NCPI Virtual Touchpoint #2: Readout (Public Version)

June 5, 2024

Introduction

On the afternoon of June 5, 2024, the NIH Cloud Platform Interoperability (NCPI) program hosted the second Virtual Touchpoint; this event focused on the Interoperability Projects and was attended by Interoperability Project award institutions and companies, as well as NIH Office of Data Science Strategy (ODSS) NCPI Program representatives and individuals representing NCPI Working Groups. The meeting was supported by the NCPI Administrative Coordinating Center (NCPI ACC).

The goals of the second Virtual Touchpoint included:

- Continue the momentum of engagement from the NCPI Virtual Touchpoint #1 that occurred in January 2024
- Build understanding of current Interoperability Projects across the NCPI community
- Discuss common challenges, potential areas of integration, or opportunities for shared learning across the Interoperability Projects
- Seek topics and activities for the upcoming NCPI Workshop (September 2024)

This summary has been prepared by the NCPI Administrative Coordinating Center (NCPI ACC) and serves as a high-level synopsis of event, capturing key discussions, challenges and ideas shared during the session.

Participant Synopsis

Approximately two-thirds of the Touchpoint participants were from ICs or institutions not directly involved with NCPI program management or coordination.

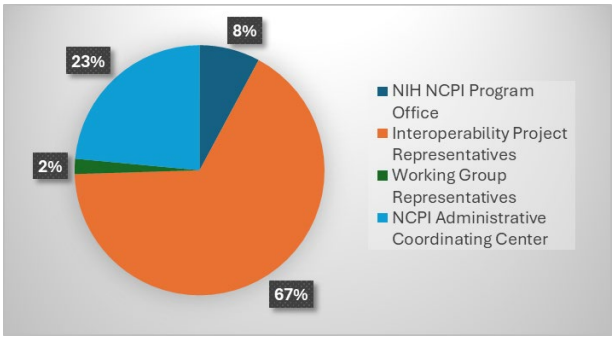

Figure 1: Participant Representation

Session Synopses

Welcome and Overview of NCPI Program

This session served as a brief welcome and introduction to the event by the NIH ODSS/NCPI Program Official (PO). The NCPI program is overseen by NIH ODSS and is a collaboration between NIH data repositories and analysis systems. This collaboration includes NIH representatives, analysis platform team members, and researchers using those platforms. The NCPI program exists to help realize the NIH's vision of a trans-NIH federated data ecosystem, facilitated by the collaboration of NCPI Partner Systems: AnVIL, BDC, CRDC, dbGaP, and KFDRC. The program's goal is to expand and connect additional data systems across NIH—aligned with the principles of data findability, accessibility, interoperability, and reusability (FAIR).

The purpose of this touchpoint was to introduce the five newly awarded interoperability projects and explore collaborative opportunities across projects. The Virtual Touchpoint served as a precursor for a deeper dive into this topic during the upcoming NCPI Workshop, which has been scheduled for September 25-26th, 2024. The workshop will expand on highlights from NCPI Interoperability Projects, explore the evolution of the federated data ecosystem, and revisit our NCPI guiding principles to ensure that NIH is providing data that is FAIR.

Introduction to Interoperability Projects - P01-002 AnVIL-BioData Catalyst Project

In the introduction to the AnVIL-BioData Catalyst Interoperability Project, the team provided an overview emphasizing the integration of two significant datasets: the GTEx and TOPMed's Freeze 1RNA. They started by contextualizing the use of Genome-Wide Association Studies (GWAS), which are traditionally powerful in identifying disease correlations with common genetic variants but are limited when addressing rare variants found in less than 5% of the population.

They highlighted the role of the GTEx project, which has compiled extensive DNA and RNA sequencing data from 838 healthy individuals across up to 49 tissues. While GTEx offers valuable biological insights, its scope has been primarily non-clinical. This project aims to address these limitations by leveraging GTEx data alongside the more clinically oriented TOPMed Freeze 1 RNA dataset, which contains genetic information, matched DNA and RNA data, from about 6,000 patients. This work will be conducted through two aims:

- **Aim 1:** Enhancing analytical capabilities by developing workflows that merge GTEx with TOPMed data. This work includes adapting Watershed, a probabilistic model

that integrates multiple genomic and transcriptomic signals to predict variant function, for use on cloud environments, making it suitable for application on the BioData Catalyst platform and enhancing its support for structural variants. The cloud-ready Watershed will then be used to identify rare variants in the GTEx and TOPMed datasets.

- **Aim 2:** This work will be enhanced by performing re-analysis of TOPMed and GTEx datasets, utilizing new references including the comprehensive telomere-to-telomere and human pangenome references, to improve the precision in alignments, variant calling, and RNA quantification.

Introduction to Interoperability Projects - P01-003 Facilitating understanding of shared disease mechanisms

The team introduced the two major scientific use cases for the project:

- **Aim 1:** Focus on phenotypic expansions of Congenital Diaphragmatic Hernia (CDH) and other structural birth defects utilizing data from Kids First and the Undiagnosed Disease Network (UDN) WGS.
- **Aim 2:** Investigate the genetic and environmental determinants that impact asthma severity in children.

A significant project goal discussed was the alignment on a shared representation of phenotypic data as well as facilitating data integration into the CAVATICA environment for downstream analysis. This interoperability focused work is crucial to support the outlined scientific use cases. Initial progress includes team onboarding within CAVATICA and preliminary efforts in FHIR harmonization.

The latter part of the session concentrated on Aim 2 (investigation of asthma severity in children). This aim seeks to analyze how the interactions between genetic and environmental factors contributes to asthma and asthma severity. The integration uses FHIR to query data from ImmPort, dbGaP, and the Gene Expression Omnibus (GEO) into CAVATICA for downstream analysis.

Initial Steps Taken:

- **Data Review:** Examined datasets across Eureka, ImmPort, and dbGaP, including whole genome sequencing data from about 1000 patients and various study files detailing environmental exposures and other clinical data.
- **Data Preparation:** Identified overlaps in data IDs between dbGaP and ImmPort and began preparation for data integration.

- **Data Access:** Secured access to dbGaP's whole genome sequencing data, essential for upcoming analyses.

Introduction to Interoperability Projects - P01-004 AnVIL/Velsera-CGC interoperability project

The overall goal for this project is to increase our understanding of the disparity in the higher rates of colorectal cancer incidence in Hispanic populations compared to Non-Hispanic Whites. This project seeks to elucidate the genetic and transcriptomic factors that could be contribution to this disparity. To achieve this, the teams are working to integrate datasets across multiple platforms. One of the major goals is to take genomic data sets from the Thousand Genomes Project on AnVIL, where 3202 samples that have been sequenced from every major continental population around the world and integrate those datasets with cancer-specific datasets brought in through the Velsera Cancer Genomics Cloud platform. The team has identified hundreds of Hispanic Colorectal Cancer patient datasets on The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) databases.

The three aims of the project include:

- **Aim 1:** Identify genetic and transcriptomic variations in Hispanic populations.
- **Aim 2:** Catalog variants associated with colorectal cancer.
- **Aim 3:** Identify variants linked to the increased risk of colorectal cancers in Hispanic populations.

In their presentation, the team elaborated on Aim 1 of the project, focusing on identifying genetic and transcriptomic variations in Hispanic populations. They highlighted the ancestry bias in human genetics, noting the predominance of data from European ancestries. To address this, the team analyzed RNA-seq data from 730 samples across 26 diverse subpopulations from the 1000 Genomes Project, aiming to understand gene expression diversity and the genetic basis across these groups. 'Hispanic' is not biologically defined, prompting their work on disentangling ancestry-specific effects. The team has begun estimating ancestral origins of individual samples and mapping genomic variations to gene expression or splicing variants across populations through association studies. Their diverse approach helps distinguish between variants causing observed changes and those merely associated.

Introduction to Interoperability Projects - P01-007 Connecting FHIR, the CDA and DRS across NIH cloud resources

The team highlighted their focus on the identification of cancer subtypes through The Cancer Genome Atlas (TCGA) and the establishment of the Tumor Molecular Pathology Working Group, which created machine learning models to recognize these subtypes. These models demonstrated potential in controlled environments and were further evaluated through a transfer learning experiment to assess their effectiveness on different datasets, such as transitioning from RNA sequencing to microarray data. This discussion underscored the challenges in applying models trained on well-annotated datasets to broader applications and the necessity of accessing relevant biological samples and data across NIH consortia.

To address the challenge of finding and accessing biological samples across NIH consortia, the primary objective of this project is to develop a FHIR aggregator. The goals of this aggregator would be to aid researchers in accessing a comprehensive index of all available FHIR endpoints, enable them to identify relevant samples for their research, and allow them to utilize APIs for data retrieval. The Cancer Data Aggregator serves as an effective model for this work, with the team seeking to extend these capabilities to non-cancer resources through a common data model. Initial surveys indicated that mapping the Cancer Data Aggregator to FHIR could be relatively straightforward, showing good alignment between the data schema and requiring minimal transformation.

Since the initial proposal, the team has advanced their work by developing additional tools to integrate external datasets into FHIR. They have successfully converted genomic data from the Cancer Data Aggregator and have begun incorporating other datasets, such as the Cellosaurus—a comprehensive cell line dictionary. This integration allows for mapping dose response curves and other experimental data, facilitating easier access and mapping of drug line experiments into FHIR so that researchers could include more comprehensive measures in their analysis.

Introduction to Interoperability Projects - P01-008 Integrated data exploration of AnVIL and KidsFirst

The overall goal of the project is to make discovery and analysis of data from both Anvil and Kids First possible. This goal is broken down into 3 specific aims:

- **Aim 1:** A common set of APIs that make data access across AnVIL and KidsFirst uniform.

- **Aim 2:** A harmonized subset of metadata across AnVIL and KidsFirst data, making combining datasets more straightforward.
- **Aim 3**: The ability to analyze arbitrary combinations of KidsFirst and AnVIL datasets within both the Velsera/CAVATICA and Terra/AnVIL platforms.

These interoperability goals are driven by two scientific use cases:

- **Scientific use case 1:** Sex differences in health and disease
- **Scientific use case 2:** Characterizing the genetic basis of Ollier disease and Maffucci Syndrome

The team then reviewed first scientific use case and emphasized the crucial role of understanding underlying sex differences in healthy tissues and their implications in disease scenarios. They highlighted the necessity of adopting a sex chromosome complement aware alignment approach to accurately detect and analyze sex-linked genetic variations. They discussed how not utilizing sex chromosome complement aware alignment can lead to significantly underestimating mapping of expression variants on the X chromosome. Next, the team discussed the need to find the DNA and RNA datasets to enable interrogation of possible sex differences. They noted that while this can be done manually looking across GTEx, KidsFirst, and TOPMed it can lead to researchers giving up on projects because they can't find the datasets. Thus, a goal of this project is automating that process of finding and bringing together datasets.

Facilitated Discussion on Common Challenges and Opportunities

Identity mapping: There was a discussion about potential challenges to identifying common subjects and samples across datasets and databases. For example, two databases may have different data from the same individual, which may undermine the statistical robustness of studies. Currently, the process of harmonizing identifiers across different datasets/subjects is done in a manual fashion—if we need to do this how do we ensure that this work is only done once? The data may not be good enough for a machine learning model to do this identity mapping, and data models/types/ID formats often differ between platforms. Participants also noted that there may be implications for potential re-identification of subjects (e.g., subjects with a rare genomic variant). A key distinction was observed between subjects that may be enrolled within multiple parts of a single IRB protocol, and subjects that may fall into multiple IRB protocols.

Communication: The group discussed the importance of communicating both across projects and externally with the larger community to get awareness on early successes, progress, and learnings. This starts from simple high-level awareness of what the projects are doing, accomplishing, and learning so that other researchers can apply the lessons to

their own use cases, as well as sharing outputs and code. Some tools discussed include: Github, regular project meetings, inviting projects to other existing meetings within the community, and a potential Interoperability Project Slack channel.

Standardized Formats: The project teams also discussed standards and best practices for sharing the specific back-end technical tools and code. The projects could benefit from deciding on a template, model framework, or structure for technical outputs to make it easy to read and apply them across projects. However, the project representatives expressed a desire to strike a balance. They didn't want to let perfectly organized Github repos or perfectly structured documents get in the way of sharing products and lessons learned regularly. They also explored the possibility of best practice software development, including Github actions, training implications, conformance testing standards, or a resource for new projects.

Generalization of Project Products: The teams discussed the importance of making sure that interoperability project outputs can benefit other researchers. This includes making models that consistently work in a variety of environments and could be easily applied to other use cases or datasets.

Upcoming NCPI Workshop

This short session focused on collecting ideas and discussion topics for the upcoming In-person NCPI Workshop. The Workshop has been confirmed for September 2024 in Rockville, MD. In-person attendance will be limited to Lead-PI and up to 2 Co-PIs per project. There will also be a virtual attendance option. More details will be shared with invitees when they are finalized..

Closing Remarks

The Touchpoint Closeout included a summary of touchpoint themes and next steps that emerged from the day's sessions. In particular, the closing remarks highlighted the opportunities to work together to identify data to reduce double counting, collaborate across projects through tools like Slack and GitHub, avoiding vaporware and perpetual development, and exploring conformance packages.