

Data Submission and Consortia Engagement

29OCT2021



Brian O'Connor and Frederick Tan



Overview



- Data submission
 - Current data ingestion system
 - Recent improvements
 - Onboarding progress
 - Vision for future evolution of data ingestion

- Consortia engagement
 - Current process
 - Consortia engaged
 - Vision for future engagement

12:35-1:50 Session 1: Breakout rooms

Data submission and consortia engagement <i>Moderators: Dr. Adam Resnick (Children's Hospital of Philadelphia) and Ms. Valentina Di Francesco (NHGRI)</i>		Analysis tools <i>Moderators: Dr. Marylyn Ritchie (University of Pennsylvania) and Dr. Ken L. Wiley, Jr. (NHGRI)</i>	
12:35-12:40	Moderator introductions	12:35-12:40	Moderator introductions
12:40-12:55	AnVIL presentation: <i>Dr. Brian O'Connor (Broad) and Dr. Frederick Tan (Carnegie)</i>	12:40-12:55	AnVIL presentation: <i>Dr. Vincent Carey (HMS) and Dr. Ira Hall (Yale)</i>
12:55-1:40	Discussion	12:55-1:40	Discussion
1:40-1:50	Prepare breakout report	1:40-1:50	Prepare breakout report

Breakout room: Data submission and consortia engagement

Moderators: Ms. Valentina Di Francesco and Dr. Adam Resnick

Dr. Elizabeth (Liz) Blue
Dr. David Crosslin
Dr. Iftikhar Kullo
Dr. Tara Matise

Dr. Aleksandar Milosavljevic
Dr. Minoli Perera
Dr. Steven (Steve) Rich
Dr. Kenneth (Ken) Rice

Data Submission



Submitting data - dbGaP versus AnVIL



- **dbGaP**

- Used by data submitters
- Phenotypes and genotypes into short read archive (SRA)

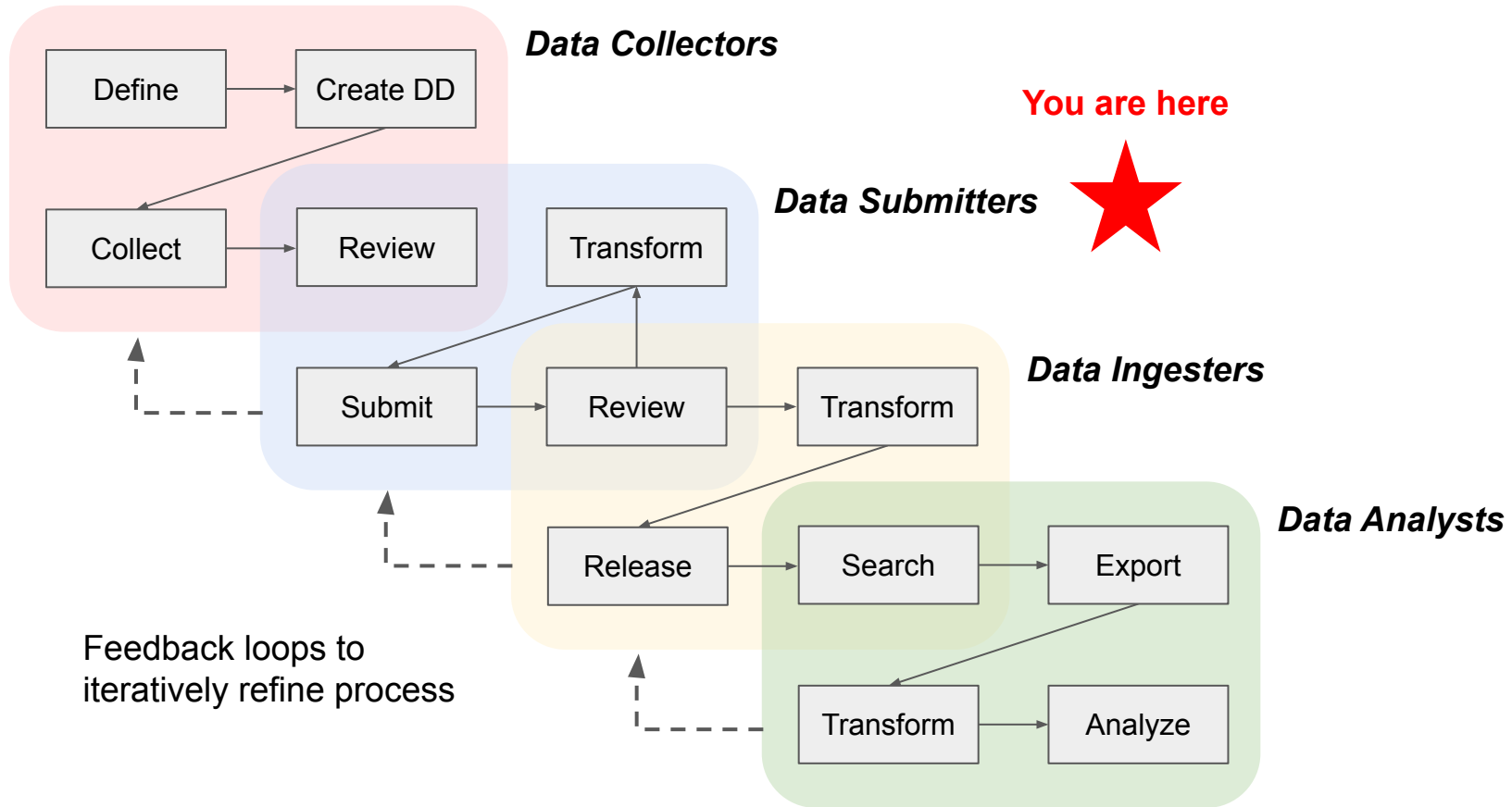


- **AnVIL**

- Used by data submitters (registered on AnVIL)
- Phenotypic data in integrated tables (CSV or TSV in workspace bucket)
- Genomic data files in workspace buckets



Data Lifecycle - Simplified Overview





Data Submission

- Step 1: Obtain Approvals
- Step 2: Develop Data Model
- Step 3: Prepare Data
- Step 4: AnVIL Data Ingestion

- **Prerequisite steps for the consortia**
 - Study registration
 - Map data to the AnVIL data model
 - AnVIL minimum data required
 - Define how data will be parsed
 - Per consent code
 - Per group, per consent code

The screenshot shows the AnVIL website's 'Learn' section, specifically the 'Data Submitters' page. The page title is 'AnVIL Data Submission Guide'. The content includes a welcome message, a goal statement, and an overview of the submission process. The navigation menu includes 'Overview', 'Learn', 'Datasets', and 'News'. The left sidebar shows a table of contents for the 'Data Submission Guide'.

AnVIL NHGRI Analysis Visualization and Informatics Lab-space

Search

Overview Learn Datasets News

Learn

Introduction Data Analysts Investigators Data Submitters

- ▼ Data Submission Guide
 - Submission Process
 - Overview
 - 1 - Register Study/Obtain Approvals
 - 2 - Set Up a Data Model
 - 3 - Prepare for Submission
 - 4 - Ingest Data
 - 5 - QC Data
- ▶ Data Submission Resources

AnVIL Data Submission Guide

Welcome to the Data Submitters docs on AnVIL. We're excited to have you here and helping to push the frontiers of biomedicine.

Our goal is to help researchers by hosting robust and large datasets and making it easier for researchers to find and analyze the data they need. By contributing datasets, you are helping us achieve this goal.

To make the data useful, especially for cross-study analysis requires standardized formatting and careful review. We are asking submitters to help us in this endeavor, by following the instructions in this guide.

Overview

In order to submit data into AnVIL you will need to do the following:

1. [Register with dbGap/Obtain required approvals.](#)
2. [Set up your data model.](#)
3. [Prepare your data for submission.](#)
4. [Ingest your data into AnVIL.](#)
5. [QC ingested data.](#)

1 Year AnVIL Planned Data Ingestion

Primary Column	Q4			Q1			Q2			Q3			Q4		
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1000G															
HPRC	HPRC														
- GTEx															
v9	v9														
open access	open access														
bi-sulfite sequencing															
recount3	recount3														
Genome in a Bottle	Genome in a Bottle														
T2T															
CCDG	CCDG														
NIA	NIA														
WGSPD & ConvergentNeuro															
Dementia Long-Read	Dementia Long-Read														
Identification of risk factors for ALS and FT	Identification of risk factors for ALS and FT														
CMG	CMG														
GAFK	GAFK														
TARN	TARN														
Clinical WGS	Clinical WGS														
GREGoR (MRGC)	GREGoR (MRGC)														
PMDG	PMDG														
eMERGE II & III															
eMERGE PRS	eMERGE PRS														
CSER	CSER														
IGNITE II	IGNITE II														
PRIME	PRIME														
COVID19 Prospective genomic stud															
ENCODE	ENCODE														
Genetic testing standard of care vs. clinical WGS	Genetic testing standard of care vs. clinical WGS														
NIMH - National Institute of Mental Health - InPSyght - Whole Genome D	NIMH - National Institute of Mental Health - InPSyght - Whole Genome D														

- Currently engaged with over 20 consortia
- Continued data submissions from established consortia
- Potential for new types of data in AnVIL
 - Imaging
 - RNA-seq
 - Proteomics

Nearly 300,000 Available Genomes

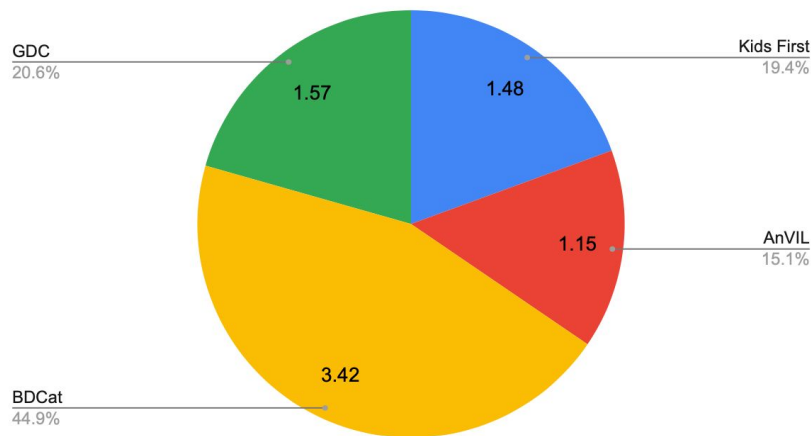
Consortium	Cohorts	Samples	Participants	Size (TB)
1000 Genomes	1	3,202	3,202	73.00
CCDG	198	272,306	256,318	2,623.69
CMG	41	18,593	16,599	97.15
Convergent Neuro	2	304	304	5.32
GTE _x (v8)	1	17,382	979	182.00
HPRC	1	57	47	195.00
PAGE	4	690	690	17.00
T2T	1	0	3,219	503.00
WGSPD1	5	1,504	9,943	176.85
Totals	254	314,038	291,301	3,873.01

Current data ingestion process

- In one year we have increased the amount of data in AnVIL ~4 fold
 - 1 petabyte to 3.9 petabyte of data
- Process is becoming more automated but still requires hands-on-keyboard work by the AnVIL project managers
- New onboarding slides standardized the information and steps of AnVIL onboarding
- Review meetings with consortia to gain new insights on how to improve onboarding and data submission

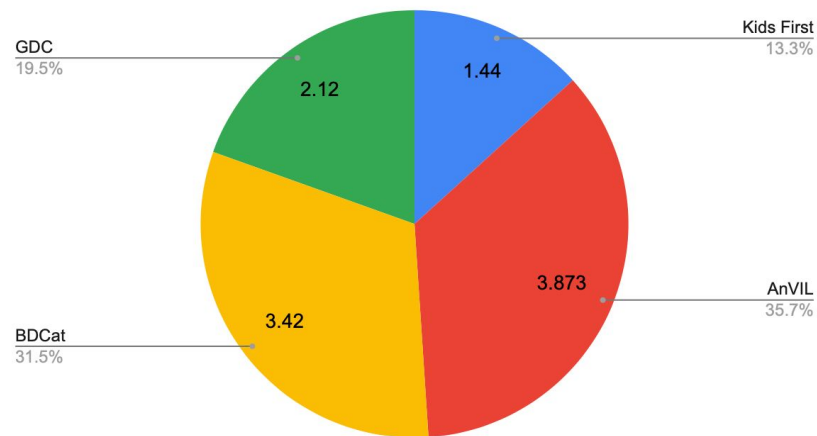
NCPI-wide data ingestion

Data Size (PB)



Beginning of 2021

Data Size (PB)



Now!

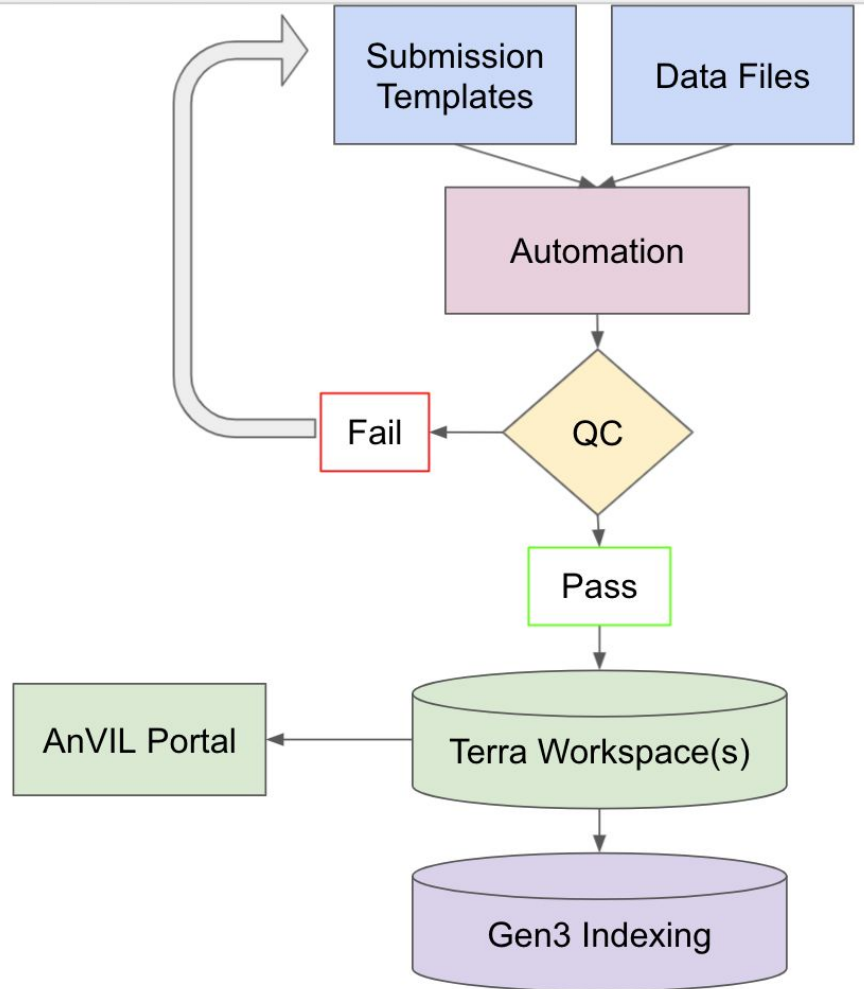
Recent improvements to process

Data Ingestion

- Developed basic scripting to push dataset attributes to workspaces
- Adopted one data model to harmonize data

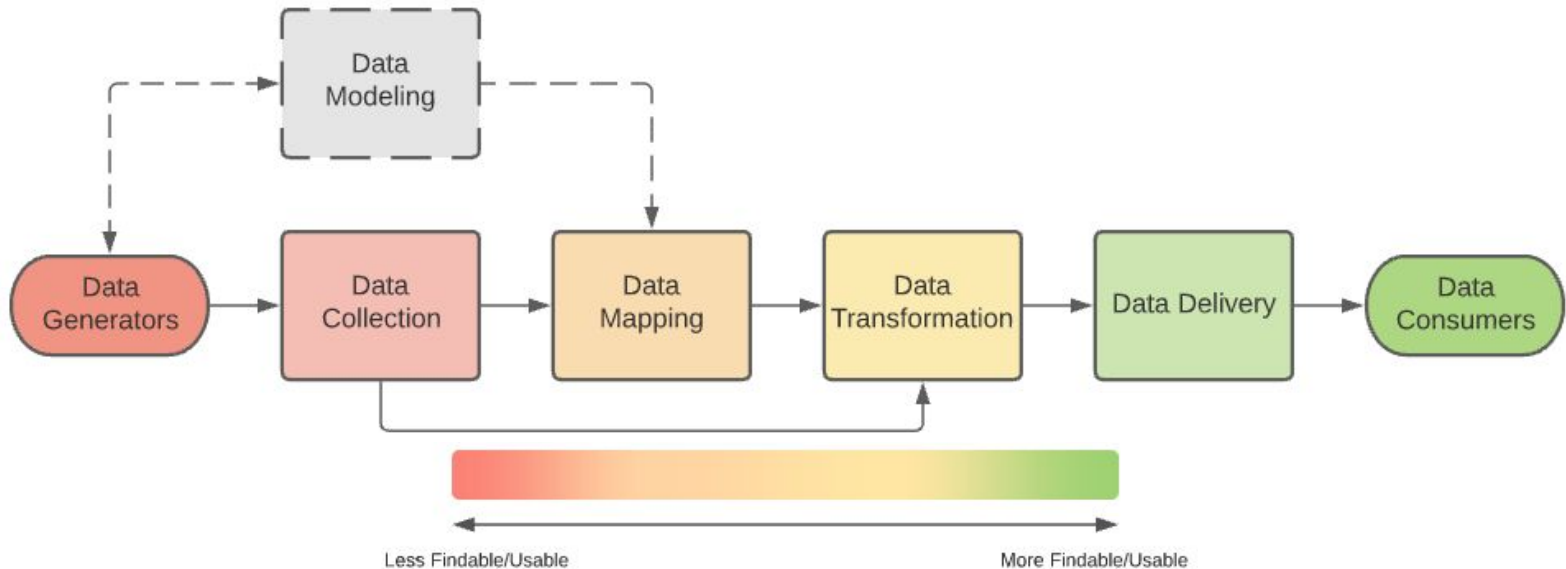
Automation

- Automated workspace creation
 - Data submission kicks off automated workspace creation
- Templated workspaces



Future of AnVIL data ingestion

- Focus: Improve user experience
 - Refine Data Submitter instructions on AnVILproject.org
 - Refine critical path for data submission and ingestion
 -
- Focus: Improve data ingestion turnaround time
 - Additional data submission tooling and automation
 - Building Terra data ingestion pipelines
 - Creating AnVIL Data Model mapping and transformation capabilities



Consortia Engagement



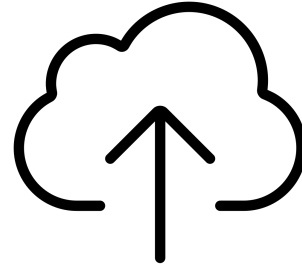
Consortia Engagement Process



Awareness



Recruitment



Submission



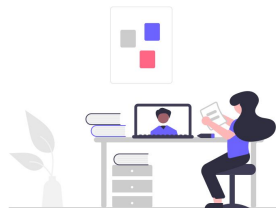
Analysis

Helping Data Managers & Submitters Get Started

PIs



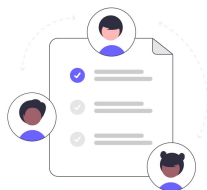
Analysts



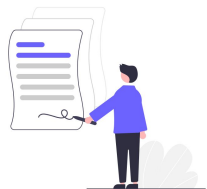
Consortia

Data Managers
Data Submitters

PIs
Analysts



Teachers



NHGRI Analysis Visualization
and Informatics Lab-space



Learn

[Introduction](#) [Data Analysts](#) [Investigators](#) [Data Submitters](#)

▼ Data Submission Guide

[Submission Process Overview](#)

- 1 - Register Study/Obtain Approvals
- 2 - Set Up a Data Model
- 3 - Prepare for Submission
- 4 - Ingest Data
- 5 - QC Data

► Data Submission Resources

Overview

In order to submit data into AnVIL you will need to do the following:

dbGaP

[dbGaP/Obtain required approvals.](#)

Data Model

[Data Model](#)

[Data Model](#)

Submission & QC

[Data Model](#)

[Data Model](#)

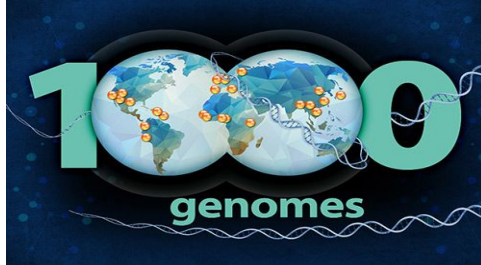
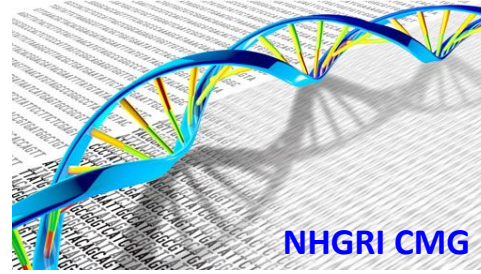
General Data Requirements

Make sure your data conforms to these overall data requirements, or contact the AnVIL data team.

Reference Genome

All submitted genomic data should be based on Human reference genome

The Forefront of Genomics



- [GREGoR](#), [PRIMED](#), TARN, [GAFK](#), [IGVF](#)
- Global FTD-ALS, NIH COVID RECOVER Platform Interoperability (NCRPI), NIA Dementia Long-read Project, NIAID Central Sequencing Program ... all of NHGRI! :)



The screenshot shows a web browser interface for a workspace titled "Telomere-to-Telomere (T2T) Consortium's AnVIL_T2T Workspace". The workspace is powered by Terra and is currently empty, with 0 submissions and 0 access levels. The "ABOUT THE WORKSPACE" section describes the consortium's goal to assemble the first complete reference human genome from the CHM13 hydatidiform mole. The "Currently Available Data" section lists one dataset: "Adds over 80 million base pairs of sequence that can be effectively used for variant calling with long reads". The "WORKSPACE INFORMATION" table shows the creation date as 2/23/2021 and the last updated date as 3/6/2021. The "OWNERS" section lists the workspace administrator as slzarate96@gmail.com. The "TAGS" section is currently empty.

WORKSPACE INFORMATION	
CREATION DATE	2/23/2021
LAST UPDATED	3/6/2021
SUBMISSIONS	0
ACCESS LEVEL	Writer
EST. \$/MONTH	\$2340.52
GOOGLE PROJECT ID	anvil-datasc...



☰ README.md

WDLs for T2T Variants

This directory contains the WDL files used for large-scale short-read small-variant analysis.

Data ingestion

- `wdls/download_aspera.wdl` : Downloads FASTQ files from the European Nucleotide Archive (ENA), given accession numbers

Read alignment

- `wdls/t2t_alignment.wdl` : Given a reference FASTA file, sample name, paired-end FASTQ files, BWA index, and dedup distance (default = 100), performs alignment as described in [Aganezov, Yan, Soto, Kirsche, Zarate, et al. \(2021\)](#)

Variant calling

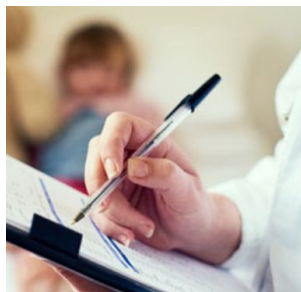
- `wdls/haplotype_calling_chrom.wdl` : Given a reference FASTA (plus corresponding index and dict), a sample CRAM (plus corresponding index), the sample name, the sex of

Clinical Genomics Engagements

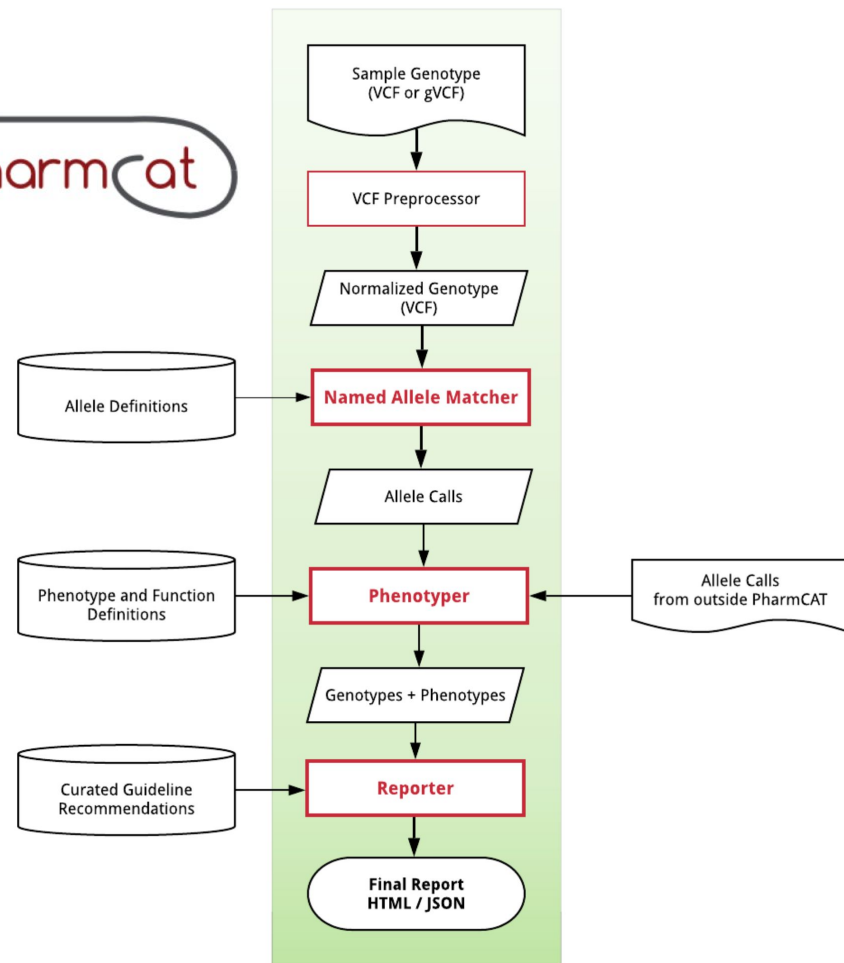
American Heart Association



eMERGE
(and more...)



Social Determinants of Health



Synchronous Events

**GSP // MaGIC
Jamboree**

June 2020

**RCMI //
VADSTI**

April 2021

**AnVIL Office
Hours**

November 2021

**GDSCN
Kickoff**

March 2021

Consortia
Kickoff
Meetings:
PRIMED,
GREGOR,
BIOPLEX,
IGVF, SDOH,
...

**AnVIL Office
Hours Pilot**

September 2021

NHGRI GSP -- MaGIC Jamboree

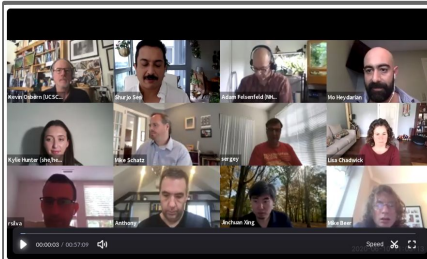


Day 1 (June 10th, 2020)

11:00 AM	NHGRI welcome and introduction (lecture)	
11:05 AM	AnVIL introduction (lecture)	Slides , Video
11:30 AM	Terra - data access and discovery (lecture)	Slides , Video
12:00 PM	AnVIL data catalog and exploration (lecture)	Slides , Video
12:30 PM	Break	
1:30 PM	Linking billing accounts, eRA commons account on Terra (hands-on)	Slides , Video
2:00 PM	Using AnVIL to access, browse, and share data (hand-on)	Slides , Video
2:30 PM	Breakout sessions	
3:00 PM	Closing	

Attendees

Day 1: 106, Day 2: 91



Low cost of training
\$1.35/ participant

Day 2 (June 11th, 2020)

11:00 AM	Data analysis on AnVIL - use cases (lecture)	Slides , Video
11:15 AM	Dockstore and WDL (lecture)	Slides , Video
11:45 AM	Terra - for data analysis (lecture)	Slides , Video
12:00 PM	Concept of Workspaces (hands-on*)	Slides , Video
12:30 PM	Break	
1:30 PM	Batch processing of data with Workflows (hands-on*)	Slides , Video
2:00 PM	Exploratory analysis with Notebooks & R Studio (hands-on*)	Slides , Video
2:25 PM	Jamboree closing	
2:30 PM	Breakout sessions	

*"The **hands on exercises** were really informative and presented smoothly. I thought they were presented in a way that was really easy to follow."*

*"It's really useful platform and have **a lot of resource!** Really appreciate your hard work!!!"*

*"I liked the **interactive slack conversation** with[out] interrupting proceedings"*

*"I am blown away. **AnVIL is amazing.** Congratulations. I think your educational outreach, availability to customer questions, tutorials, etc. will be an important part of making AnVIL becoming widely used."*



Research Centers in
Minority Institutions



Module 7

Tools for Applied Data
Science Using Cloud-
Based Platform

Thursday, April 22, 2021 &
Friday, April 23, 2021
11:00 AM – 1:30 PM EDT

[Download Calendar
Notice»](#)

[read more»](#)



Module 8

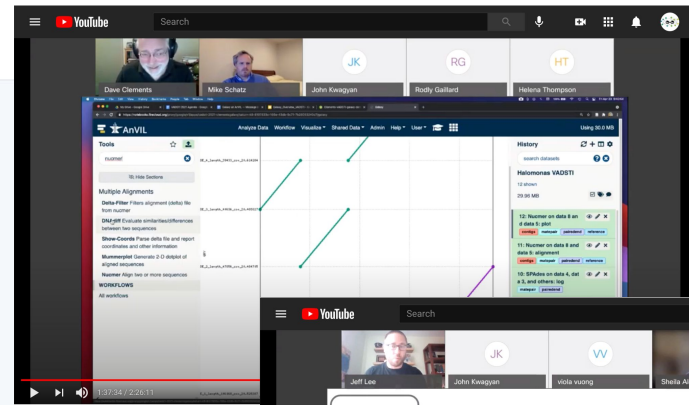
Current Research Topics
Seminar: Biomedical,
Clinical & Genomic
Application

8.1 – Thursday, March 11, 2021
8.2 – Thursday, March 25, 2021
8.3 – Thursday, April 8, 2021
8.4 – Thursday, April 29, 2021

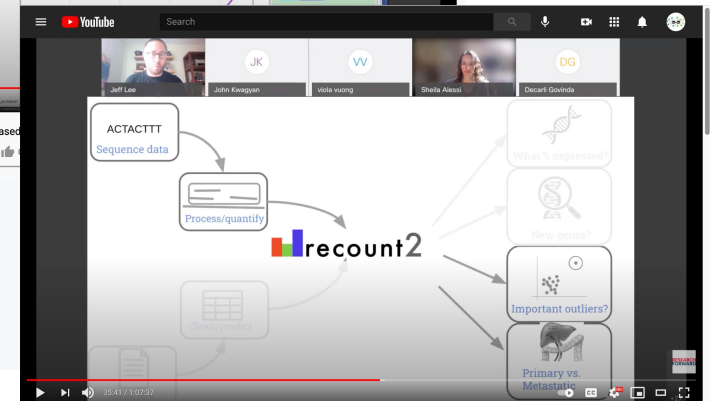
11:00 AM – 1:00 PM EDT

[Download 8.2 Calendar
Notice»](#)

[Download 8.3 Calendar
Notice»](#)



Tools for Applied Data Science Using Cloud-Based



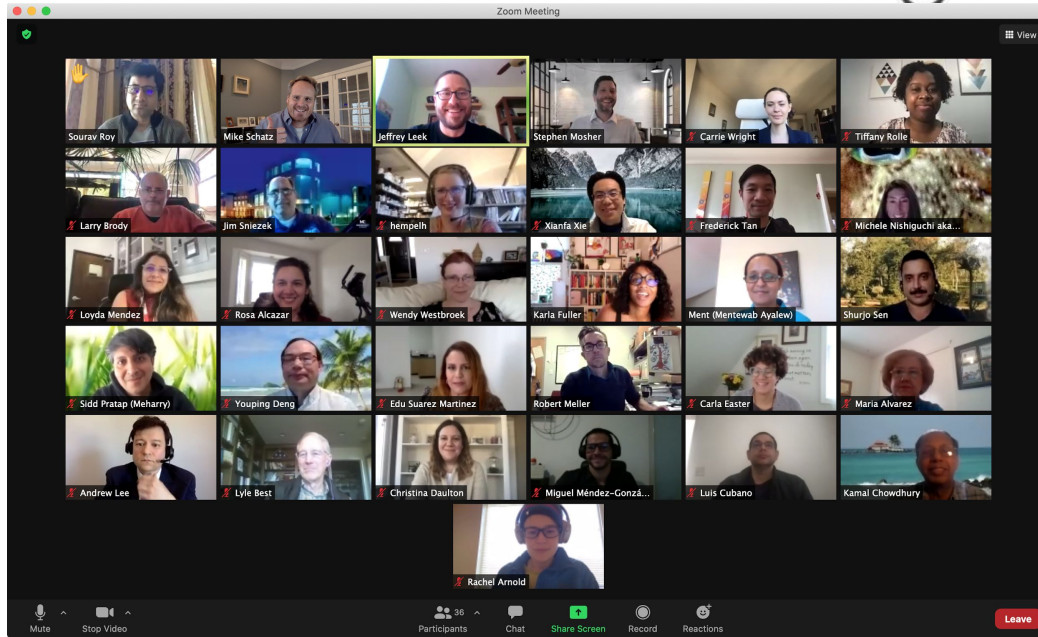
Research Seminal Presentation 8.2

13 views · Mar 25, 2021

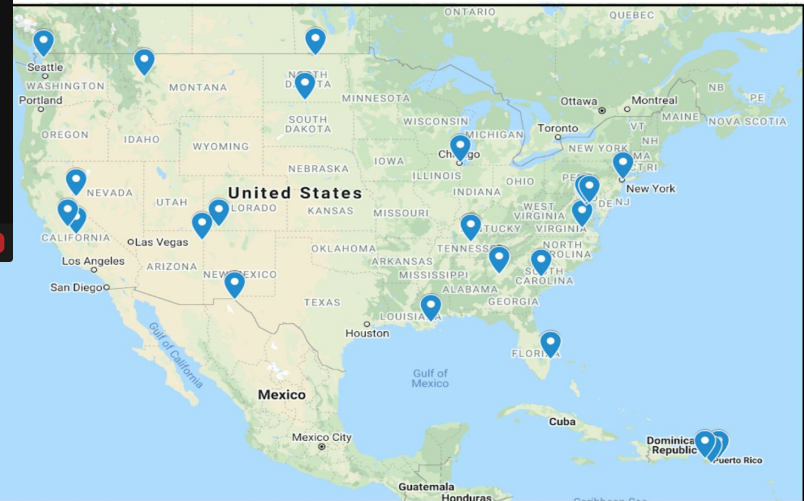
1 0 SHARE SAVE ...

Neural Networks Models -
Module 5, Day 2
Howard ICMI
4 views · 1 month ago

GDSCN -- Diverse Faculty and Institutions



- Network
- White Paper
- Curricula




Office Hours Access for Consortia

Technical Working Group

Chairs: Michael Schatz (JHU)
Brian O'Connor (Broad)

Data Access Working Group

Chairs: Stacey Donnelly (Broad)
Carolyn Hutter (NHGRI)



Outreach Working Group

Chairs: Jeffrey Leek (JHU)
Frederick Tan (Carnegie)



Data Processing Working Group

Chairs: Eric Banks (Broad),
Ira Hall (WashU/Yale)

Portal Working Group

Chairs: Michael Schatz (JHU)
Benedict Paten (UCSC)



Phenotype Working Group

Chairs: David Crosslin (eMERGE - UW)
Robert Carroll (VUMC)



Data Ingestion Committee

Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

AHA/AnVIL Working Group

Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

Vision for Future Engagement

