine or operator. For instance, Dr. Subramaniam's group ran a "ringtest" where they sent the same samples to multiple cores for analysis. In many cases the results from different cores had significant variation suggesting that even properly harmonized data might be difficult to re-combine with other studies. Still, resources such as mwTab and RefMet have greatly increased our ability to cross-reference study data, and have made data portable across platforms.

The Metabolomics program developed the Reference list of Metabolite Names (RefMet) as a way to standardize metabolite names across studies and experiments. Depending on the instrument used to identify molecules (e.g. nuclear magnetic resonance (NMR), mass spectrometry), the method used to separate fragment and separate metabolites (e.g. mass spectrometry (MS), liquid chromatography (LC), time of flight (TOF), etc), and the electric charge of the substrate used to separate them, the same starting metabolite will be discovered as one of many different breakdown products across experiments. For this reason, metabolites can often be accurately referred to by multiple different names in the literature. This makes it exceedingly difficult to determine whether any two studies found the same metabolic compounds unless they used exactly the same methods. To create RefMet, the NMDR recorded ~220,000 variable metabolite names from across ~1200 MS and NMR studies. By mapping all the variations in metabolite names to standard RefMet designations, the NMDR was able to collapse most (about 200,000) of the variable names into 95,000 standardized metabolite species.

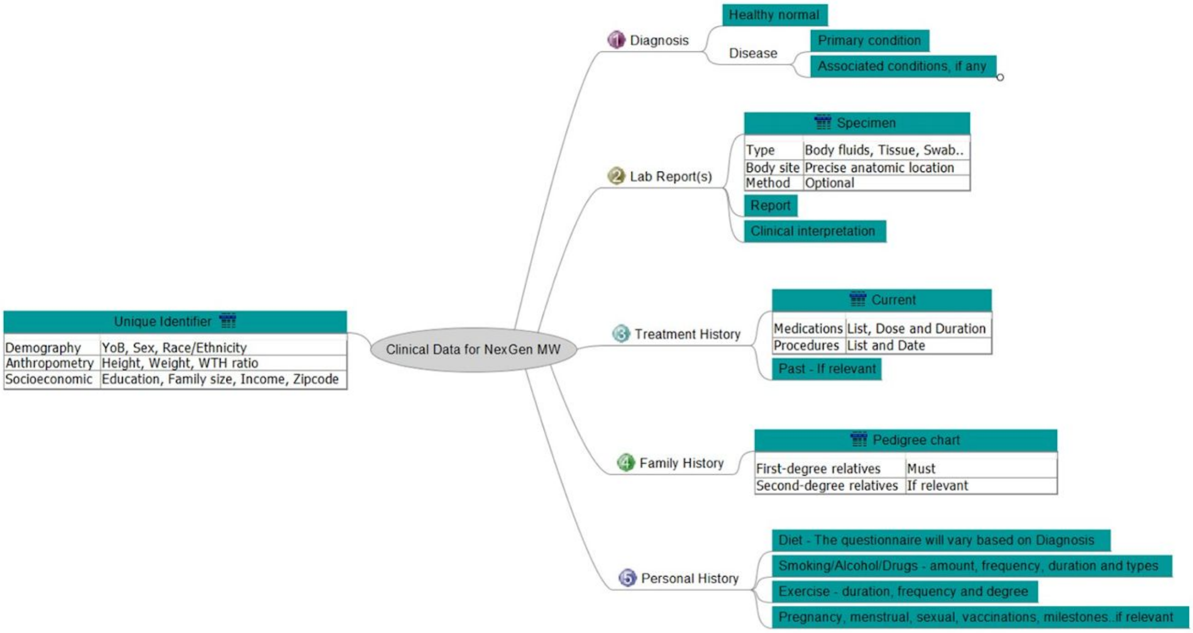
RefMet is a huge international effort. The database is entirely human curated, and the work is governed by a board consisting of eight international experts in metabolomics, including collaborators like PubChem. Users submitting data to the NMDR are strongly encouraged to use common metabolite names from RefMet in uploaded data. An online tool to map current metabolite identifications to corresponding RefMet species in metabolomic data is available to assist with this process via the MW.

Another data harmonization effort undertaken by the Metabolomics program is the creation of the mwTab file format. It was designed to serve as a “common currency” for sharing all of the relevant metadata and data from an experiment in a single file, for any metabolomic assay, and is the basis of the NMDR data repository, as shown in their online data/metadata submission flowchart below. The format is both machine and human readable, and very flexible. It contains sections for high level metadata such as project information, PI contact and publication information, as well as more standard metadata (e.g. sample collection, treatment, sample preparation, analysis).The data section of the file is broken into a number of optional data blocks, so a single file can accommodate any combination of methods. Each data block has a number of standardized, optional, fields which correspond to every possible setting for various instruments. A third section allows for mostly freeform additional comments. Since the format is machine readable it can be read by all R based tools, and since it includes all relevant metadata, it can be automatically ingested into the NMDR as part of data submission.



Since its creation, the mwTab format has become popular in the field, and is the file format used by the MetabolomeXchange, an international data aggregation and notification service for metabolomics. The MetabolomeXchange was originally funded by the European Commission-funded COSMOS project, and is used by five countries. The site indexes metabolomics data with mwTab and connects a number of metabolomics data repositories across Europe and the USA. The MetabolomeXchange also provides an API which can translate mwTab into several other formats. In effect, all data in the MetabolomeXchange, as well as any data that is properly formatted into an mwTab file, is automatically harmonized with all of the data at Metabolomics and can instantly be used in any of their tools.

In addition to developing mwTab format and RefMet, the Metabolomics group encourages and facilitates data harmonization in other ways such as through the use external standards whenever possible. The NMDR has developed a metadata model to support large scale human studies in the Metabolomics Workbench. The model, shown here, was made in consultation with TOPMED, MoTrPAC, CHEAR and the European Bioinformatics Institute:


Clinical Data for NexGen MW
Unique Identifier
Demography | YoB, Sex, Race/Ethnicity
Anthropometry | Height, Weight, WTH ratio
Socioeconomic | Education, Family size, Income, Zipcode
1 Diagnosis
Healthy normal
Disease
Primary condition
Associated conditions, if any
2 Lab Report(s)
Specimen
Type | Body fluids, Tissue, Swab..
Body site | Precise anatomic location
Method | Optional
Report
Clinical interpretation
3 Treatment History
Current
Medications | List, Dose and Duration
Procedures | List and Date
Past - If relevant
4 Family History
Pedigree chart
First-degree relatives | Must
Second-degree relatives | If relevant
5 Personal History
Diet - The questionnaire will vary based on Diagnosis
Smoking/Alcohol/Drugs - amount, frequency, duration and types
Exercise - duration, frequency and degree
Pregnancy, menstrual, sexual, vaccinations, milestones..if relevant

Note that although the Unique Identifier contains a number of clinical variables, all human data in the MW is completely de-identified.

Sustainability

As the Metabolomics program plays a central role in international efforts to standardize and popularize metabolomic research, they have begun to make plans to ensure their work survives the projected end of Common Fund funding. For RefMet, they are hoping to involve some other institute in sustaining the effort. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) has indicated interest, however there is currently no guarantee that funding will appear. The NMDR is actively in talks with NIDDK to explore ways to continue funding, but are also interested in other opportunities if they become available. Due to its integration with the MetabolomeXchange, it seems likely any required maintenance will be tied to that effort.

The fate of the Metabolomics Workbench, the portal that allows end users to access and analyze NMDRs data, is less certain. All of the pipelines for the project are in Docker containers or other similar shareable systems. The idea is to provide everything in a way that lets sophisticated users do their analysis, and lets companies (the machine creators) engage with the software and tools they've built. However, the widespread popularity of the workbench, paradoxically, makes it more difficult to find an Institutional Center or other organization to fund it. Dr. Subramaniam has a long history of obtaining funding for projects in systems biology and metabolomics and is pursuing alternative sources of funding for the Metabolomics program after the FY18-FY21 funding period ends.

Training

Internal

Currently, all training and nearly all help requests that occur at the NMDR are related to data submission; primarily on how to create and properly format mwTab files. The team at UCSC has created a detailed format specification and has developed some training materials which they occasionally offer as workshops, however the bulk of their interaction with these users is one-on-one: when a data creator is trying to format their mwTab file, they contact the NMDR for help and advice, or the NMDR reaches out to investigators who have submitted an ill-formatted document. They field approximately 100 queries per day, almost all of which are related to data submission, often from users submitting to the NMDR for the first time.

External

The MW site has a high volume of traffic, mainly from users querying the metabolite and RefMet databases, or using the statistical analysis tools. Remarkably, they report having almost no support burden from end users despite having over 270,000 page views per year. The NMDR team told us that this is where they would most like to expand their training efforts. They are very interested in reaching out to end users, finding out what they're doing with the data, and offering training or help to increase data use. Currently, the only training the NMDR offers end users is written documentation on using the search and analysis features. The NMDR indicated that they are very interested in getting assistance from the CFDE with training, and engaging end users to incorporate NMDR data into larger studies. In stage 1 of the Metabolomics project, two groups in the consortium had funding allocated specifically for training. However, in stage 2, no groups are solely tasked with training, and there is no substantial funding allocated to training efforts. The Metabolomics Consortium Coordinating Center (M3C) at the University of Florida, which handles overall coordination for the Metabolomics program, does host end user trainings. However, these are typically broader training on metabolomics experimental techniques and data analysis. At the time of our visit, 15 training workshops, courses, and conferences were scheduled from January 2020 to November 2020 and ranged in length from two days to several weeks. The events have costs ranging from free (one event) to $5,000, typically have ~100 attendees and are live-tweeted. They are generally organized by the various centers of the Metabolomics program but also include links to training efforts outside the consortium such as at Birmingham Metabolomics Training Centre, and The Association for Mass Spectrometry.

FAIR

The NMDR team is committed to making their data FAIR, and has made great strides in improving the FAIRness of metabolomic data, both in their own database and across the field. RefMet, their database for standardizing metabolite names, increases the Findability of data, and allows researchers to Find and Reuse datasets that might be Interoperable. mwTab, their file format specification, makes all of the data at the NMDR at least technically Interoperable with any other data in that format. As mwTab is used by metabolomic data repositories in six countries, this represents a huge proportion of available data. The format is documented at fairsharing.org

Data at the NMDR also has a number of basic FAIR features. For instance, metadata is formatted using Schema.org which allows it to be found and indexed by search engines, and DOIs assigned to each submitted dataset study. The Metabolomics Workbench website is completely open access. Raw and processed data can be searched, analyzed and downloaded in a variety of formats, without a login, making all of the data freely Findable and Accessible. Additionally, all metabolomic data and RefMet information can be programmatically accessed via an API. Users can also Reuse any of the NMDR tools on their own uploaded data either on the MW website, or using Dockerized versions.

The NMDR staff did express some frustration with current efforts to measure FAIRness. They told us that rubrics developed to test FAIRness are generally not designed by people who implement FAIR principles. This means that the priorities of the data maintainers and the constraints they face when implementing FAIR are often not accounted for. This mismatch between the reality of data managers and the goals of FAIR rubric creators sometimes results in metrics that can’t be satisfied, and other times in metrics that are easy to meet, but don’t actually materially improve the FAIRness of the data.

Cross-pollination

The members of the Metabolomics NMDR are highly supportive of cross-pollination between Common Fund Programs and are already involved in a number of collaborations with major metabolomics data-producing entities. At the direction of NIH they have ongoing collaborations with MoTrPAC and TOPMed. However, there is no funding specifically allocated to support these efforts. In addition to their previously mentioned RefMet and mwTab work, the Metabolomics program also collaborates with the Children’s Health Exposure Analysis Resource (https://cheardatacenter.mssm.edu/) at the Icahn School of Medicine at Mount Sinai. They also have a strong collaboration with LIPID MAPS. Notably, ~65% of all the studies on the MW relate to lipids.

SSO (Single Sign-on)

The NMDR does not use an SSO service, and only requires a login for certain users. Data generators are required to create a log-in and register before they can submit data. New data submitters are manually screened before being authorized for access. Submitters are responsible for cleaning their data prior to uploading it in accordance with Federal Health Insurance Privacy and Portability Act (HIPAA) requirements. In particular, human data must be de-identified, with patient IDs and any other personally identifiable information removed. As no private data is in the portal, end users are not required to create an account or log in to the MW in order to download or access data, or to use tools.

As data upload is only permitted by people with an account, and other users do not currently require a login, moving the workbench to a SSO system would require several large changes to their workflow while making access more difficult for end users.

Outcomes

Infrastructure and Resource Reuse

The MW tools are, by design, highly reusable, and will work on any data that is coerced into the mwTab format. Additionally, they have ported the R pipeline to Python based Docker containers that could be easily reused by other programs with similar data types. The mwTab format created by the Metabolomics group also contributes significantly to the MW platform's reusability. Any data in mwTab format, independent of the data's source, can be submitted to the MW and analyzed using the available tools.

Dr. Subramaniam has a wide breadth of knowledge about other Common Fund programs. He is an advisor for TOPMed, LINCS, and MoTrPac. Over the course of two days, he related several potential Use Cases 'Horizontal Integrations' that could be completed with various collaborations, explained how to create a matrix view of Common Fund datasets that could be used to identify new potential collaborations, and presented a way to structure all Common Fund data into a nested ontology. Dr. Subramaniam and his lab were clearly thinking deeply about both the theoretical and practical aspects of building a Common Fund Data Ecosystem long before we arrived, and their help will be invaluable for community building within the CFDE.

Challenges

During the meeting the Metabolomics NMDR team identified a number of challenges it is currently facing or anticipates facing in the future.

Unfunded Mandates. One challenge is the lack of funding for collaboration efforts. For example, the collaborations between Metabolomics and MoTrPAC, and Metabolomics and TOPMed have been directed by NIH, however no specific funding has been allocated to collaboration efforts. Any time dedicated to collaboration, such as the time required to convert data into shareable formats or standardize names using RefMet, must be absorbed by the Programs. Similarly, projects are often required or expected to host training, but with little or no support. Although two groups were allocated funds specifically for training in the first stage of the project, there are no centers with funding solely for training in-specific funds for stage 2.

Researcher Focus. Metabolomics researchers tend to be “islands” and stick to metabolomic or lipidomic studies as opposed to multi-omic or other cross-discipline studies. This is not unique to metabolomics researchers: researchers in other areas rarely think about using metabolomics.

Barriers to Training Access. One reason that metabolomics data is underutilized is that the data can only be collected on specialized equipment that is often restricted to only expert users, and analyzing that data requires specialized knowledge that is difficult to acquire and has a steep learning curve. Researchers generally learn by being in a lab that already does metabolite work, but for an outsider, there are few ways to get even a cursory overview of what metabolomics is, or to learn why a researcher might want to start incorporating that kind of data into their experiments. Christine Kirkpatrick lamented that after an exhaustive search, she couldn’t find a single example of metabolomic non-fiction. The closest she could find to an intro on the topic for a novice was a book chapter written by Eoin Fahy. There are training workshops and other events about metabolomics, however they are costly, aimed at people who already do metabolite research, and not broad enough. For example, a ‘Best practice in operating mass spectrometers in Metabolomics’ workshop, offered by the West Coast Metabolomics Center is a beginner course, but teaches only how to run a single machine, with no data analysis or practical application. Registration is also $1500 and the course is a four day commitment, an amount of time and money that would likely deter a researcher who simply wants to see if it makes sense to add metabolomics to their skillset.

STRIDES. The NMDR team shared that although they are happy with their STRIDES collaboration, they are concerned about how the program might be affected by the conflicting goals of the NIH and commercial interests. They noted that the cloud providers are often the ones writing the policies and that the administrators at the NIH who are brokering the deal may not appreciate their long-term consequences. For instance, part of the STRIDES discount is made by putting most of the user support burden on the CF Programs. This means that a DCC will be responsible for answering questions not just about their own data and pipelines, but about the cloud infrastructure and other things that would normally have been answered by the cloud provider. The savings from the discount, however, are unlikely to make up for the increased support burden at DCCs.

Potential Solutions

We also discussed a number of ways in which the CFDE could offer support to help address some of the challenges faced by the Metabolomics program.

Funding More Cross-talk between Groups and Researchers. In addition, the CFDE plan to fund new collaborations, funding could also go towards activities such as expanding existing collaborations. These efforts could include other data-generating organizations or outreach aimed at metabolomics researchers to encourage the use of standard formats/naming or other data harmonization efforts.

Another potential direction would be funding projects specifically aimed at increasing cross-collaboration. This could be an initiative similar to the NSF Convergence Accelerator (https://www.nsf.gov/od/oia/convergence-accelerator/index.jsp) which specifically funds projects with a convergent approach to promote interdisciplinary collaborations.

Training and Community Building. The CFDE could help with two main training efforts: training for existing users, and outreach to broaden the use of metabolite data.

- Expanding Metabolomics Workbench-specific training for data submission and mwTab format would help reduce the user support burden. Currently, the NMDR gets about 100 requests per day for help with data submission.

Outreach efforts aimed at researchers outside the metabolomics community to encourage the broader use of metabolite data. This requires outreach and education about the basics of metabolomics techniques and analysis, as well as examples of how this kind of data can enhance experiments. A number of ideas were discussed for specific training workshops that the CFDE could offer. Of note, Christine Kirkpatrick stated that she has a pool of people that might be able to organize and lead CFDE trainings.

- Introducing metabolomics with a focus on topics of public interest to engage participants. One possibility is a ‘What are metabolomics and metabolite signatures good for’ workshop. Potential topics:
    - Analysis of metabolites in blood/urine/saliva for non-invasive disease diagnosis.
    - How do specific pharmaceutical drugs alter the metabolome? Look for deviations from homeostasis and effects in other metabolites beyond the expected direct targets of the drug.
    - How does the environment alter our metabolism (exposome)?
    - Interrogate tissue specific metabolism.
    - Interaction of metabolomics with gene expression and other biological processes.
- Novice friendly training workshop in metabolomics. This could be an ANGUS-affiliated workshop or hackathon covering:
    - Processing and analysis of metabolomic data

    - How to use mass spec data
    - Naming of metabolites
    - Connecting metabolites to functions and pathways
    - Common pitfalls in metabolomics research
- A “data storytelling” workshop to train current experts in metabolomics how to effectively communicate their findings and research to be interesting to a broader audience.
    - This could be combined with a storytelling workshop and evening where NIH program managers are invited to hear researchers present their stories. This would allow NIH to hear about the impact of the CFDE through users instead of the CFDE.
- An ‘Opportunities and Challenges of Metabolomics” workshop or hackathon.
    - This could be done as a collaboration between MoTrPAC and Metabolomics.
    - Ideas for workshop topics:
        - Materials development/brainstorming workshop around data integration with other data types.
        - Bring your own data and learn how to intersect it with metabolomics data.
        - An exercise incorporating metabolomics into published research.

Infrastructure and Data Reuse. In order to increase reusability, the CFDE could support an effort to standardize metabolomics data and have other DCCs routinely use RefMet and mwTab format. This would be facilitated by making an API that can directly translate mwTab into a C2M2 compatible format.

Potential Projects

As previously noted, the NMDR is already engaged in a number of efforts across funding institutions and nations, and is actively working to create new collaborations with MoTrPAC and TOPMed.

Game Changers

Crowdsourcing Data Harmonization Use Cases. Dr. Subramaniam and the Metabolomics team had many intriguing ideas for crowdsourcing use cases. Some of the ideas discussed were:

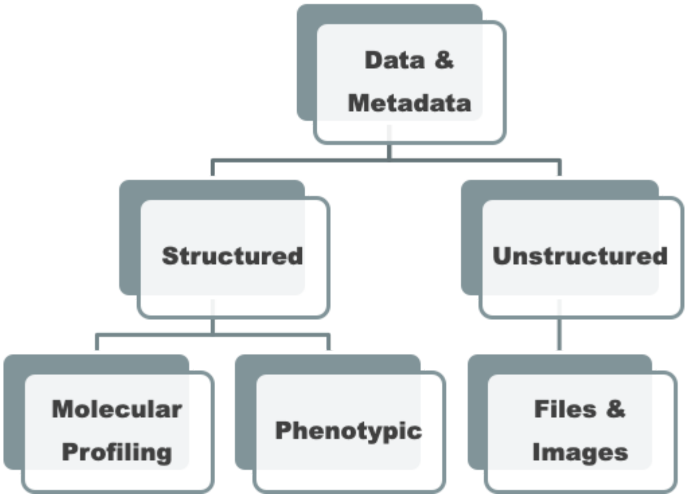
- Developing a platform where PIs can submit ideas for use cases. NIH funded postdoc positions would be dedicated to working closely with the relevant DCCs on implementing these use cases. One goal would be to make the analysis reusable for use in future projects.
- Presenting researchers with a matrix of Common Fund datasets and collecting ideas for use cases involving the data as well as developing a platform to match them with postdocs that could help implement their ideas.
    - One way to implement this would be to bring together a group of bioinformaticians and clinical researchers to discuss what types of data are available in the common fund.

- Have a 30-minute presentation at an IC council meeting by the CFDE providing a matrix of the types of data available in the Common Fund to researchers in a variety of research areas (HLB, diabetes, etc.). Ask researchers to think of use case scenarios that would be useful in advancing their research.

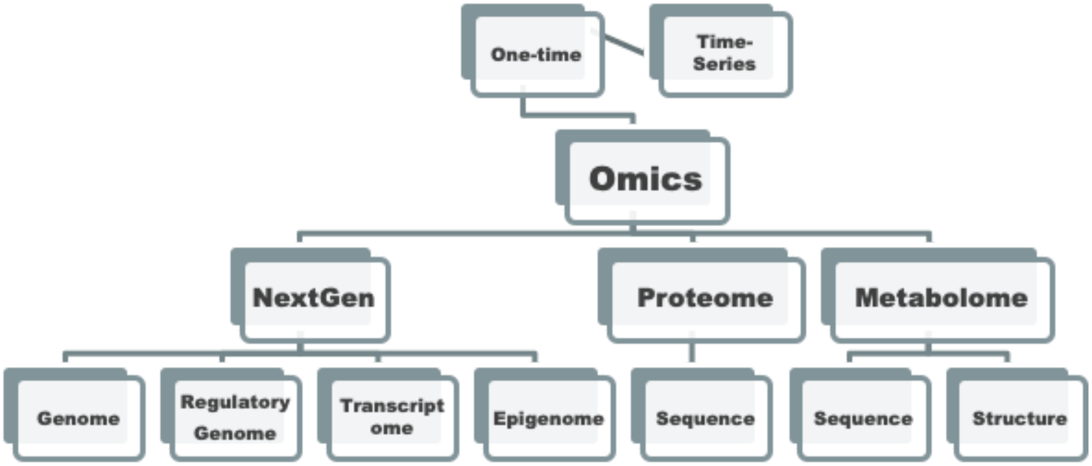
Common Fund 10 Big Ideas. Similar to the NSF 10 Big Ideas, create a collection of topics for CF DCCs to coalesce their collaborations around.

Informatics-specific Funding. We live in a data world. Doing good science depends on having reliable, accessible, well managed data, but the NIH rarely dedicates money to data. NIH Institutes don't have a specific line item for data science, data management or informatics, and so money often ends up being used for the science, at the expense of the data. The CFDE could advocate for the NIH to allocate funding specifically for data management, data science and informatics at a high level, i.e. having a certain portion of the NIH budget dedicated to it. This is being done elsewhere in other federal agencies., NASA, for example, has a percentage of their funding that goes specifically to data management.

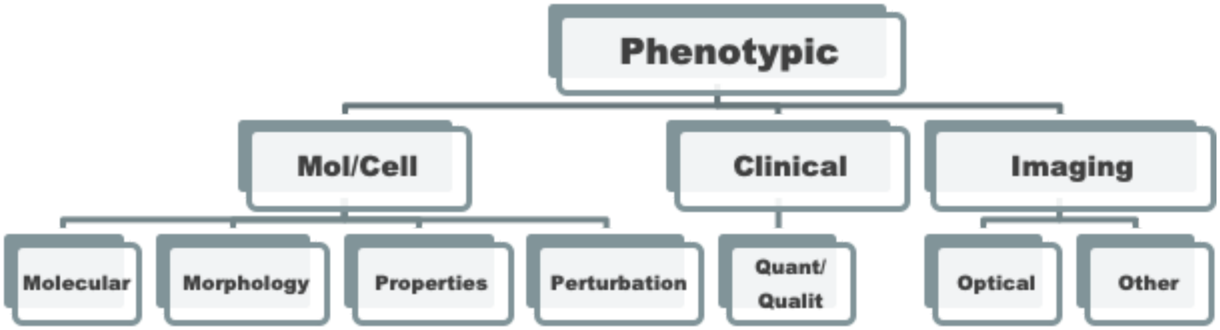
Innovative ways to view Common Funds Data Portfolio. Dr. Subramaniam suggested several ways to condense and contextualize the varied datasets at the Common Fund so that a Program Officer or PI at a program could easily determine how their project fits into the overall portfolio, and who might be good targets for collaboration. For instance, the CFDE could create a matrix view of datasets, where programs could be easily compared by the data types, biological systems, or other useful terms that describe their work. He further suggested that one could imagine the Common Fund portfolio in terms of nested hierarchies describing the assets. At a very high level, data could be grouped according to its structure:



Data might be conceptually categorized by aspects of the technology and protocol used to create it:



Or by its scientifically interesting properties:



The overall idea is that if we built these kinds of trees, and mapped Common Fund programs onto them, it would be easy to not only see what exists, but to find overlaps between them.

Scientifically Motivated Cross Cutting Questions. Dr. Subramaniam shared several 'Horizontal Integrations'; scientific questions that he thinks could be answered with current Common Fund data, by someone with sufficient time and expertise to combine datasets:

> LINCS project has generated proteomic and transcription factor data following drug perturbation of cells (e.g. cancer cells). ENCODE has ChIP seq data on the same cells. GTEx provides tissue specific expression on the tissue that embeds the cancer cells.
>
> Can we seek:
>
> - What are the transcription regulatory pathways and genes that are affected in the tissue based on extrapolating the drug perturbation data from the primary cancer cells?

- What CRISPR experiments can be designed based on the function/pathway analysis to explore tissue specific role of specific genes?
- From LINCS time series protein data, can we infer a “pseudo” time series that can be explored by specific kinase inhibitors?
- Can the mechanisms inferred be explored in other tissues from tissue-specific expression data?

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. Suggested topics to cover: 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?

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
- Move data to cloud
- Metadata harmonized within DCC
- Metadata harmonized with _________
- Metadata harmonized across Common Fund
- Implement new service/pipeline ____________
- Increase number of eyeballs at 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: as a group, or separately, host's choice

1 - 2pm Open discussion (with breaks)

Using the results of the morning's 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:

- 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?

NIH Cloud Guidebook:

- What would you like to see included?
- What would be better left to individual DCCs to decide?
- Would you be interested in contributing to it?

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

Day 2

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, Thoroughness checking
DCC reflection on suggestions, open discussion to find shared solutions.
Touch on any questions not covered previously, ensure we have information on

- datatypes they maintain
- formats etc of same
- tools / resources they think might be useful for the 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 C: MoTrPAC Deep Dive

Stanford Molecular Transducers of Physical Activity Consortium (MoTrPAC) Bioinformatics Center (BIC) Site Visit

Location: Stanford University, Falk Cardiovascular Research Center

Date: Tuesday, October 9, 2019

Attendees: Representatives in attendance from the CFDE were Amanda Charbonneau (UCD), Nathan Gaddis (RTI), Titus Brown (UCD), Owen White (UMB) and Anup Mahurkar (UMB). The representatives from MoTrPAC BIC were Euan Ashley (Co-PI), Matt Wheeler (Co-PI), Ashley Xia (NIH Project Scientist), Steve Hershman (Director of mHealth), Malene Lindholm (Postdoc), Karen Dalton (Software Developer), Jimmy Zhen (Software Developer), Young Kim (Software Developer), Shruti Marwaha (Research Engineer), David Jimenez-Morales (Computational Biologist), David Amar (Biostatistician), Archana Raja (Computational Biologist), and Elizabeth Chen (Graduate Student, Biostatistics).

Meeting Logistics

The CFDE engagement team met with representatives of the Molecular Transducers of Physical Activity Consortium (MoTrPAC) Bioinformatics Center (BIC) on Tuesday, October 8, 2019 and Wednesday October 9, 2019 at the Falk Cardiovascular Research Center at Stanford University to discuss their work in support of the NIH Common Fund's MoTrPAC program. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the day.

The engagement team began by reviewing their goals for the meeting, which included learning about the structure and goals of the MoTrPAC BIC, including specifics about the data they host, as well as information about training, organization, and the overall set of priorities for their group. In turn, MoTrPAC BIC representatives provided us with a comprehensive overview of their work on the MoTrPAC program to date, including descriptions of the types of data generated by the consortium, the BIC pipelines for data QC, processing, and analysis, and a demonstration of the web portal for data distribution.

MoTrPAC Overview

MoTrPAC is an NIH Common Fund program tasked with creating a "map" of the molecular changes that occur during and after exercise with the goal of understanding the mechanisms by which physical activity improves health and prevents disease. The $200 million project funds 3 preclinical animal study sites, 7 clinical centers, 7 chemical analysis sites, a consortium coordinating center, and the BIC, our hosts for this site visit. The consortium coordinating center provides overall management of consortium activities, including protocol development, establishment of standards for data collection, intra-consortium communications, and general administrative tasks. The preclinical animal study sites and clinical centers carry out acute exercise testing, exercise training programs, and collection of biospecimens and other physiological measurements in rats and humans, respectively. The chemical analysis sites then process the biospecimens and generate a variety of omics data, which are passed to the BIC for processing, QC, analysis, and distribution to the scientific community.

The human studies arm of MoTrPAC is a mechanistic randomized controlled trial consisting of 2280 adult participants (1980 sedentary, 300 highly active) and 300 pediatric participants. All adult participants undergo baseline acute exercise testing and biospecimen collection. For the acute exercise test, muscle and adipose biopsies and blood samples are taken at various time points surrounding a bout of acute exercise. A variety of physiological measurements are also taken, including basic history and physical exam, graded exercise test with 12-lead ECG, fasted blood screening, anthropometric measurements, bone density scans, and behavioral questionnaires. The 1980 sedentary participants are randomized to one of three intervention groups: endurance (n=840), resistance (n=840), or control (n=300). They undergo a 12-week supervised exercise program (endurance and resistance) or no intervention and then are subjected to a second round of acute exercise testing and biospecimen collection. A similar study design is being applied to the pediatric participants, but with a smaller number of participants and no resistance arm. In total the trials will collect ~53,100 blood samples, ~19,500 muscle biopsies, and ~9900 adipose biopsies. In addition to the core studies described above, a variety of ancillary studies will be carried out.

The rat arm of MoTrPAC consists of two phases, an acute exercise time course study and a training time course study. For the acute exercise time course, tissue samples are collected at seven time points following an acute bout of exercise. Including controls, tissues will be collected from 216 rats (12 male, 12 female per time point; 24 male, 24 female controls). For the training time course, tissue samples are taken after four different lengths of training (120 total rats - 12 male, 12 female per time point; 12 male, 12 female controls). For both phases of the rat studies, 21 tissues are collected from each rat, including blood, heart, lung, kidney, and brain.

A diversity of omics data are being generated from the collected human and rat tissue samples, including genomics (RNA-Seq, ATAC-seq, Methyl-cap, RRBS, and WGS), proteomics (global, phospho, acyl/acetyl, and redox), and metabolomics (targeted and untargeted reversed phase chromatography, positive and negative HILIC, and lipidomics) data. There is great variation among the many MoTrPAC sites in the technologies being used to generate the omics data, including heterogeneity within data types and some proprietary technologies.

The BIC is responsible for the daunting task of harmonizing the large quantity and diversity of data and metadata being generated by the consortium and performing meaningful integrative analyses across these omics data types. Although they are fairly early in the process, they have made significant progress towards these goals.

Program Lifestage

The MoTrPAC BIC is still in the startup phase of its lifecycle, approximately one year into its projected six year funding period. The full BIC team has only been in place for a few months, and most team members have been in place for less than a year. They have developed data processing, QC, and analysis pipelines for some of the data types they will be receiving. In addition, they have established a public data portal website (https://motrpac-data.org/) and in October had their first external data release, which included raw data and counts from 6-month old rats who had performed an acute bout of endurance exercise.

Much of the BIC effort in their first year has focused on establishing relationships and interacting with researchers at the MoTrPAC research sites. They noted that this “social” aspect of the project has taken up more time than anticipated. The interactions with MoTrPAC researchers have largely been aimed at gaining a thorough understanding of the data that is being generated. There has also been a significant amount of back-and-forth communication regarding pipeline and analysis decisions, as well as data submission formats, standards, and timelines.

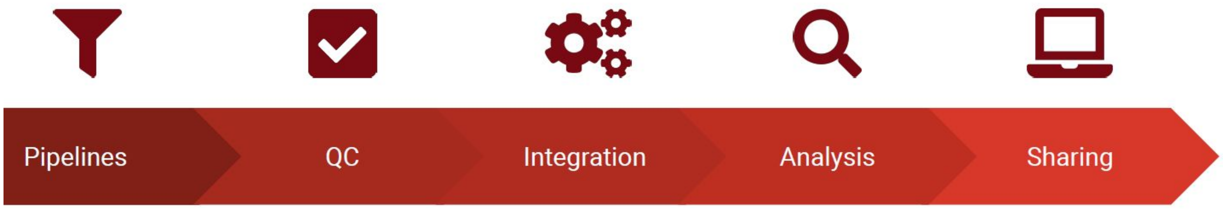
Another area requiring an unexpectedly large amount of effort in the first year has been dealing with the heterogeneity of sites, technologies, and techniques within the consortium. In many cases, there are multiple ways a certain data type is captured across sites, resulting in different formats for the data files the BIC receives. Adapting pipelines to work with the varying inputs has added an extra layer of complexity to an already complex endeavor.

Given the early stage of the BIC, their priority has understandably been ensuring that they deliver high quality data to the research community and that the pipelines and procedures they are developing are robust and appropriate for accomplishing the MoTrPAC goals. The immensity of those tasks has so far prevented focus on issues such as internal and external user support, ensuring that all FAIR principles are addressed, inter-DCC collaborations, and long-term sustainability, particularly with respect to data egress costs. However, there is enthusiasm for tackling these other issues if/when resources are available to do so.

Data Platform

The flow of data within the BIC is illustrated in Figure 1. When datasets are received, they are first processed using standard pipelines specific to the data type. The data then undergo quality control checks and are integrated with existing datasets via multisite intra- and inter-omic analyses. The next step is to analyze the data both within the dataset, e.g., differential expression and enrichment analyses, and in the context of other datasets in integrative multi-omics analyses. Finally, both raw and analyzed data are shared with internal and external users.

Figure 1. Data flow within the MoTrPAC Bioinformatics Center.



The data submission process involves MoTrPAC researchers uploading datasets and associated metadata to a Google Cloud bucket. When a dataset is uploaded, BIC analysts receive a notification and the data undergo basic automated QC and completeness checks. BIC analysts then work with the data submitter to resolve any issues, fill in missing information, and get clarification where necessary. The data then enter the process described in Figure 1.

The primary endpoint for accessing the MoTrPAC data is the MoTrPAC Data Hub (https://motrpac-data.org/). The first data release occurred in October and consisted of raw data and counts from the rat arm of the study. Future releases will contain human data as well as analyzed data from both species. The data release buckets are static and contain data, metadata, and documents describing any processing done on the data. Advanced users can load datasets directly into their data analysis environments from the data release buckets.

A variety of tools for searching, analyzing and visualizing the MoTrPAC are under development or planned for the future. The first tool that will be released is a faceted search tool to assist users in identifying datasets of interest. A shopping cart feature for assembling data downloads is also in development. In addition to accessing data directly, there will be tools for customizable web-based data visualizations provided on the site in the upcoming year.

Infrastructure

The MoTrPAC BIC uses the Google Cloud Platform (GCP) to host their systems and data. Among the GCP services the BIC utilizes are Buckets for data storage, in addition to Google Compute Engine (GCE) for website VMs and services (including pipelines), Cloud DNS, and Google Container Registry. The choice of GCP was largely driven by the fact that Stanford has an existing Business Associates Agreement (BAA) with Google for PHI data, but not some of the other major Cloud providers (e.g., Amazon Web Services). The BIC has not worked with the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative in setting up their Cloud services and has no immediate plans to do so.

The technology stack the BIC utilizes for their data pipelines, QC, and analysis includes the following components:

- **Workflow Definition Language (WDL).** Workflow language used to define data-processing pipelines.
- **Cromwell Server.** Workflow management system that can run WDL scripts and supports GCP and other cloud platforms. Exposes REST endpoints for managing workflows. The plan is to move to a tool called Caper (Cromwell Assisted Pipeline ExecutoR) which is a wrapper Python package for Cromwell.
- **Docker.** Provides containers configured with software and dependencies needed for running pipelines.
- **GitHub.** Version control and hosting for pipelines and notebooks.
- **Jupyter Notebook (Python/R).** Full documentation of methods in an executable notebook format.
- **GCP.** Hosting of data, infrastructure, services, and Cromwell server.

For the MoTrPAC Data Hub, some of the technologies being used are as follows:

- **Terraform.** Tool for building, configuring and managing infrastructure required for the web portal.
- **React/Redux.** Javascript-based web framework.
- **Auth0.** Single sign on (SSO) Identity as a service (IDaaS) solution which can be configured to allow login from multiple providers.
- **D3.** JavaScript library for generating real-time dynamic charts.
- **Services.** Various services are used for functionality of the web site. For example, a service was created by the MoTrPAC developers to generate temporary signed URLs for data downloads.
- **APIs.** Used to retrieve data from other data resources with public-facing APIs . For instance, the MyGeneInfo API provides gene details that are displayed on the site.
- **GitHub.** Version control and hosting for code.
- **GCP.** Web and data hosting.

The BIC plans to use dbGaP to distribute human sequencing data, but have not yet initiated this process. Deposition to a variety of relevant external resources such as Metabolomics Workbench are planned.

Analysis

The MoTrPAC BIC supports all levels of data analysis for the MoTrPAC consortium, including processing of raw data, first- and second-level QA/QC and normalization, multivariate analysis, differential expression analysis, and integration of multiple omics.

The BIC has devoted a great deal of effort in their first year towards developing standard data processing pipelines that are portable and ensure reproducibility. They noted that their goal is not necessarily to create pipelines that are compatible with those used by other resources, but rather to create pipelines that do what is required for the purposes of MoTrPAC. That being said, they have generally tried to utilize and adapt existing pipelines where possible.

The most mature pipelines at this point are those for RNA-Seq, Reduced representation bisulfite sequencing (RRBS) and ATAC-seq. The pipelines start with raw FASTQ files and process all the way through QC and metrics reports. Originally, the BIC planned to use the ENCODE pipelines for all of these data types, but found that they did not fit the MoTrPAC needs for RNA-Seq and RRBS. For these two data types, they instead adapted Snakemake pipelines created by Mt. Sinai to the WDL/Cromwell framework. At the time of our visit, they had already processed 1280 samples with their RNA-Seq pipeline. For ATAC-seq, the ENCODE pipeline proved to be sufficient for their needs.

The proteomics pipelines pose a greater challenge than the sequencing pipelines due to proteomics being a less mature field in general. The BIC has decided to adapt Windows pipelines created by PNNL, a leader in the field of proteomics and MoTrPAC consortium member, to the WDL/Cromwell framework. These pipelines are under development.

The most challenging pipelines to develop are those for metabolomics. There is currently little consensus within the MoTrPAC consortium about what should be done with these data. Making things more complicated, there are 6 different sites generating the metabolomics data, all with different instruments and software. The BIC is actively working with the Common Fund Metabolomics Program as they develop their approach to the metabolomics data.

The BIC will eventually make all of their data processing and analysis pipelines public through GitHub. However, they are concerned that doing so might increase their support burden as more people use, and have questions about, the workflows. They do not have the personnel/budget to provide that kind of support. They cited ENCODE as an example of a provider of open source pipelines who gets pinged with a lot of questions. Providing this type of support is something the BIC would definitely like to see the CFDE tackle.

For data analysis, the BIC will perform a variety of analyses within and across datasets and make results available to the research community, but also eventually plans to provide tools with which users can perform their own analyses of the MoTrPAC data. The types of activities that the BIC intends to undertake include the following: QC, normalization, merging of datasets, comparison of data from different sites, differential expression analyses, enrichment analyses, interomic factorization analyses, and meta-analyses. Their activities will be aimed at answering the research questions of interest to MoTrPAC. As of now, no decisions have been made regarding what analysis platform and tools to provide users for their own investigations.

While the primary focus of the BIC is to cover the specific analysis needs of the MoTrPAC consortium, they are open to engaging in cross-DCC analyses provided that additional funding is provided to cover these efforts. In addition, they would like to see the CFDE identify cross-DCC collaboration opportunities and take the lead in organizing such efforts.

Access

As discussed earlier, the BIC provides access to MoTrPAC data via the web-based MoTrPAC Data Hub. Their goal is to rapidly release raw and analyzed data to the greater research community. To download data, non-consortium users must create an account and sign a Data Use Agreement, which includes the following conditions:

- Users may not use the data for publications of any sort or publically host or disseminate the data prior to the embargo deadline. For the first data release, the length of the embargo is 15 months from the date of external release.
- Users are permitted to use the data for analyses supporting grant submissions prior to the embargo deadline.
- Users must cite MoTrPAC when using the data in a publication.
- Users must notify MoTrPAC of publications that use the MoTrPAC data.

Currently, the BIC is paying egress costs associated with data downloads out of their budget. Due to the potential high costs associated with downloads of the raw data, the links to download the raw data are “hidden” to discourage unnecessary downloads, i.e., users have to click an extra link to get to them. The BIC recognizes that this approach may not be sustainable long term and that some other mechanism may be required to cover the egress costs.

Given that the BIC is still in the ramp-up stage of their lifecycle, they have not made concrete plans for other mechanisms of data access, e.g., providing a workbench such as Cavatica or Terra where users can bring their own data and perform cross-study analyses with the MoTrPAC data.

Harmonization and Metadata

At this point in their development, the BIC and the MoTrPAC consortium in general are focused on establishing standards and harmonizing data within the consortium and are not ready/do not have sufficient resources to address incorporating external standards or ontologies or mapping to other resources. Within MoTrPAC, IDs are assigned to all participants and biospecimens, and the provenance of all data is carefully tracked. They have developed standard protocols and forms for data collection across sites, e.g., ~80 forms for measurements in the human participants. They also use standard codes for tissues and other metadata. For data submissions, there are guidelines for what metadata to provide and a format for how the metadata should be supplied, but not metadata forms per se. The BIC is still in the process of surveying the data landscape and determining the specific metadata they will collect for their various data types and technologies. As mentioned earlier, the BIC is also developing standard pipelines for each data type to ensure uniform processing and enable integration across datasets. Overall, within MoTrPAC, the level of standardization is quite high. Their guiding principle in their approach to metadata and standardization is enabling meaningful analyses of their data.

One topic of conversation around harmonization was whether it is a reasonable goal to harmonize data processing pipelines across different DCCs to enable easy integration of data. The BIC put forward a number of arguments against such an approach (or at least major impediments):

- Different DCCs have different needs and data analysis goals, and pipelines must be tailored for the specific goals of each DCC.
- Pipelines are not static, and changes to “universal” pipelines would require reprocessing of all datasets at great cost.
- Pipelines don’t fully solve the problem of integrating across resources. Other obstacles, e.g., batch effects, are not accounted for by standard pipelines.

As an alternative, the BIC proposed having some tool or resource that compares analysis pipelines and provides an assessment of the compatibility of data emerging from different pipelines. They noted that knowing that pipelines are, e.g, 95% compatible is sufficient in many cases and a more attainable and cheaper goal than trying to apply a standard pipeline to every dataset across the Common Fund projects.

Sustainability

The BIC is funded under the U24 Cooperative Agreement mechanism for a period of six years. Given that the MoTrPAC BIC is in a very early stage in its lifecycle, the long-term sustainability of its resources has not been a major focus. That being said, the use of GCP and primarily open source solutions means that the resource is not intrinsically tied to Stanford and could be transferred to and maintained by the NIH or a third party following the end of the performance period. Obviously, given the necessary funding, Stanford could also continue to maintain the resource at the end of the performance period.

In terms of short-term sustainability, the BIC noted that several aspects of the project have taken a greater amount of effort than anticipated, which has limited the amount of their budget they have available for other key areas. Specifically, the extensive heterogeneity of data types, platforms, and software used by the research centers has added significantly to the burden on the BIC. In addition, the “social” aspect of the project has been far greater than expected, i.e., establishing relationships and working with the data generators around pipeline development, metadata, data submissions, and other issues. They also forecast that other, unstarted, activities mandated by their award are also likely to take more resources than their award was written for. Directives such as ‘put your data in dbGaP’ seem simple on paper, but require a lot of hidden work. Aside from the work of gathering the data and preparing it to meet dbGaP requirements, it can require weeks or months of time setting up dbGaP contacts, learning the database, and other ‘soft’ work to get ready to begin preparing the data. As a result, less explicitly mandated areas such as internal and external user support and training have by necessity had lower priority. Devoted attention to complying with FAIR principles has also not been possible. In addition, as discussed earlier, the BIC has concerns about the sustainability of their model for dealing with egress costs. These are areas where the BIC would welcome help from the CFDE.

Training

Internal

The BIC indicated that support/training *within* the MoTrPAC consortium is one of the areas where they could definitely use assistance. As noted above, interactions with the data generators have taken up much more time than they anticipated in the first year. There is a wide range of technical abilities among the data generators, including those who only use Microsoft Excel, and many need assistance with tasks such as format conversion and data submission. The extra time spent providing support to consortium data generators has interfered with the ability of the BIC to develop the technical solutions necessary for such a complex project and has made it extraordinarily strenuous to meet data release deadlines while maintaining high data standards.

External

Support/training for external users of the MoTrPAC data is another area where the BIC would like to have support from the CFDE, even voicing that they would be happy to have the CFDE fully take over this task. They expressed concern that they do not have the personnel or budget to handle what could potentially be a large volume of support requests from consumers of the data. One consequence of this concern is that the BIC is not planning to immediately make their data processing pipelines public because they know that they are not able to adequately provide external-to-MoTrPAC-consortia support for them at this time.

The BIC supports the idea of establishing a set of best practices guidelines that outline common pitfalls and misinterpretations for different data types and sharing these expectations with users who wish to publish with the MoTrPAC data. However, they were not overly concerned about data consumers misusing the data, i.e., performing analyses or drawing conclusions based on the MoTrPAC data that aren’t valid or statistically sound. The BIC staff indicated that they view this issue as being largely out of their control and not worth devoting much effort to. They are much more interested in creating and providing quality datasets than policing their users science.

FAIR

Complying with Findable-Accessible-Interoperable-Reproducible (FAIR) principles is another task for which the MoTrPAC BIC would be interested in receiving support from the CFDE. They expressed uncertainty about what rubric(s) are used to judge if a DCC is “FAIR” in practice and were concerned that they will be judged on FAIRness without fully understanding the criteria they will be judged on. In addition, they were worried about the fairness of applying uniform FAIR expectations across DCCs given that individual DCCs may be limited by restrictions specific to their project. As with training and support, the BIC felt that they do not have sufficient resources to address the FAIR principles given the high burden associated with carrying out their core responsibilities. They would be happy to offload FAIR compliance to the CFDE.

Cross-pollination

The MoTrPAC BIC is interested in the concept of DCC cross-pollination, but does not feel that they are currently in a position to shepherd or drive such efforts. In addition, they do not think that their budget would accommodate such efforts and that supplemental funding would be needed. They also feel that they lack sufficient knowledge of the other DCCs to allow them to easily come up with collaborative cross-cutting projects. The BIC would like to see the CFDE take the lead in this area and identify potential areas of collaboration, as well as provide supplemental funding for such endeavors.

Of note, the BIC PIs, Euan Ashley and Matt Wheeler, are also PIs for one of the clinical sites of the Common Fund Undiagnosed Diseases Network (UDN), which could facilitate collaboration between the BIC and UDN DCCs.

SSO (Single Sign-On)

The MoTrPAC BIC uses Auth0 as their SSO solution for accessing the data portal. Initially, they enabled login with Auth0 using Google credentials, but not eRA Commons or other credentials. However, they noted that they may have to switch to using Red Hat SSO (an enterprise version of Keycloak) due to a possible internal mandate from the Stanford Medicine Technology and Digital Solutions group. The BIC would prefer to stay with Auth0. At this point, login with Google credentials via Auth0 has been temporarily disabled until this issue is resolved.

Outcomes

Infrastructure and Resource Reuse

Like many of the other Common Fund DCCs, the MoTrPAC BIC has opted for a cloud-based infrastructure and primarily open-source technology solutions. Consequently, their system is inherently fairly portable and reusable. Some aspects are specific to their cloud provider, GCP, but equivalents of the GCP services they use are provided by most major cloud providers.

A good illustrative example of the reusability of the infrastructure being developed by the BIC is their data processing pipelines. The pipeline scripts are written in WDL and run using the Cromwell workflow management system, both of which are open-source workflow solutions offered by the Broad Institute. Unlike many workflow management systems, Cromwell can be used with many of the most popular cloud providers. Environments (containers) for executing the pipelines are created using Docker, which means that they can be run on any local or cloud system that is running Docker. Given these components, it would be fairly simple for others to set up their own environment for running the data processing pipelines.

Challenges

The MoTrPAC BIC identified a number of areas that have been challenging in the early stages or that they anticipate being challenging as they move forward. They also indicated several key areas where they think the CFDE could substantially contribute to their success. We engaged in a productive conversation about possible solutions for these challenges. Below is a summary of some of the challenges that were discussed. Many of these were described in more detail in earlier sections.

Startup. The BIC encountered a variety of unexpected challenges associated with the startup of their center that were time-consuming and prevented them from producing as polished a product as they desired for the first data release. These challenges included getting a handle on the heterogeneity of data, technologies and software across the MoTrPAC consortium and providing support to the data generators.

FAIR Compliance. As discussed earlier, the BIC identified concerns about how to ensure that their resource is FAIR compliant. On the most basic level, they are unsure about what FAIR means in practice and what the criteria will be used to assess the FAIRness of their data. They also expressed worry about applying the same FAIR standards across all DCCs without taking into account the restrictions that individual DCCs might be limited by.

Interoperability. The BIC is very supportive of the idea of making their resources interoperable with other databases and resources. One of the main challenges identified around this issue is that there is currently not even a standard mechanism to find out what assets are maintained by each CF program.

Internal User Support & Training. The MoTrPAC data generators have a wide range of technological abilities, and some need help with the most basic of data-related tasks. Supporting these internal users has occupied more of the BIC's time that anticipated.

External User Support & Training. The BIC is concerned that they lack sufficient personnel, resources and budget to adequately support and train external users of their data. As a result, they are hesitant to add to the support burden by, for example, making their data processing pipelines public.

Testing. There are several challenges the BIC faces with respect to testing. Given the complexity of the data, the support burden within the consortium, and difficulty getting data generators to comply with deadlines, they felt that their deadline for the first data release did not leave sufficient time for comprehensive testing. In addition, there are few researchers in the consortium that are qualified to assess the full catalog of MoTrPAC data, and they often do not have the time or incentive to assist in the testing process.

Harmonization. The BIC has some misgivings about the feasibility of standardizing data processing pipelines across resources. They are making use of existing standard pipelines where possible, but are finding that some modifications are necessary to suit the specific needs of MoTrPAC. Also, they have concerns about the implications of having a standard pipeline, for instance the cost of reprocessing all data when switching to a standard pipeline or when there are changes in the pipeline.

Infrastructure. The BIC indicated that they are hesitant to tie themselves to an analysis platform such as Terra or Cavatica because the world of cloud analysis tools is constantly shifting. As

MoTrPAC is young and has few end users asking for analysis tools, the BIC is concerned about the potential risks of choosing a platform now. They do not want to tie themselves to an analysis platform that may not exist in a few years due to loss of support, for example. Or to put a great deal of effort into engaging a platform now only to move to a new system that better suits their end users in a couple of years.

DCC Cross-pollination. The BIC is interested in cross-DCC projects, but they do not know much about the other Common Fund DCCs, and consequently identifying areas of synergy would be difficult and time-consuming. Nor do they have the budget or personnel to tackle this task.

Financial. In general, the BIC feels that there are a number of implied responsibilities and hidden costs associated with their role in MoTrPAC that are not accounted for in the budget because they were not made explicit. They gave several specific examples such as the hidden time and effort required to get data into dbGaP; the harmonization of pipelines with other sites, which could entail the expensive task of reprocessing of all the MoTrPAC data; and the issue of egress costs for data downloads, which the BIC is currently covering, but could become unmanageable depending on the level of interest in the MoTrPAC data.

Potential Solutions

We had a productive conversation about possible solutions to the challenges facing the BIC, and they were very forthcoming about the areas where they think the CFDE could most benefit their efforts. Below are some of the solutions discussed for the challenges described above.

Startup. There is a vast amount of experience among different institutions in setting up and managing DCCs, but there is not an easy way for new DCCs to access this base of knowledge. Even the process of establishing contact with the appropriate personnel at existing DCCs to ask questions can be daunting. We discussed the possibility of the CFDE establishing a knowledge base of accumulated DCC wisdom. Possible knowledge base content includes the following: expertise and contact information of personnel at existing or former DCCs; standard metadata fields for different data types; standards and ontologies; pipelines for data processing, QC and analysis; FAIR guidelines; and technological solutions for building biological databases.

FAIR Compliance. The BIC expressed interest in offloading the task of ensuring that their data is FAIR compliant. Part of the issue is a lack of understanding of what exactly being FAIR involves in practice. We discussed the idea of the CFDE establishing detailed FAIR guidelines that more clearly define for DCCs how to satisfy the FAIR guidelines. Providing clear criteria will allow the DCCs to better plan and budget for this responsibility.

Interoperability. To address the lack of standard mechanism for discovering the assets present Common Fund DCCs, we discussed creating an easily queryable API that provides asset inventories in a standard format.

Internal/External User Support & Training. The BIC would definitely like to see the CFDE take on a large part of the support and training burden for their internal and external users. Specific ideas included the following:

- Seminar providing an overview of Common Fund data resources
- Tutorials/webinars for a variety of tasks users might want to perform
- Centralized help desk

Testing. The CFDE could provide dedicated staff with the required expertise for testing releases. The BIC was very enthusiastic about this possibility.

Harmonization. The BIC expressed interest in a tool or resource that would provide an assessment of the compatibility of data produced by different processing pipelines. Alternatively/additionally, the BIC would like to see the creation of a set of “blessed” pipelines that would be used across resources. However, they worry about the cost of having to reprocess data using these “blessed” pipelines. The BIC also suggested that the NIH provide incentives to encourage compatibility of data from different resources, e.g., RFAs that promote integration of data from different sources.

DCC Cross-pollination. The BIC indicated that they would like to see the CFDE coordinate pilot projects between DCCs. The CFDE would identify opportunities for collaboration and cross-resource analyses and provide supplemental funding to the DCCs to carry out these projects.

Financial. There was general support for the CFDE providing supplemental funding to cover the costs of creating an integrated Common Fund data network. The budget for the BIC is largely committed to fulfilling their core responsibilities for the MoTrPAC consortium.

Potential Projects

The BIC would be interested in participating in cross-DCC analyses and collaborations, given that the CFDE identify areas of synergy and specific ideas for projects and provide supplemental funding to cover the costs. One specific idea that was mentioned was performing a meta-analysis of RNA-Seq data from different resources.

Game Changers

We discussed several different ideas that could be game changers for the CFDE:

Common Fund Data Ecosystem Data Ingestion/Egress System (CoFundIES). Karen Dalton proposed that the CFDE develop a system through which all Common Fund data passes. Her description of the system is as follows:

> Centralized Ingress/Egress location... similar to an Identity as a Service provider like Auth0 provider (which offloads the integration of over 40 different social and enterprise

identity services on the developers' behalf), CoFundIES provides a way for Common Fund Data providers to write to a standardized format and that data is distributed to appropriate third party/Common Fund/NIH resources. Each DCC can select the output destination(s) as mandated by their purpose.

Each Common Fund Program no longer has to find and maintain customized connections to each data resource. Data comes to them. This allows tracking and location and findability of data (both it's home DCC origin and where a subset of the information is sent in transit and will live [in perpetuity or simply for now, while the tertiary resource has funding]).

CoFundIES does not need to hold the data, except during the transfer/transformation process. This may be represented in a systems diagram, similar to the Pub/Sub system used in Apache Kafka.

Tool/Technology Library. The CFDE could develop standard implementations of tools and technologies (e.g., using Docker) that new and existing DCCs could use to rapidly add new functionality without reinventing the wheel. Karen Dalton proposed the concept as “Common Tools for Common (Fund) Tasks”

Harmonization Tool. The CFDE could develop a harmonization tool that performs some meaningful comparison of data processed using different pipelines and produces an assessment of their compatibility that includes a measure of statistical uncertainty.

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. Suggested topics to cover: 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?

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
- Move data to cloud
- Metadata harmonized within DCC
- Metadata harmonized with _________
- Metadata harmonized across Common Fund
- Implement new service/pipeline ____________
- Increase number of eyeballs at 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: as a group, or seperate, whatever is convenient for MoTrPAC staff

1 - 2pm Open discussion (with breaks)

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:

- 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 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?

NIH Cloud Guidebook:

- What would you like to see included?
- What would be better left to individual DCCs to decide?
- Would you be interested in contributing to it?

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

Day 2

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, Thoroughness checking

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

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

- datatypes they maintain
- formats etc of same
- tools / resources they think might be useful for the 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.