CIRM

CALIFORNIA'S STEM CELL AGENCY

CNS Consortium 2022 Workshop

California Institute for Regenerative Medicine

February 24 & 25, 2022

Virtual Workshop

DAY 2 SUMMARY – DATA INFRASTRUCTURE

(Day 1 summary is presented in a separate document)

Table of Contents

Executive Summary

Day 1 & Day 2

As articulated in its 2022-2027 strategic plan, CIRM's mission is to accelerate world class science to deliver transformative regenerative medicine treatments in an equitable manner to a diverse California and world. One of the three major themes of the Strategic Plan calls for the advancement of world class science by leveraging collective scientific knowledge to inspire collaborative research that addresses Californian's unmet medical needs. To achieve this vision, the goals are (1) to develop next-generation technology **competency hubs** and (2) to build **knowledge networks,** fostering a culture of open science.

The goal of the CNS Consortium Workshop was to solicit feedback regarding the feasibility, opportunities, and best approaches to realize these goals of CIRM's Strategic Plan. **Day 1** discussions focused on **Shared Resources Labs** and **Day 2** discussions focused on **Data Infrastructure**.

Discussions at this workshop were focused on applications to the central nervous system (CNS) as a use case, but the resulting initiative(s) would be implemented broadly across cell types, organs, and diseases.

Day 1 – Shared Resources Labs for Stem Cell-Based Modeling

There is abundant interest and expertise in the California research community to capitalize on the promise of stem cell-based modeling. To assess needs in the field and utility of a possible shared resources funding program, the goal of the first day of the workshop was to identify challenges related to stem cell-based modeling and how a network of Shared Resources Labs may help to address them. To consider outcomes and lessons learned from previous efforts, **Session I** featured speakers who described their involvement and use of **shared resources that CIRM created under Proposition 71**, i.e., Shared Laboratories, a human induced pluripotent stem cell (hiPSC) Repository, and a Stem Cell Genomics Initiative. These shared resources provided the stem cell research community with access to infrastructure, tools, datasets and training, and fostered effective collaborations between laboratories with different areas of expertise. This made stem cell research more accessible to researchers at a time when federal funding for human embryonic stem cell (hESC) research was restricted, and made entry into the emerging field of human pluripotent stem cell (hPSC, umbrella term for hESC and hiPSC) research accessible to a broader research community. Even after CIRM funding for these programs ended, many remained sustainable and valuable resources for California's stem cell research community.

Session II on the first day featured a moderated discussion with 22 subject matter experts and stakeholders who were asked to identify hurdles to effective stem cell-based disease modeling and explore potential strategies for CIRM to help researchers overcome these challenges through the implementation of Shared Resources Labs. Two major hurdles to meaningful stem cell-based modeling were considered: **(1) limited reproducibility of findings** and **(2) uncertainty about the predictive value for human biology and disease.**

Discussants emphasized that **limited reproducibility** across projects employing similar stem cell-based models poses a major hurdle to scientific advancement. Suggested approaches to improving reproducibility ranged from: (a) technical solutions, such as automation and standardization of materials and protocols, (b) networking among researchers to share best practices and protocols, to train those new to the field and to replicate studies across labs, (c) scaling research for increased statistical power, and (d) data-related considerations, such as sharing outcomes data using FAIR (Findable, Accessible, Interoperable, Reusable) principles, providing detailed and consistent metadata, and deploying machine learning for analyses.

Discussants argued that there remains a need to continue to innovate and **improve stem cell-based models to increase their predictive value**. To better understand how a stem cell-based model relates to human biology and disease, discussants pointed to a need for deep clinical phenotyping of cell donors and for obtaining molecular and cellular information from relevant post-mortem human tissues as ground truth for analysis of various omics datasets generated from hPSC-based models. Another approach involves validating a candidate hPSC-based model by testing whether a drug elicits cellular and molecular phenotypes in vitro consistent with the drug's known effects in vivo. Developing more predictive hPSC-based disease models may be achieved by e.g., integrating multiple relevant cell types to better mimic complex biology, but this may also add variability to the experiment. Discussants commented that standardization of simpler models and innovation toward more complex models may both be needed to advance the field.

Discussants considered two distinct goals for a possible Shared Resources Labs network: **(i) to drive innovation toward optimizing and standardizing cutting-edge stem cell-based models** and **(ii) to lower barriers of entry into the stem cell-based modeling field**. An argument was made that both goals could be pursued. In addition to effectively sharing stem cell-based modeling expertise, recommendations for approaches to **building a network of Shared Resources Labs** included providing access to well characterized unmodified and modified hPSC collections, providing access to new technologies and equipment that may be too expensive or specialized for a single laboratory to acquire, and providing help with navigating the relevant stem cell and gene editing intellectual property (IP) landscapes. Discussants emphasized the importance of creating a network of Shared Resources Labs that will be sustainable in the long-term.

Day 2 – Data Infrastructure

The second day of the workshop focused on evaluating the best approach to promoting data sharing in California and determining which role, if any, CIRM should have. The day was divided into 2 sessions. During the first session (**Session III**), presenters outlined the **principles of the data biosphere**, which provides a framework for creating an open, compatible, and secure approach to data storage and collaboration designed for biomedical research, and they described **examples** of scientific initiatives that have **successfully implemented Knowledge Platforms**, deploying a cloud-based data and software ecosystem based on data biosphere principles. The session wrapped up with an overview of the "Data Use Oversight System" (DUOS) to semi-automate and efficiently manage compliant sharing of human subjects data. The goal of Session III was to illustrate how the concept of a Knowledge Platform could be applied to address technical and collaborative needs of California researchers, many of which were identified during Day 1 of the workshop.

The second session (**Session IV**) was a moderated discussion dedicated to understanding the best approach to promoting data sharing in California. The discussion was framed around the current needs and obstacles for existing collaborative Knowledge Platforms. A pre-workshop survey provided insight into the limited knowledge with regard to Knowledge Platforms among respondents, who were mainly researchers in the regenerative medicine field but who also showed a general interest in and openness to sharing data and collaborating across laboratories.

The moderated discussion with 22 subject matter experts and stakeholders focused on considerations for the most optimal design and implementation of a potential collaborative Knowledge Platform and for designing a Data Coordination and Management Center (DCMC).

The Knowledge Platform would provide a cloud-based data and software ecosystem, including tools, applications, and data processing workflows, to allow collaborative analysis in a shared computing environment with defined data access and security protocols. For a Knowledge Platform to be successful, discussants commented that it is important to understand and consider the needs and goals of those who will contribute data and those who will use the data. To allow efficient data sharing, it is also important to address administrative and technical challenges related to accessing and retrieving data, and discussants stated that data processing standards, metadata specifications, and naming conventions for each of the anticipated data types should be addressed early in a program's execution to enable interoperability of data from different sources. Discussants argued for tiered metadata to accommodate core information needed to replicate analyses and ensure interoperability of datasets while also allowing optional metadata as needed for specialized experiments. Importantly, discussants emphasized that cloud-based collaboration is new to many researchers, and that those who generate and contribute data need to be supported during data submission, and researchers should be incentivized to collaborate in the cloud.

The final part of the discussion was focused on considering several options for structuring data coordination and management responsibilities among researchers who generate data and a DCMC that would be responsible defining the conventions used, and for storing the data and making it available to researchers. Discussants considered several DCMC models and favored a model that would entail the inclusion of data type-specific expertise within the DCMC and also within the sites that produce raw data.

Conclusion

Overall, the presentations and discussions during this 2-day workshop reaffirmed the needs of the research community for shared competency hubs and a collaborative data infrastructure. By supporting these resources, CIRM can help democratize data analysis, improve access to hPSC models for human biology and disease, and increase collaboration across laboratories with diverse areas of expertise.

Workshop Summary

The mission of CIRM is to accelerate world class science to deliver transformative regenerative medicine treatments in an equitable manner to a diverse California and world. One of the three major themes of the Strategic Plan calls for the advancement of world class science by leveraging collective scientific knowledge to inspire collaborative research that addresses Californian's unmet medical needs. To achieve this vision, the goals are (1) to develop next-generation technology **competency hubs** and (2) to build **knowledge networks,** fostering a culture of open science.

Box 1 - What is a competency hub? An entity that shares a specialized skill or resource (competency) at any stage of the drug development pipeline with other investigators in a collaborative manner.

The goal is to empower and connect California's research ecosystem and facilitate validation and standardization of research platforms.

Shared Resources Labs are one form of competency hub.

Box 2 - What is a knowledge network? Shared scientific knowledge across discovery, translational, and clinical research.

The goal is to maximize the impact of research by facilitating and incentivizing data sharing.

The Data Infrastructure (see Day 2, Figure 4) would enable cloud-based, collaborative analyses across data shared from CIRM-funded and other research through the creation of a **Data Coordination and Management Center** (DCMC) and a **Knowledge Platform**.

The goal of the CNS Consortium Workshop was to solicit feedback regarding the feasibility, opportunities, and best approaches to realize these goals of CIRM's Strategic Plan. **Day 1** discussions focused on **Shared Resources Labs** and **Day 2** discussions focused on **Data Infrastructure**.

Day 1 summary presented in a separate document

Day 2 – Data Infrastructure

The second day of the workshop focused on obtaining expert and stakeholder input to inform CIRM about opportunities for building knowledge networks (*see Box 2*) as envisioned in CIRM's 2022-2027 Strategic Plan. The goal is to maximize the impact of research by fostering a culture of open and collaborative science. This can be achieved by building a **Knowledge Platform** that would leverage CIRM-funded research outcomes by enabling cloud-based, collaborative analyses across shared data (*see Box 4 for definition of 'cloud'*).

Box 3 - A **Knowledge Platform (**aka Data Platform**)** is a cloud-based data and software ecosystem that provides tools, applications and workflows to allow collaborative shared analysis in a computing environment with defined data access and security protocols.

A Knowledge Platform supports a knowledge network.

Note: the term "platform" is used differently in different contexts. In this document, "platform" will only be used in the context of Knowledge Platform.

The day was divided into 2 sessions. **Session III** was a series of presentations to provide an overview of existing Knowledge Platforms, and **Session IV** was a moderated discussion among experts to inform CIRM about the best ways to build and operate a potential CIRM Data Infrastructure.

Session III: Data Infrastructure Overview and Examples

In a blog post in 2017, a group of researchers introduced the vision for a data biosphere, a framework of cloud-based data storage and computation designed for biomedical research, to overcome challenges related to efficient data sharing and cross-dataset analyses. The data biosphere concept has been adopted by several scientific initiatives. Presentations in Session III introduced the data biosphere framework and examples of user experiences with existing Knowledge Platforms.

III.A. Data Biosphere: An Introduction

Presentation

Benedict Paten, UC Santa Cruz; Brian O'Connor, Broad Institute/Sage Bionetworks; and Timothy Tickle, Broad Institute

Traditionally, sharing data involves copying and moving data to different physical or cloud locations, which becomes more expensive as the size of the dataset increases. This approach often entails redundant infrastructure and siloed computing. Furthermore, data security is difficult to enforce when data are downloaded by users and stored in numerous locations. A data biosphere framework provides a cloud-based environment to store, process, and analyze data in a central location, which facilitates cost-effective and secure research with large datasets for single-site and multisite collaborations.

The data biosphere framework was designed to create a more open and secure approach to data access and analysis. The guiding principles of a data biosphere are: (1) modularity of functional components, (2) community[*]-supported development of ideas, (3) open access to software and other tools, and (4) consistency with standards developed by coalitions such as the Global Alliance for Genomics and Health (GA4GH).

The modular design of the data biosphere framework is such that each functional component, including indexes of data assets, specialized search engines, data processing engines, analytical tools, cloud-based collaborative workspaces, and the datasets themselves, are universally compatible with other components, such that they enable flexible data analysis according to the specific needs of a given research project. The GA4GH provides many interoperability standards and resources for the data biosphere, thereby promoting collaboration and helping to break down data silos.

Cloud-based platforms provide industry standard identity and access management tools that improve overall security and enable uniform logging and auditing of data access. This improves a program's ability to track and monitor the users and the data they use as compared with traditional systems in which data are downloaded and used locally. GA4GH standards provide a means for users to work across knowledge platforms that follow data biosphere principles. Using a set of authentication and authorization standards known as passports, users may login one time and access data from multiple knowledge platforms that have each defined the data access permissions of individual users. In the data biosphere framework, the data assets are managed

[*] In this document, 'community' refers to research community

by individual programs and knowledge platforms adhere to findable, accessible, interoperable, and reusable (FAIR) principles.

The data biosphere framework provides a community-focused environment that enables users to share their work with other users under open science principles. This focus promotes a diversity of ideas, data, and tool creation by and for the broader research community. Ideally, a knowledge platform that embraces data biosphere principles would incentivize users to publish their results, analytical tools, and other derived works on the platform, and those works would become searchable and reusable by other community members. These tools could then be executed and the results could be produced in an identical manner as the original author.

AnVIL was presented as a real-world example of a Knowledge Platform that adheres to the data biosphere framework and is deployed in a Federal Risk and Authorization Management Program (FedRAMP) compliant cloud environment. AnVIL is community focused, emphasizes open science, and is assembled from modular components that are based on GA4GH standards. These include BioConductor, DockStore, Terra, BioData Catalyst, Jupyter notebooks, and indexed searchable data models provided by UCSC.

The NIH Cloud Platform Interoperability (NCPI) ecosystem was presented to highlight the value of interoperability across knowledge platforms. NCPI is a trans-NIH federated data ecosystem comprising 4 NIH platforms and 11 petabytes of data from 831,000 participants in various studies. Applying principles of the data biosphere framework, the NCPI ecosystem empowers users to analyze data gathered from four integrated knowledge platforms: the BioData Catalyst program, the AnVIL program, the NCI Cancer Research Data Commons, and the Kids First Data Resource Center. NCPI has focused on applying GA4GH interoperability standards to the interfaces between knowledge platforms. NCPI supports a search strategy that facilitates finding data that resides on different knowledge platforms, combining search results into a common format, accessing data through the NIH RAS implementation of the GA4GH passport, and retrieving data that are referenced in search results from multiple knowledge platforms.

III.B. User Experiences: Examples of Cloud Collaboration

Presentation: Collaborating in the Cloud – AMP PD/Terra

Matt Bookman, Verily; David Craig, University of Southern California; and Barry Landin, Technome

The Accelerating Medicines Partnership Parkinson’s Disease (AMP PD) is a consortium that was established as a public private partnership between the NIH and several organizations in 2018 to support target and biomarker discovery in Parkinson’s Disease (PD). To achieve this goal, AMP PD has followed the data biosphere approach to create a Knowledge Platform that grants researchers access to consortium data that are compiled and harmonized from eight independent studies and shared in a single cloud location. Bookman provided an overview of how AMP PD makes clinical, transcriptomic, genomic, and proteomic PD data more accessible to researchers by providing a cloud environment to conduct their analyses through workspaces in Terra. These workspaces include starter analyses developed by AMP PD and shared scientific analyses developed by the AMP PD community. Overall, this data biosphere approach has made

AMP PD data more accessible and secure, has made shared analytical tools broadly available to data users, and decreased the cost of data storage and computation for the AMP PD community.

Notably, AMP PD's Knowledge Platform contains more than 50 analysis notebooks and workflows in cloud-based workspaces that are shared with the research community, some of which are subsidized through AMP PD funds. Data explorers and visualization tools are configured for use with AMP PD datasets and can be reused by the community via open source licensing. AMP PD provides users with training webinars, tutorials, and workflows that are developed in the AMP PD environment and shared on AMP PD's Terra cloud platform and public Github repository. Users can collaborate in the cloud using shared notebooks and workspaces, and users are encouraged to present their work in community focused webinars hosted by AMP PD. Source code is available for shared processing and analysis tools, which enables other community members to reuse and adjust according to their needs.

Craig provided a case study demonstrating how to leverage AMP PD's Knowledge Platform to create a full variant aware alignment pipeline using raw RNA sequence data in conjunction with processed genomic variant call format (VCF) files. AMP PD's Knowledge Platform enabled processing and computation of a 200 Terabyte dataset at a lower cost compared to local computing strategies, although implementation of the workflow across the entire dataset required considerable optimization at the outset. The project group also yielded a visualization tool that allows for viewing and downloading of gene-level data. These data are also integrated with multi-omics data and clinical data from both AMP PD and the Foundational Data Initiative for PD (FOUNDIN-PD). Managing access to constituent programs' datasets is simplified using the Google cloud Identity Aware Proxy, which enables the user of these tools to work with any data they have access to in the Google cloud.

Presentation: Cloud-Based Collaborative Research in Neurodegenerative Diseases
Patrick Brannelly, ADDI

The Alzheimer's Disease (AD) Data Initiative (ADDI) was created to accelerate the use of large datasets in research on AD and related dementias (ADRD), and ADDI launched the AD Workbench (ADWB) to facilitate the sharing of datasets and analytic tools with the global research community. ADDI provides users with free collaborative workspaces and free cloud-based compute time, with specific aims to increase access to datasets for researchers of varying means, to bring diverse skill sets from the research community together, to address research problems like under-represented populations in research datasets, and to aggregate large datasets that cut across multiple disorders from AD to PD, Amyotrophic Lateral Sclerosis (ALS), and Frontotemporal Dementia (FTD). ADWB provides optimized data security and data provenance in a Health Insurance Portability and Accountability Act (HIPAA) and the European Union's General Data Protection Regulation (GDPR) compliant platform, through which data contributors may control access to the datasets that they contribute to the platform. Users obtain initial accreditation to access workspace resources through ADWB, and then request data directly from data owners. These contributors define the terms of access to their data and specify which of three types of connectivity supported by the Knowledge Platform should be used for their data. Brannelly outlined the types of connectivity that the ADWB supports. Centralized connectivity, in which data and metadata are hosted on the ADWB cloud, is the most accessible to data users, whereas distributed connectivity entails remote hosting and querying of data at a separate source location. In the distributed connectivity model, participant (tissue donor)-level data (aka record-level data) are transferred to trusted ADDI workspaces. Contributors can also opt for federated connectivity, in which data are hosted by the data owner, all participant-level data remain with the owner, and all queries and computational tasks are sent to the owner's compute environment for execution at the data source. In the federated connectivity model, only approved results are released back to trusted ADDI workspaces.

ADWB has 2,396 registered users from 80 countries and is highly accessible even in low- and middle-income countries. The largest group of users are academic researchers (n=744) and the breadth of user profiles also includes patients and family members, data scientists, bioinformaticians, pharmaceutical researchers, and clinicians. ADWB further encourages use of the platform by hosting global data challenges that seek to identify specific, high-priority questions that must be answered by the ADRD field.

Presentation: NHGRI Analysis Visualization and Informatics Lab-space (AnVIL)

Ken Wiley, NHGRI/NIH; and Cornelis Blauwendraat, National Institute on Aging/NIH

Wiley provided an overview of AnVIL, which is a cloud-based federated genomic Knowledge Platform supported by the National Human Genome Research Institute (NHGRI). It provides a unified environment for data management and computation, is based on the data biosphere modular platform principles, and leverages a wide variety of tools and resources including Terra, Bioconductor, Galaxy, Dockstore, Jupyter, RStudio, WDL, and UCSC's Genome Browser. These modules offer access to datasets and established analysis pipelines alongside other tools commonly used by the research community. AnVIL has also established a meta-portal with general information about the platform.

AnVIL has ingested data from more than 300,000 samples amounting to 4,000 TB of data, including genomic and phenotypic data as well as associated metadata. AnVIL promotes data democratization by reducing cloud service costs, providing user training and outreach, and incorporating scientific and technological advances into the platform.

In terms of governance, the AnVIL's External Consultant Committee (ECC), with representation from a wide range of institutions, assesses how AnVIL can better meet the needs of the broader research community. The platform is run by multiple staff members serving in co-leadership roles, and a set of Working Groups have been established to continually improve AnVIL's data portal, data processing, technical development, data access, phenotype metadata, and outreach.

One of these Working Groups, the AnVIL Outreach Working Group, has developed training materials, projects, evaluations, and videos to help acclimate new users to the platform. All materials are open source so that they can be integrated into existing educational courses. This group is also working with NIH's Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) program to reduce cloud costs associated with AnVIL use in order to make the platform more widely accessible.

AnVIL maintains multiple data access options. Importantly, because this effort is funded through a cooperative agreement with the federal government, members of the external research community own the data hosted by AnVIL. Two types of public access datasets are available:

open public access datasets, which require no approval for access, and controlled public access datasets, which users can apply to access via the Database of Genotypes and Phenotypes (dbGaP). These access requests are reviewed by a Data Access Committee, and a list of approved requestors is managed by dbGaP and provided to AnVIL staff. AnVIL also offers consortium-level access, in which members of a given research consortium can access data from other consortium members. This access is managed by a central point of contact (agreed upon by that consortium's members) who provides a list of authorized users to AnVIL staff who apply access to the data. Consortium data that are also registered through dbGaP are made available through controlled public access to researchers outside of that consortium. AnVIL has also piloted the Data Use Oversight System (DUOS, see next presentation below), which would streamline the process to request access to datasets for secondary analysis.

Blauwendraat described a practical example of how a research project using AnVIL aimed to create an accessible structural genomic variant reference dataset for ADRD. This reference dataset can be used to explore a variety of questions related to the genetics of ADRD in order to (1) assess the role of structural variants in ADRD, (2) resolve complex genetic regions of interest in ADRD, (3) assess the impact of structural variants on gene expression in healthy and disease states, and (4) investigate methylation patterns across samples and diseases. By leveraging the AnVIL platform, the research team was able to work with data files that are difficult to manage in local systems due to their size (nearly 5,000 TB in aggregate, including 4,000 TB of ingested data plus additional processing data), including long-read sequencing across complex regions of interest. In this case, the AnVIL platform offered a cost-effective solution for analysis of these large datasets using harmonized pipelines that had been developed by experts in the field, as well as an opportunity to share the processed data with the broader research community.

Presentation: DUOS & GA4GH Standards

Jonathan Lawson, Broad Institute

Increasingly, a major challenge to data sharing is navigating the complex web of restrictions on secondary data use (i.e., researcher using data that another researcher has gathered). Human subject datasets often have complex or ambiguous usage restrictions that are derived from the original consent form; this language must be respected when sharing data. Data use restrictions are often drafted by contributing institutions independently, which creates vast inconsistencies and requires significant effort to determine if the data should be broadly shared with researchers.

Lawson provided an overview of the Broad Institute's Data Use and Oversight System (DUOS), which is a semi-automated data access management service governing secondary use of human genomics data. Lawson described how with DUOS support, the GA4GH has developed a solution to simplify data sharing requests that includes a set of data use ontologies that can be applied to help automate key aspects of access management processes. These ontologies can be used in both consent forms and in data access requests, and whose combined use helps determine automatically whether the data access request falls within the scope of the established consent. Automated decisions are made available to the Data Access Committee for review and final approval. The GA4GH has published guidance to assist researchers and compliance officers with writing consent forms that enables DUOS access request automation. When data access requests are approved, data access can then be tracked using a digital passport.

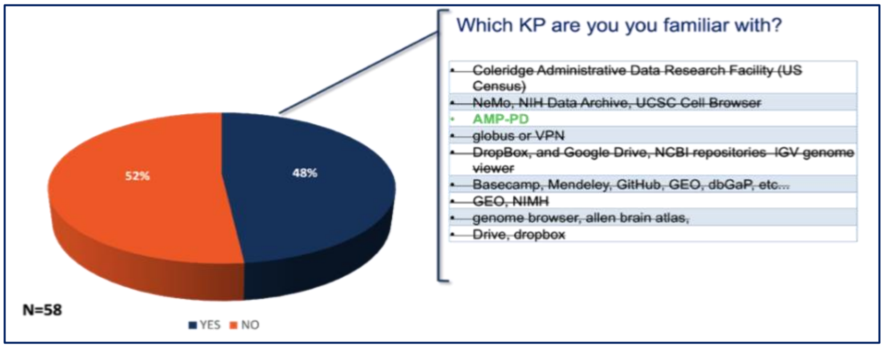
To request access to a program's data, researchers are often required to obtain a signature for a Data Use Agreement from their institution's signing official. This requirement can significantly delay the access request process. Data access approvals are expedited when signing officials provide a single broad endorsement for members of their institution, enabling researchers to directly submit their access requests.

Session IV: Moderated discussion

A moderated discussion was held to identify features that are essential to a Knowledge Platform and to consider tradeoffs of different approaches to data management. Following (A) an overview of the results from a pre-workshop survey, the discussion was divided into two parts: (B) Considerations for Designing a Knowledge Platform and (C) Considerations for Designing a Data Coordination and Management Center (DCMC), wherein expert discussants provided input on key features, questions, and recommendations. The Knowledge Platform discussion focused on challenges and solutions related to data access, data interoperability, and metadata requirements, as well as considerations for promoting adoption of a Knowledge Platform by the research community. The DCMC discussion focused on the distribution of responsibilities of researchers who generate data and dedicated data managers during data production, submission, quality control, and other phases necessary to prepare data to be shared with the research community.

IV.A. Pre-Workshop Survey Results

To start the discussion and to frame it within the context of California researcher needs, CIRM presented results from a survey of workshop attendees that was conducted prior to the workshop and was designed to gather the current needs for and obstacles to creating a collaborative data analysis platform. Nearly half (48 percent) of the respondents indicated they were familiar with the concept of a collaborative Knowledge Platform; however, when asked to name the platforms they were familiar with, only one of the answers provided (AMP PD) qualified as a program with a collaborative Knowledge Platform (Figure 1).



Which KP are you you familiar with?

- ~~Coleridge Administrative Data Research Facility (US Census)~~
- ~~NeMo, NIH Data Archive, UCSC Cell Browser~~
- AMP-PD
- ~~globus or VPN~~
- ~~DropBox, and Google Drive, NCBI repositories IGV genome viewer~~
- ~~Basecamp, Mendeley, GitHub, GEO, dbGaP, etc...~~
- ~~GEO, NIMH~~
- ~~genome browser, allen brain atlas,~~
- ~~Drive, dropbox~~

Figure 1: Familiarity with collaborative Knowledge Platforms. Workshop attendees were asked to indicate whether they are familiar with collaborative Knowledge Platforms, and if so, which platforms they are familiar with, n = 58. Crossed out indicates not a Knowledge Platform.

Survey respondents indicated a need for sharing and collaboration using various data types, including omics data (93.8 percent), imaging data (28.1 percent), electrophysiology data (15.6 percent), and clinical data (6.3 percent); 12.5 percent of respondents answered this question by citing a need for more data standardization (Figure 2).

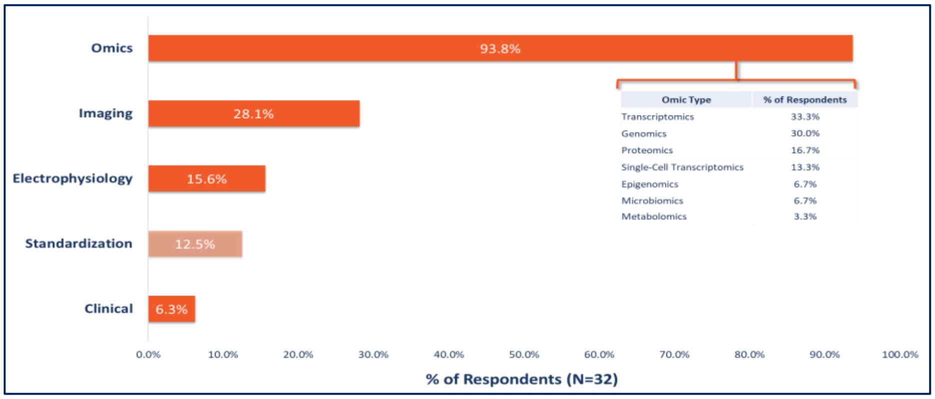


Figure 2: Preferences for shared data types. Workshop attendees were asked to indicate which data types require sharing and collaboration across laboratories, n = 32.

Most respondents (72 percent) expressed interest in accessing both raw and processed data. When asked whether there were any barriers to accessing and analyzing data on the cloud, 27.3 percent indicated that they experienced difficulties with technical, security, cost, or policy barriers (Figure 3).

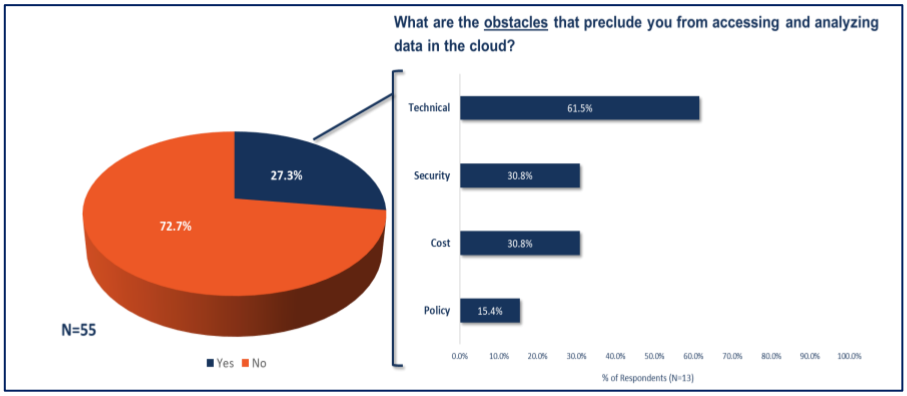


Figure 3: Barriers to accessing and using the cloud. Workshop attendees were asked to identify what barriers exist to accessing and analyzing data on a cloud-based platform, n = 55.

IV.B. Considerations for Designing a Knowledge Platform

The goal of this part of the discussion was to examine the features of a possible Knowledge Platform that empowers scientists with cutting edge data sharing, management, and software tools to harness the power of large-scale collaborative data analyses. Given the magnitude of biomedical research data that already exists and continues to be generated, discussants emphasized that a cloud-based system will be needed to feasibly enable effective data sharing and collaborative analysis approaches. However, working on data in the cloud represents a major shift in how biomedical researchers conduct collaborative science, necessitating a phased approach to implementing a Knowledge Platform and providing robust support and training of data contributors and future Knowledge Platform users. By supporting the creation of a Knowledge Platform that

Box 4 - “Cloud" or “cloud environment” refers to servers (computers) **that are accessed over the internet**, and the software (tools) and databases that run on those servers.

Prominent cloud service providers include, in alphabetical order

- Amazon Web Services
- Google Cloud Platform
- Microsoft Azure

attracts users because of its user-friendliness and the ease it provides to interrogate researchers' own data in the context of numerous other datasets, CIRM would effectively contribute to the advancement of world class science. The moderated discussion was therefore focused on a cloud-based Knowledge Platform.

A potential CIRM Data Infrastructure would include CIRM-funded data, stored in a cloud-based data repository, and would feature a **cloud-based Knowledge Platform** to enable collaborative analyses across datasets (Figure 4). The data is generated through research funded by CIRM's main pillars and harmonized to achieve interoperability within a CIRM Knowledge Platform and with other program's cloud-based Knowledge Platforms. Ideally, all data to be interacted with is hosted centrally in a cloud, but it is possible to interoperate with on-premise solutions as well if needed.

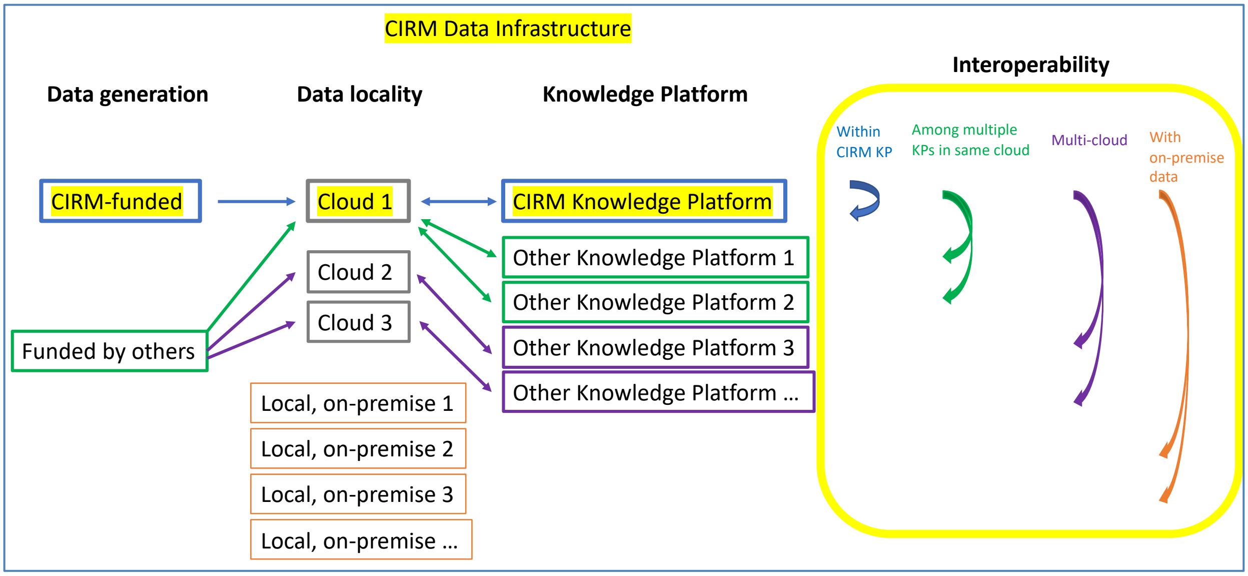


Figure 4: Yellow highlights illustrate components of a potential **CIRM Data Infrastructure** (data generation, cloud storage of data, and a Knowledge Platform), and its relationship to other data infrastructures. Various levels of interoperability are also illustrated. (KP; Knowledge Platform)

When designing a Knowledge Platform, a program should clearly define its purpose and understand the needs and goals of "data generators" and "data users" (†) when evaluating tradeoffs and making decisions. At the outset it is also important to understand what data types and datasets will be included in the Knowledge Platform, and what kinds of analyses should be supported. There should be a focus on today's needs, but plans should also prepare for future opportunities. For instance, many platforms currently target omics data, while imaging is not yet considered. Modularity and interoperability are critical to support change.

† In this document, "data generator" refers to scientists, such as wet lab researchers, whose research generates the data that populate a Knowledge Platform, and "data user" or "user" refers to researchers who use the data. Data users can be the data generators themselves or data scientists, such as computational biologists, whose research is focused on analyzing data generated by others. Different types of data users have different needs and goals.

Discussants recommended that a new Knowledge Platform adopt modular off-the shelf components over adopting an existing whole platform; these software components should use standards-based interfaces (software designed to enable communication between applications or modular components) to communicate with each other, to be interoperable with external Knowledge Platforms and to be interoperable with components of external Knowledge Platforms. Examples of off-the-shelf components include Terra, Dockstore, Galaxy, and various web portals that can be composed to form the Knowledge Platform. Discussants also commented that a program should not wait for perfection to implement their Knowledge Platform.

Several other critical aspects that need to be considered in the design of a Knowledge Platform were discussed, and include

1. Cross-Study Analyses – Data Access and Interoperability,
2. Metadata Required for Practical Data Use and for Data Harmonization,
3. Understanding Researcher Needs and Incentivizing Cloud Collaboration.

They are elaborated next.

1. Cross-Study Analyses – Data Access and Interoperability

Any collaborative analysis of data generated by different laboratories faces two major challenges, (a) accessing and retrieving data from different sources, and (b) the interoperability of data from different sources.

1a. Accessing and retrieving data from different sources

To co-analyze data from different sources, users need to navigate differences among data access policies. Discussants highlighted that the need for institutional sign offs can present a barrier to data access and for a program to negotiate blanket sign offs could help overcome this legal hurdle. Furthermore, data use agreements associated with different datasets would ideally be homogenous across studies. When they differ, a Knowledge Platform could use automated software tools to facilitate data access for a given collaborative study, such as the Data Use Oversight System (DUOS), to semi-automate compliant sharing of human subjects data described in section III.B. Discussants noted that certain governance policies like the European Union’s General Data Protection Regulation (GDPR) can make data access and transfers challenging.

Another important challenge relates to physically accessing or retrieving data that are stored in different locations. A key benefit for researchers to perform analyses within a Knowledge Platform is the fact that Knowledge Platforms are designed to retrieve data automatically from multiple sources. However, a researcher who seeks to incorporate additional data, not hosted by the Knowledge Platform they use, may have to develop code or use custom applications to retrieve the data from other locations before their analysis can begin. Discussants highlighted the importance of choosing a Knowledge Platform that can interoperate with others and that industry's tendency to silo will be countered, as more customers choose those platforms that emphasize interoperability. An important distinction relates to whether desired datasets are located on the same cloud, a different cloud, or whether they are hosted on local, on-premise servers.

Discussants described two approaches for accessing or retrieving data. One approach is to copy a dataset from one data server to another (move the data); this approach is called data mirroring and is considered practical when there is a sufficient number of anticipated users who will access the dataset in the new location. The second approach is a federated analysis approach, whereby researchers perform the same analysis on datasets that reside within several data locations (bring the compute to the data) and later combine the results; this approach is practical when the analysis tools are compatible within each source data environment. For a cross-study analysis, there is no single solution that addresses all researcher needs; each approach is a viable solution for different types of analyses.

Discussants remarked that multi-cloud support is not a priority for most Knowledge Platforms today, but not addressing multi-cloud will be a handicap for any Knowledge Platform in the future. They recommended that a program should start building a Knowledge Platform on a single cloud and add integration with other clouds later.

While multi-cloud analysis features need to be improved, discussants commented it is better for cross-study data analyses when the data are hosted in a small number of cloud environments rather than a large number of custom locations. However, discussants also noted that cloud environments are not available in many regions of the world.

1b. Interoperability of data from different sources

Box 5 - Interoperability is the ability of a dataset to be compatible with other datasets without special effort on the part of the user, through a system of shared standards and common ways of processing and using data.

For a cloud-based Knowledge Platform to enable collaborative analysis on data generated independently by multiple researchers, datasets need to be made compatible with each other. For example, different laboratories may use different data processing pipelines to align raw genomic sequencing data with a different reference genome. If, as is often the case, processed data (e.g. variant calls for genomic sequencing data) is shared for cross-lab analyses, discussants remarked that a Knowledge Platform should designate preferred analysis pipelines, but since it is unlikely that all data generators would agree to the same pipelines, information about analysis pipelines should accompany each dataset, and if custom pipelines were generated and used, the data processing code should be provided as well.

Besides information about data processing approaches, certain standard metadata (information about the data, such as demographic information about study participants, methods used to manipulate collected tissue samples, etc, see section IV.B.2) are required to ensure interoperability. Ideally, this minimal metadata, and the defined vocabulary used to describe it, should be defined collaboratively, and must be uniformly and consistently applied to submit data.

While one of the goals of a Knowledge Platform is to enable collaborative analysis of similar datasets for increased statistical power or comparative analyses, discussants highlighted that well harmonized data is required to conduct analyses across different diseases and across different data types, such as genomic, transcriptomic, proteomic, metabolomic, etc. data. Therefore, when defining minimal metadata standards, stakeholders should also consider interoperability across research areas and different data types.

While a Knowledge Platform may aggregate and harmonize ample data from many data generators, researchers may still want to access additional data from outside sources for their analyses. In addition to hurdles related to retrieving the data (discussed above), interoperability of data, such as the use of different data processing approaches or different metadata vocabularies, may hamper cross-study analysis. Datasets that are retrieved from different sources that are not part of a given Knowledge Platform need to be made compatible before the cross-study analysis can be executed.

As an example, many of the data sharing systems developed by the NIH were not designed to be interoperable; the NIH is working to change data access policies to allow systems to interoperate and to affect technical changes to retrofit systems with standards-based interfaces.

A program can manage data harmonization in a consistent and efficient manner one time, obviating the need for every researcher who is interested in using that data to harmonize it themselves.

2. Metadata Required for Practical Data Use and for Data Harmonization

Metadata is data that provides information about experimental data, to meet researchers' needs when interpreting results (practical data use) and to enable collaborative use of multiple datasets (data harmonization to achieve interoperability). Discussants categorized metadata broadly into three tiers:

i. Minimal core metadata, that are defined across the Knowledge Platform and must accompany all data; these metadata are needed to meet researchers' needs, replicate analyses, and ensure interoperability of datasets (e.g., information about data processing pipelines, information that accompanies the original tissue sample, such as tissue donor's demographic and disease diagnostic information and consent);
ii. Metadata that are also defined across the Knowledge Platform but are only included for specified experiments or collaborations or may otherwise be optional; these metadata are needed to support specific research questions (e.g., information about differentiation protocol, information about genetic modification protocol); and
iii. Optional metadata that are defined by data generators.

Discussants emphasized the importance of ensuring that defined, minimal core metadata are uniformly applied to all data contributed to a future CIRM Knowledge Platform. Similarly, participant identifier and sample naming conventions are important to define at the onset, as affecting a change becomes costly to institute later. Metadata standards should not be over-designed and where possible, a CIRM Knowledge Platform should work with trusted organizations, such as the Global Alliance for Genomics and Health (GA4GH) for genomic data, to adopt metadata standards. When defining metadata specifications, discussants noted that the needs of both human and machine interpretation (machine learning) should be considered.

Discussants noted that data processing and sharing of some data types, such as transcriptomes, is becoming standard (e.g., work on RNA standards through ENCODE), and a program should have a governing body of domain specialists who should be able to agree on data processing and metadata specifications, address changes in their field, affect changes to a program's metadata specifications at an appropriate pace, and help data generators prepare for the effects when more significant changes are required.

While changes to core metadata specifications should be exceedingly rare, a program should expect some change to occur as standards within specialized fields evolve. For newer and more specialized data types, such as imaging or spatial transcriptomics, there are options in data generation that are important to address collaboratively and agree to for a program, since the resulting data and a researcher's ability to interpret them can be significantly impacted by those decisions. Therefore, to keep pace with a given field, certain metadata or processing standards for a program may have to be updated more frequently.

To promote metadata consistency and quality, methods to support data hygiene during data generation, such as the use of electronic notebooks, should be promoted, and validation processes should be implemented to prevent incomplete or erroneous data submissions to the data repository. The DCMC also plays a role in harmonizing data and monitoring metadata quality before they are released or shared with the community.

In cross-laboratory data analyses, it is important to recognize when data obtained by different researchers originated from the same study participant. Discussants considered the use of a global unique identifier (GUID) as is the practice for the National Institute of Mental Health (NIMH) Data Archive. While ideal for connecting different data from the same study participant, GUIDs raise privacy concerns. CIRM could learn from NIMH's GUID approach when navigating the required regulatory processes and approvals.

3. Understanding Researcher Needs and Incentivizing Cloud Collaboration

The cloud is ideal for building a data sharing community, but discussants acknowledged that cloud-based collaboration represents a major shift in how biomedical researchers work together, and user adoption will take time. While some data users, such as computational biologists, may be already interested in using cloud-based platforms, many scientists still need to be attracted to this concept, and a Knowledge Platform should be designed and implemented in such a way that researchers choose and prefer cloud-based analysis. Special attention needs to be paid to potential barriers to adoption and how to overcome them.

To address concerns related to moving data to the cloud, clear data access controls and sharing policies need to be implemented, and distinct policies should be developed for data generators who contribute data to a Data Infrastructure versus other end-use researchers.

Discussants touched on data security concerns for data residing in the cloud. While efforts are made to prevent unauthorized downloads, they acknowledged that ultimately it is impossible to prevent users from pulling data from the cloud. On the other hand, federated data analysis strategies (see IV.B.2), where the compute is brought to the data, have the advantage that data are never transferred to users, and only one copy of the data exists in one cloud.

Supporting data generators

Since cloud-based collaborations are new to many researchers, discussants suggested a phased approach to establishing a Knowledge Platform and stressed the importance of supporting data generators.

The defined processes to prepare data for submission to a data repository should be streamlined as much as possible, but still can be time- and cost-intensive and may differ from data management processes a researcher may use for their own in-house analyses. Data submitters should be prepared early so they collect and store the data in a way that lends itself to data sharing. In addition to appropriately budgeting for this cost, discussants emphasized that dedicated data wranglers should be part of a Data Infrastructure, to be available to assist data submitters, who may not always be well versed in the metadata or the purpose of the metadata fields. Data wranglers help navigate the submission process and ensure quality and adherence to metadata standards. Metadata quality should also be actively monitored at the level of the data repository itself.

The help of data wranglers is particularly important for submission of newer data types, like proteomics and metabolomics, where the platforms evolve more quickly, are less standard, and are more highly specialized.

Incentivizing researchers to collaborate in the cloud

Despite the current shift towards more open data in biomedical research, some academic researchers may continue to be financially disincentivized from using cloud resources to perform their analyses because they have access to on-premise compute resources at their institutions that are free for them to use. Discussants suggested that providing a meaningful bank of compute credits for use in a Knowledge Platform may help overcome this hurdle. For researchers interested in access to cloud-stored data, creating barriers to retrieving data from the cloud, such as making it technically difficult and passing on the data download (egress) cost, may further incentivize cloud computing.

When designing a Knowledge Platform, much attention should go into making it the preferred choice for users. Discussants suggested CIRM consider bringing in outside groups who are not invested in a particular solution or researchers with experience using existing platforms to provide guiding feedback. Some approaches for driving a cloud-first strategy include the development of easy-to-use cloud-based tools, such as visualization and explorer tools, and to create incentives such as developing a uniform data use agreement to cover all access requirements within a platform and other tools to manage the complexities of specialized governance policies for constituent study datasets. Users may also appreciate that the provenance of resulting analyses is more transparent through standardized metadata and those analyses are more readily shared in the cloud.

Enabling research opportunities that are highly desired but hard to implement may be a strong driver toward adoption of a Knowledge Platform. Discussants highlighted researchers’ interest in linking clinical research (clinical trial) and healthcare (electronic health record, EHR) data with omics data. Other potential research benefits of a shared cloud-based Knowledge Platform include the ability to compare data across different diseases, highlighting as an example

Parkinson's Disease, Alzheimer's Disease, and ALS, and across different data types including data from wearables. Harmonized data from multiple studies may also aid in efforts to improve reproducibility in a field of research (see Session II in Day 1 Summary).

Since researchers still typically prefer accessing external datasets by downloading them to their local servers, providing training for using a Knowledge Platform is important. A discussant suggested that training the next generation of researchers to use cloud tools over local on-premise tools would be an efficient way to foster adoption of cloud-based analyses.

IV.C. Considerations for Designing a Data Coordination and Management Center (DCMC)

The goal of this part of the discussion was to examine how a future DCMC may be structured to implement the data flow within the Data infrastructure.

Prior to the workshop, a poll was shared with discussants that described potential responsibilities of a DCMC; respondents were asked to mark whether these are services that a DCMC should always, sometimes, or never manage. The purpose of this exercise was to establish the types of activities a DCMC would typically manage and to frame the discussion of how those management activities should be distributed for a potential CIRM Data Infrastructure.

A super majority of respondents (66% or more marked "always", remainder marked "sometimes") recommended that a DCMC should manage:

- Data releases,
- Metadata specifications,
- Data repositories,
- Knowledge Platform,
- Site content,
- User registration,
- User support,
- Systems and data security, and
- Technical Vision.

Poll responses indicated less certainty (30 – 50% of respondents marked "always", remainder marked "sometimes") as to whether a DCMC should manage:

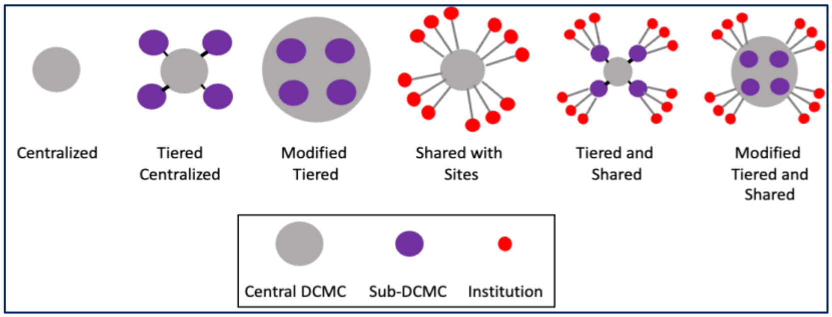
- Data transfers,
- Data harmonization,
- Cloud costs / budgets, and
- Cloud infrastructure.

Finally, there was no agreement from respondents (marking "always", "sometimes" or "never") as to whether a DCMC should manage

- Data generation,
- Data quality control (QC), and
- Systems and Data Governance.

To guide the discussion, the CIRM facilitator described hypothetical scenarios (models) for the division of responsibility for a DCMC. During this exercise, the audience is to assume that the raw data generation occurs outside of these models; data management under these scenarios begins the moment wet lab preparation transitions to dry lab raw data. Discussants were to consider the distribution of data coordination and management responsibilities as they will apply to further processing, preparation, and release for consumption by members of the research community.

The DCMC models considered during the discussion are depicted below. In the models, the centralized DCMC is represented by gray nodes, sub-DCMC tiers are represented by purple nodes, and data generation sites are represented by red nodes. These models represent the human architecture, as one might define in a Responsible, Accountable, Consulted, or Informed (RACI) matrix, and the discussion centered on the distribution of accountability across functions.



Fully Centralized

In a Fully Centralized framework, all data management functions are performed by a centralized DCMC. Several discussants expressed concern that a single, central DCMC would be unable to offer specialized expertise in processing the wide variety of data types that the CIRM Data Infrastructure might include.

Tiered Centralized | Modified Tiered

In a Tiered Centralized framework, specific data management responsibilities are assigned to specialized sub-DCMCs and a centralized DCMC is responsible for overall data management. A discussant suggested that sub-DCMCs could take the form of working groups composed of data generators and other experts for each data type, in a manner similar to the AMP PD program. Under this model, the DCMC is informed by experts in their fields but the responsibility to drive those working groups and the accountability for reaching decisions falls on the DCMC.

Noted benefits to the Tiered Centralized model include that all data types supported by the sub-DCMCs have established the infrastructure necessary for data ingestion, processing, and release/sharing; specialized expertise for each data type supported within sub-DCMCs may better serve the needs of data generators and data users for those data types; and specialized knowledge within sub-DCMCs may lead to the identification of novel ways to work with given data types, data sources, and tools that add value to the generated data.

One discussant noted that specialized working groups will be required regardless of the DCMC model and divisions of the DCMC into sub-DCMCs are a contracting or organizational concern more so than a concern of operational responsibility; a single

DCMC contract award would naturally result in a hierarchical distribution of operational tasks, given the divergent expertise needed for managing different data types. In other words, a Centralized DCMC would create its own sub-DCMCs to manage different data types (Modified Tiered).

While simpler to contract and manage, a drawback to combining the tasks into one master DCMC (Modified Tiered) is to lose the opportunity to compete each sub-DCMC award individually for optimal expertise. However, a single award to a Modified Tiered DCMC would have the advantage of a team forming organically.

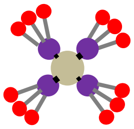
Overall, discussants felt that a drawback of the Tiered Centralized or Modified Tiered model was that this model does not address that accountability for several DCMC responsibilities including management of data generation and aspects of quality control should be distributed in-part to data generation sites.

Shared with Sites

Under the Shared with Sites model, certain data management functions would be managed by a centralized DCMC, while other data management responsibilities would be distributed to the sites (institutions) that produce raw data. A discussant suggested an example of this model is the Answer ALS program, wherein each site is responsible for defining the data and metadata specifications, generating and processing the data, and applying quality control standards to the particular type of omics data that it generates, and the centralized management team aggregates and manages releases after assuring those standards are uniformly applied. Several discussants (data generators) noted that this is a workable model and recommended that data generation and a degree of quality control functions fall under the responsibilities of the data generation sites, though the DCMC could define data and metadata requirements and quality control standards associated with those functions.

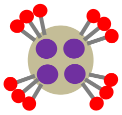
Because there will be multiple research groups generating the same data types, specialized sub-DCMCs may be required for data processing and management. The Shared with Sites model would require the central DCMC to develop and validate certain uniform quality control standards that are applied to all submitted datasets, and data generators would need to fulfill certain quality control standards prior to data submission. Data generators would have the necessary expertise to process the raw data, but data generators may struggle to agree on consistent processing practices without a sub-DCMC or working group to facilitate the development of a program’s standards. A discussant cautioned that more complex models, such as Tiered and Shared (see below) could result in confusion among data generators regarding allocation of responsibilities, noting the simplicity of the Shared with Sites model is attractive. Several discussants identified the need for collaboration amongst data generators to inform the specifications that all data generators for their specific data type would have to apply; the DCMC could facilitate those collaborations and be responsible for the definition of the specifications. All discussants who responded to the poll agreed that quality control is a shared responsibility that has requirements at the sites and requirements at the point of aggregation and organization for release to the research community.

Tiered and Shared | Modified Tiered and Shared



In a Tiered and Shared framework, some data management functions would be fulfilled by specialized sub-DCMCs while other data management responsibilities would be performed by sites who generate raw data. This design is a hybrid of the Tiered Centralized and Shared with Sites models that would harness the expertise of data generators while leveraging sub-DCMCs to facilitate consensus on specifications relating to specific data types.

A discussant indicated that this model, which is used by the NIH's Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative – Cell Census Network (BICCN), is complex, but also noting that this complexity may be unavoidable due to the degree of specialization of tools, data, or personnel for different data types. Each division of management responsibility creates a degree of added complexity, some of which may need to exist for a shorter period of time than others. The divisions may need to be flexible to support new data types or to serve distinct purposes that change over time. Interoperability is also more complicated when there are more entities and more components; when there are more entities, it becomes more important to clearly define applicable data, metadata, and interface standards early in a program's design.



Discussants offered a modification to the Tiered and Shared model that entailed the inclusion of data type-specific expertise within the centralized DCMC. This "Modified Tiered and Shared" model would have fewer formal entities, thereby reducing the number of contracts and operational overhead required of the infrastructure and enabling the DCMC to adapt new divisions to a program's evolving needs.

Overall, discussants favored the Tiered and Shared and Modified Tiered and Shared models, acknowledging the need for specialized expertise within the DCMC structure and the need for data generation site involvement in data coordination and management.

Concluding Remarks

CIRM thanks the presenters and discussants for the time spent to prepare and participate in the workshop, and very much appreciates the dynamic and informative discussions and the invaluable insights provided. The outcomes of this workshop, summarized in this document, will inform CIRM as we develop and implement our strategic vision to advance world class science.

Appendix A: Acronym Definitions

ADDI	Alzheimer’s Disease Data Initiative
ADWB	Alzheimer’s Disease Workbench
AnVIL	Analysis Visualization and Informatics Lab-space
AMP PD	Accelerating Medicines Partnership Parkinson’s Disease
BICCN	BRAIN Initiative – Cell Census Network
BRAIN	Brain Research through Advancing Innovative Neurotechnologies
CIRM	California Institute for Regenerative Medicine
CNS	central nervous system
DUOS	Data Use Oversight System
FAIR	Findable, Accessible, Interoperable, and Reusable
GA4GH	Global Alliance for Genomics and Health
hESC	human embryonic stem cell
hiPSC	human induced pluripotent stem cell
hPSC	human pluripotent stem cell
iPSC	induced pluripotent stem cell
NCPI	NIH Cloud Platform Interoperability
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
UCSC	University of California, Santa Cruz
VCF	variant call format

Appendix B: Agenda

February 24, 2022

11:00 – 11:25 AM Introduction, Background, Purpose, and Goals for the Workshop
Rosa Canet-Avilés, CIRM

Session I: Overview of CIRM-funded Research Resources

11:25 – 11:40 AM Overview of CIRM-funded Research Resources
Uta Grieshammer, CIRM

<u>Case Studies</u>

11:40 – 11:55 AM RFA 07-01: CIRM Shared Research Laboratory Grants and Stem Cell Techniques Course
David Schaffer, UC Berkeley

11:55 – 12:10 PM Leveraging Large iPSC Cohorts and Population Scale Stem Cell Models to Study the Effect of Genetic Variation on Cellular Phenotypes
Sulagna Ghosh and Ralda Nehme, Broad Institute

12:10 – 12:25 PM CIRM hiPSC Repository: NAFLD Lines for Disease Modeling
Jacquelyn Maher, UC San Francisco

12:25 – 12:40 PM CIRM hiPSC Repository: Machine Learning & Engineered iPSCs for Unraveling the Complex Biology of CNS Disease
Ajamete Kaykas, insitro

12:40 – 1:00 PM CIRM Genomics Stem Cell Hub: Experimental-Computational Collaboration to Characterize Cortical Organoids
Aparna Bhaduri, UC Los Angeles; and Max Haeussler, UC Santa Cruz

1:00 – 1:30 PM BREAK

Session II: Moderated Discussion – Building Shared Resources for Stem Cell-Based Modeling

1:30 – 1:45 PM Summary of Pre-Workshop Survey Results
Uta Grieshammer, CIRM

1:45 – 3:45 PM Discussion
Moderated by Uta Grieshammer, CIRM

3:45 – 4:00 PM Summary and Closing Remarks for Day 1
Rosa Canet-Avilés, CIRM

4:00 PM ADJOURN FOR DAY

February 25, 2022

9:00 – 9:20 AM Introduction to Data Infrastructure: Outcomes from September 2021 Expert Meeting
Rosa Canet-Avilés, CIRM

Session III: Data Infrastructure Overview and Examples

9:20 – 9:50 AM	Data Biosphere: An Introduction *Benedict Paten, UC Santa Cruz; Brian O'Connor, Broad Institute/SageBionetworks; and Timothy Tickle, Broad Institute*
9:50 – 10:00 AM	Data Biosphere Q&A

User Experiences: Examples of Cloud Collaboration

10:00 – 10:30 AM	Collaborating in the Cloud – AMP PD/Terra *Matt Bookman, Verily; David Craig, University of Southern California; and Barry Landin, Technome*
10:30 – 10:45 AM	Cloud-based Collaborative Research in Neurodegenerative Diseases *Patrick Brannelly, ADDI*
10:45 – 11:15 AM	NHGRI Analysis Visualization and Informatics Lab-space (AnVIL) *Ken Wiley, NHGRI/NIH; and Cornelis Blauwendraat, CARD, LNG, NIA/NIH*
11:15 – 11:30 AM	User Experiences Q&A
11:30 – 11:40 AM	BREAK

Data Access

11:40 – 12:00 PM	DUOS & GA4GH Standards *Jonathan Lawson, Broad Institute*
12:00 – 12:10 PM	Data Access Q&A
12:10 – 12:40 PM	LUNCH BREAK

Session IV: Moderated Discussion – CIRM CNS Data Infrastructure

12:40 – 2:40 PM	Discussion *Moderated by Rosa Canet-Avilés, CIRM*
2:40 – 3:00 PM	Summary and Closing Remarks *Rosa Canet-Avilés, CIRM*
3:00 PM	ADJOURN

Appendix C: Workshop Participants

Day 2 Presenters and Discussants

Anton Arkhipov, PhD, Associate Investigator, Allen Institute for Brain Science
Cornelis Blauwendraat, PhD, Investigator, National Institute on Aging
Matt Bookman, MS, Cloud Solutions Architect, Verily
Patrick Brannelly, MBA, Director of Partnerships and Business Development, ADDI
Jonah Cool, PhD, Science Program Officer, Chan Zuckerberg Institute
David Craig, PhD, Professor and Co-Director, USC Institute for Translational Genomics
Sonya B. Dumanis, PhD, Executive Vice President, Coalition for Aligning Science
Steven Finkbeiner, MD, PhD, Senior Investigator, Gladstone Institutes
David Glazer, BS, Engineering Director and Terra CTO, Verily Life Sciences
Maximilian Haeussler, PhD, Associate Research Scientist, UC Santa Cruz
David Haussler, PhD, Scientific Director, UC Santa Cruz Genomics Institute
Barry Landin, BS, Solutions Architect, Technome
Jonathan Lawson, BA, BS, Senior Software Product Manager and Data Access Committee Vice Chair, Broad Institute
Tetsuyuki Maruyama, PhD, Executive Director, ADDI
Brian O'Connor, PhD, Principal Investigator, Data Sciences Platform, Broad Institute
David Panchision, PhD, Chief, Developmental & Genomic Neuroscience Research Branch, National Institute of Mental Health
Benedict Paten, PhD, Associate Director, UC Santa Cruz Genomics Institute
Ekemini Riley, PhD, Managing Director, Aligning Science Across Parkinson's
Todd Sherer, PhD, Executive Vice President, Research Strategy, Michael J. Fox Foundation
Michael Snyder, PhD, Director, Center for Genomics and Personalized Medicine, Stanford University
Leslie Thompson, PhD, Professor, UC Irvine
Timothy Tickle, PhD, Head of Scientific Partnerships, Data Sciences Platform, Broad Institute
Kendall Van Keuren-Jensen, PhD, Professor, TGen
Ken Wiley, PhD, Program Director, Division of Genomic Medicine, National Human Genome Research Institute

Meeting Organizers

Rosa Canet-Avilés, PhD, Vice President of Scientific Programs, CIRM
Uta Grieshammer, PhD, Senior Science Officer, Discovery Program, CIRM
Mitra Hooshmand, PhD, Senior Science Officer, Special Projects and Initiatives, CIRM
Shyam Patel, PhD, Director of Business Development, CIRM