Data Submission and Consortia Engagement

29OCT2021

Brian O'Connor and Frederick Tan

AnVIL

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		Analysis tools	
Moderators: Dr. Adam Resnick (Children's Hospital of Philadelphia) and Ms. Valentina Di Francesco (NHGRI)		*Moderators: Dr. Marylyn Ritchie (University of Pennsylvania) and Dr. Ken L. Wiley, Jr. (NHGRI)*	
12:35-12:40	Moderator introductions	12:35-12:40	Moderator introductions
12:40-12:55	AnVIL presentation: *Dr. Brian O'Connor (Broad) and Dr. Frederick Tan (Carnegie)*	12:40-12:55	AnVIL presentation: *Dr. Vincent Carey (HMS) and Dr. Ira Hall (Yale)*
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

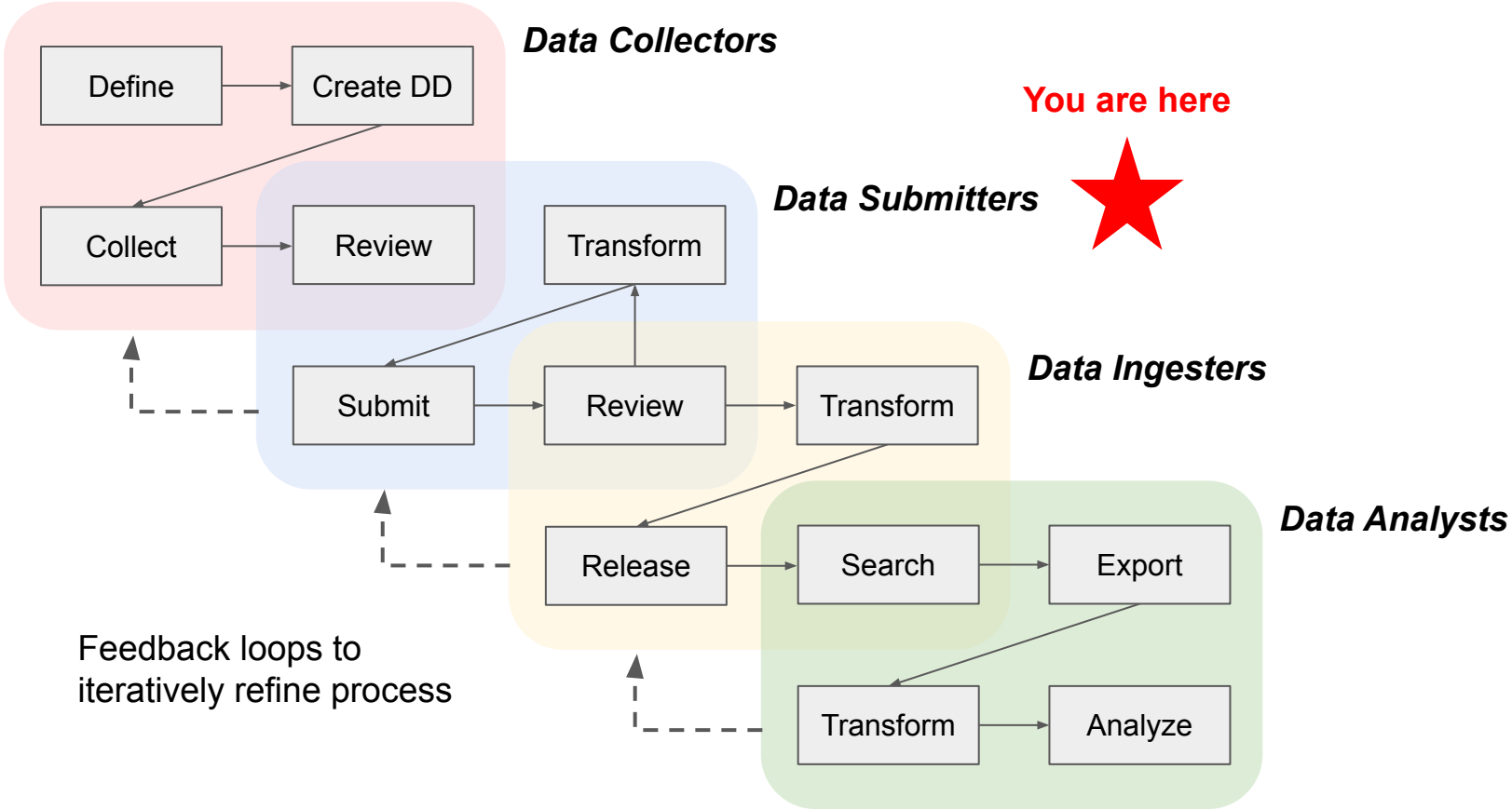
Data Submission

Submitting data - dbGaP versus AnVIL

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

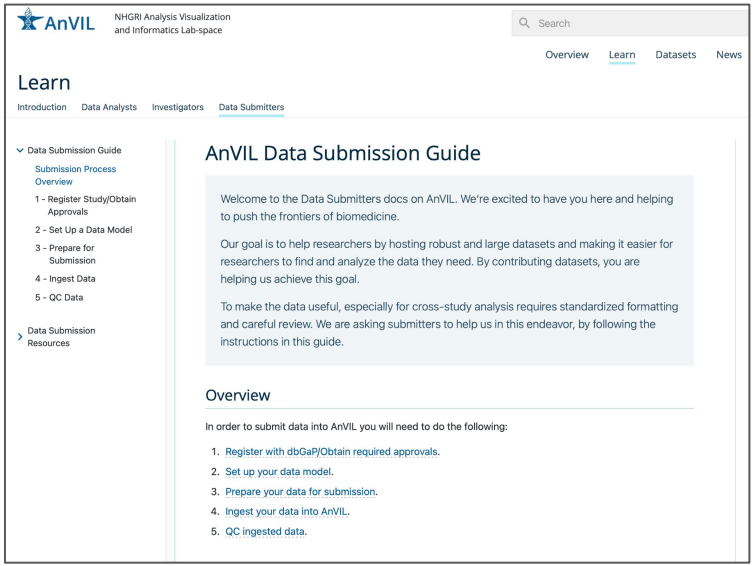
- **AnVIL**
  - Used by data submitters (registered on AnVIL)
  - Phenotypic data in integrated tables (CSV or TSV in workspace bucket)
  - Genomic data files in workspace buckets

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

1 Year AnVIL Planned Data Ingestion

- 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
GTEx (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**

anvilproject.org/data

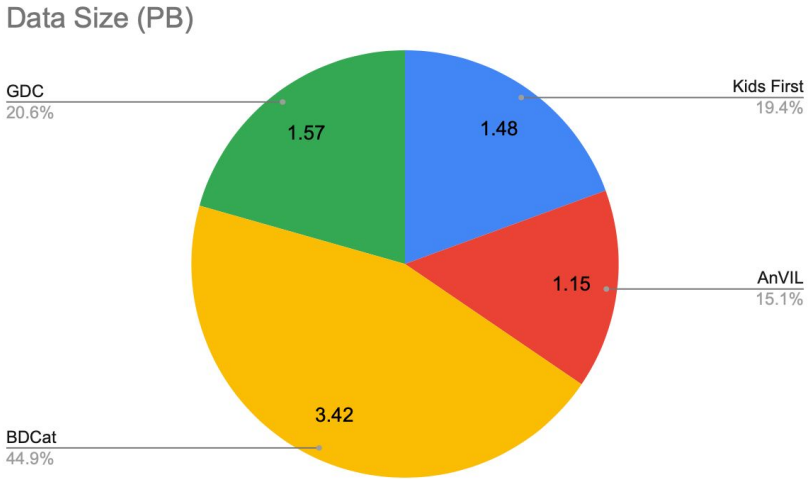
Current data ingestion process

- 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

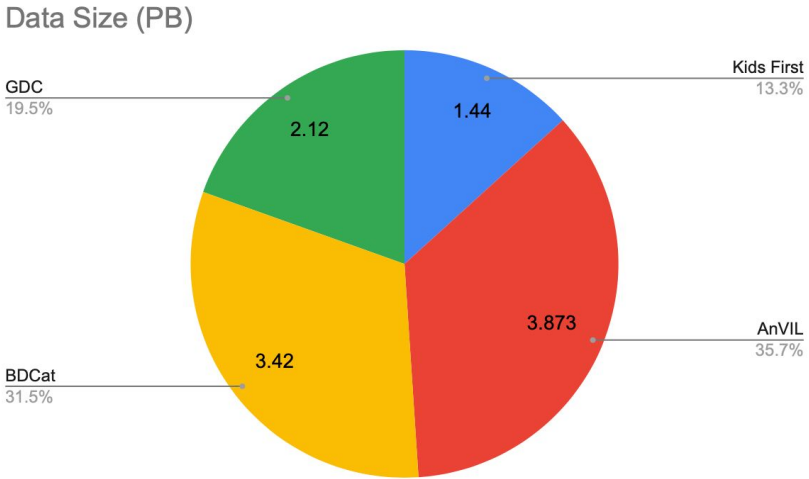
Current data ingestion process

- 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



Beginning of 2021



Now!

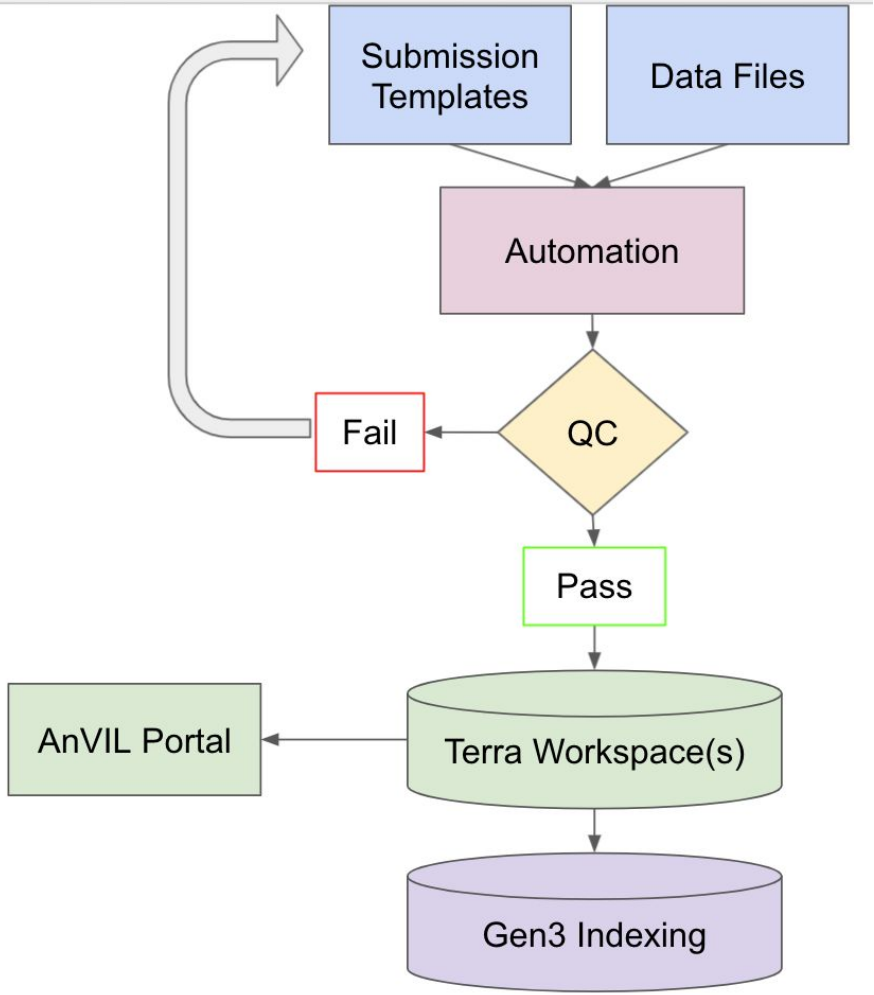
Recent improvements to process

Data Ingestion

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

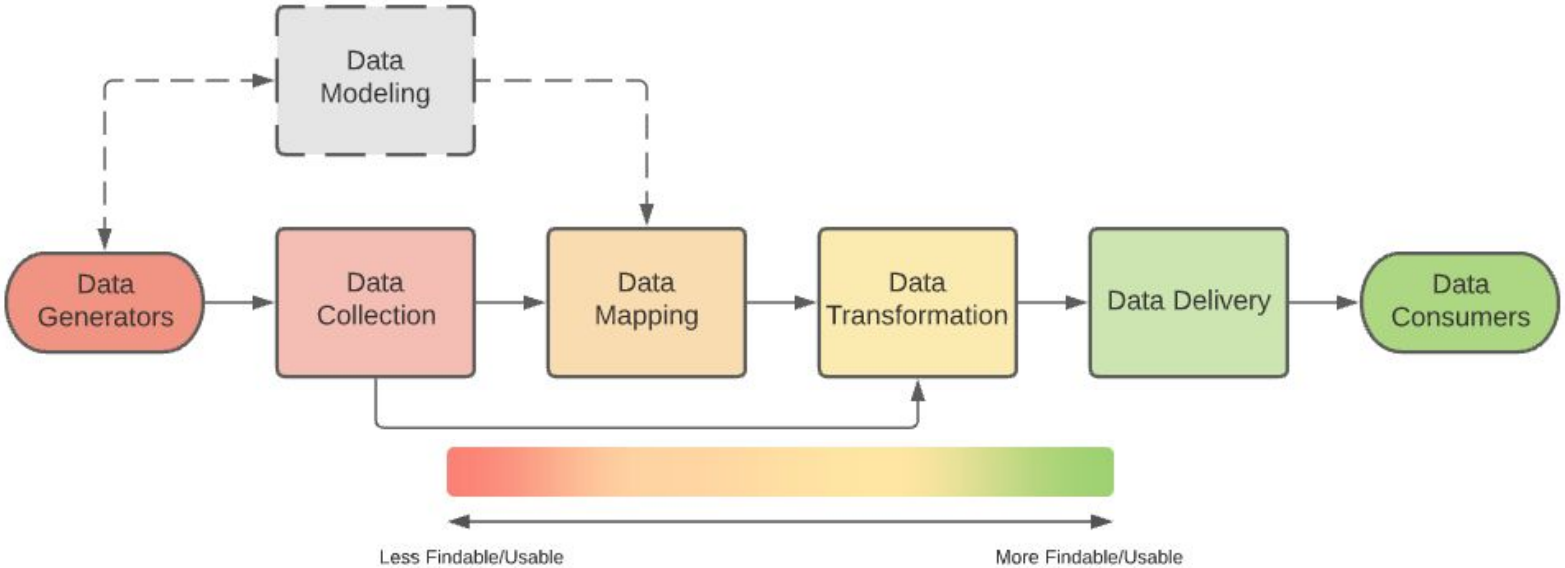
Automation

- 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

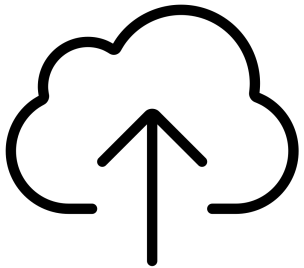


Consortia Engagement

Consortia Engagement Process

Awareness

Recruitment



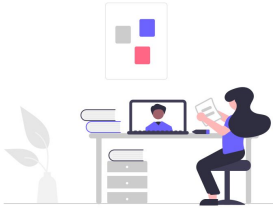
Submission

Analysis

Helping Data Managers & Submitters Get Started

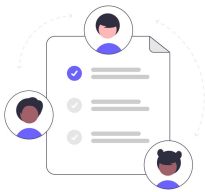
PIs

Analysts

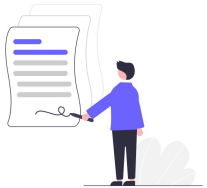


Consortia

Data Managers
Data Submitters
PIs
Analysts



Teachers



AnVIL NHGRI Analysis Visualization and Informatics Lab-space

Search

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

dbGaP

Data Model

Submission & QC

Overview

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

- GaP/Obtain required approvals.
- a model.
- ta for submission.
- into AnVIL.
- ta.

General Data Requirements

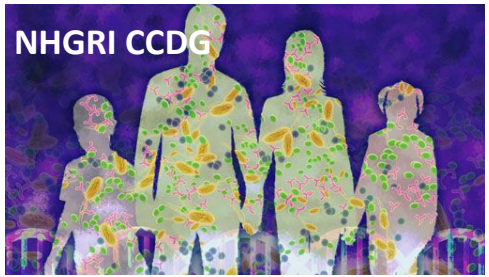
Make sure your data conforms to these overall data requirements, or contact the AnVIL data team.

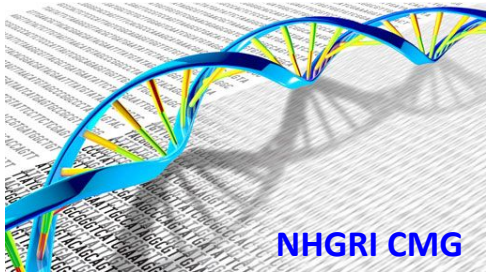
Reference Genome

All submitted genomic data should be based on Human reference genome

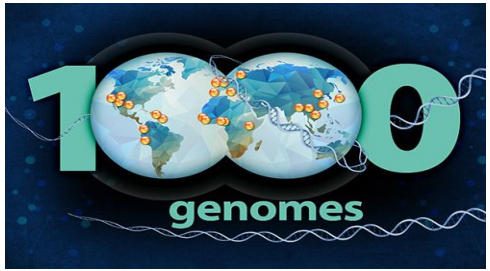
jhudatascience.org/AnVIL_Book_Getting_Started

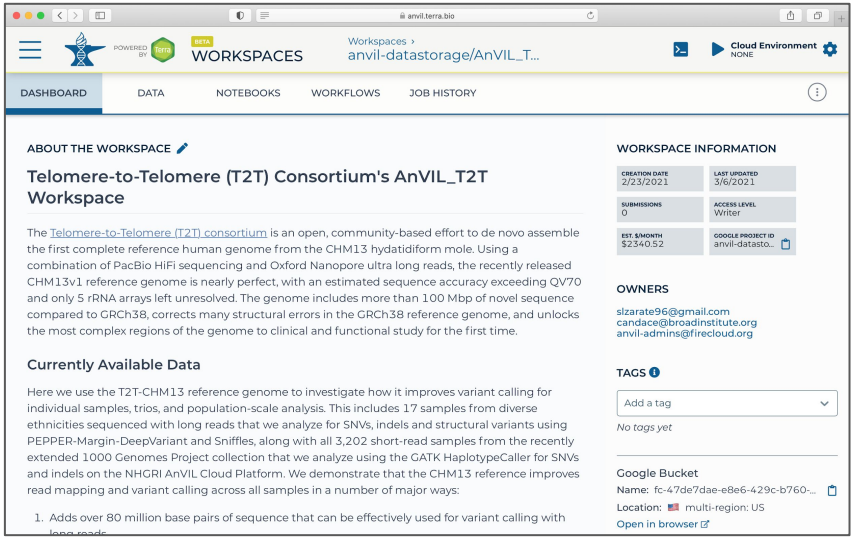
The Forefront of Genomics





TELOMERE-TO-TELOMERE CONSORTIUM



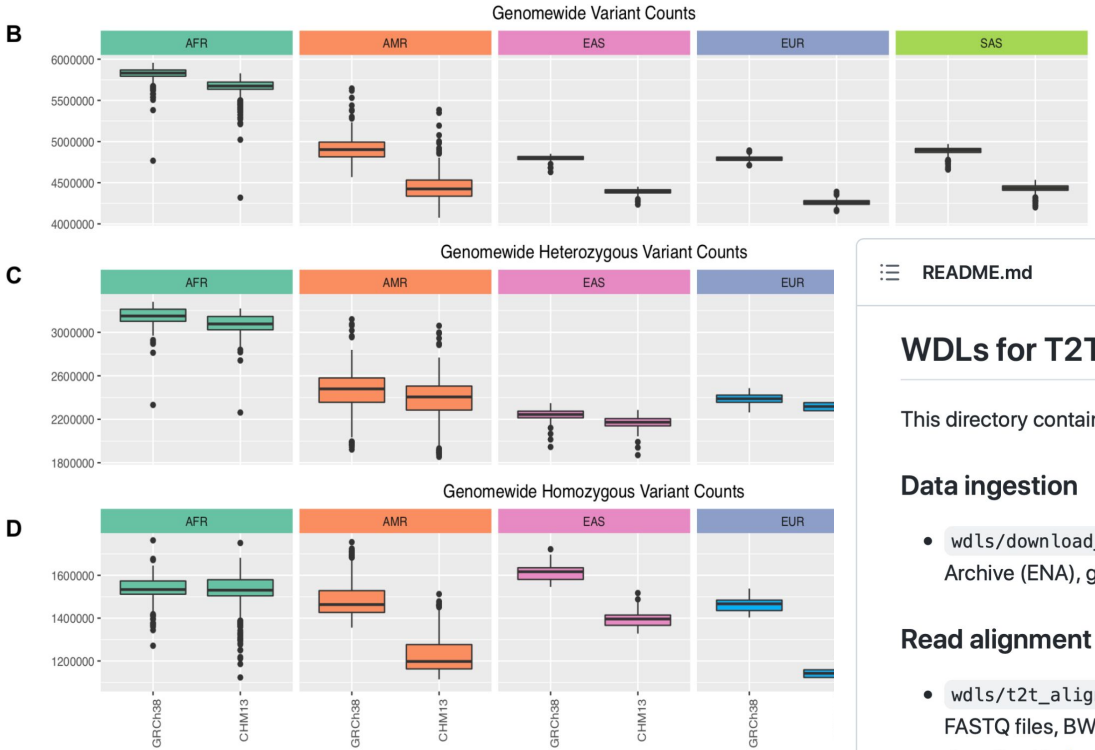


- GREGoR, PRIMED, TARN, GAFK, IGVF
- Global FTD-ALS, NIH COVID RECOVER Platform Interoperability (NCRPI), NIA Dementia Long-read Project, NIAID Central Sequencing Program … all of NHGRI! :)

anvilproject.org/data

anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T

Re-analyzing Diversity in 3,202 Samples



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

Aganezov, *et al.*, 2021 bioRxiv // github.com/schatzlab/t2t-variants

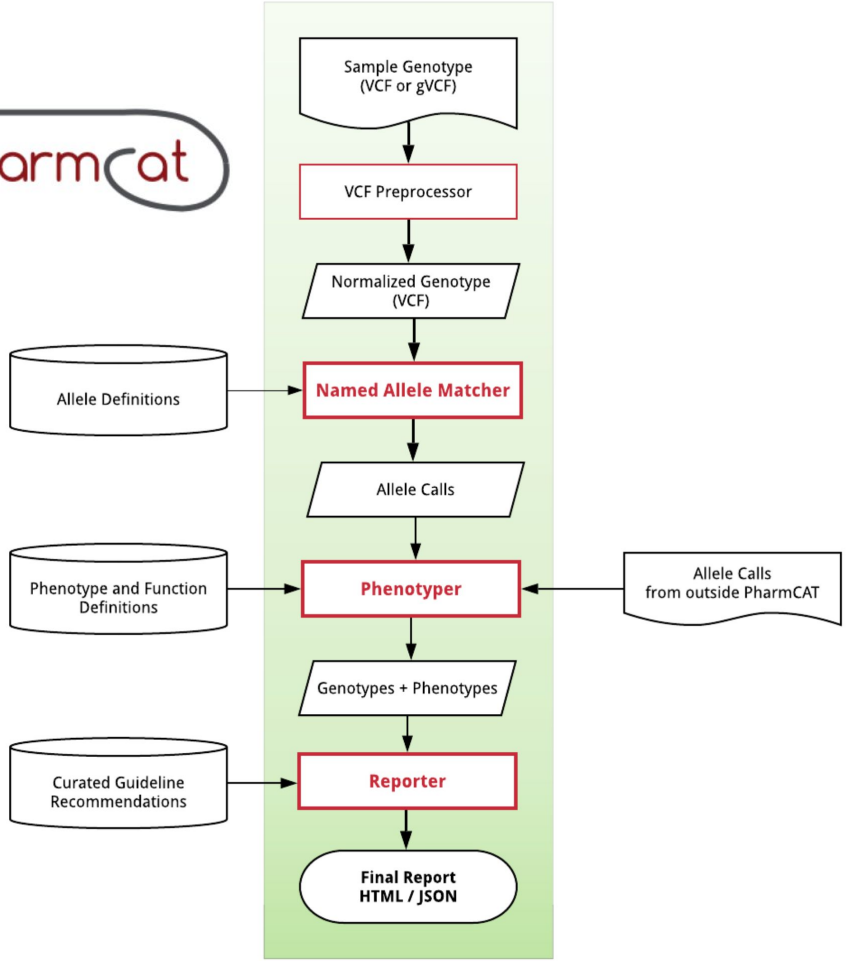
Powered by AnVIL

Clinical Genomics Engagements

American Heart Association

eMERGE
(and more…)

Social Determinants of Health



pharmcat.org

Synchronous Events

GSP // MaGIC Jamboree

June 2020

GDSCN Kickoff

March 2021

RCMI // VADSTI

April 2021

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

AnVIL Office Hours Pilot

September 2021

AnVIL Office Hours

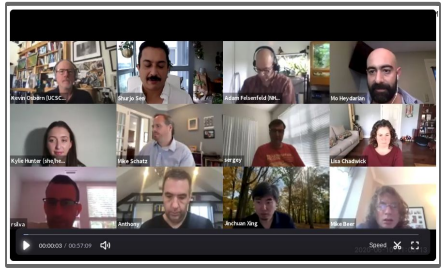
November 2021

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."

anvilproject.org/events/magic2020

RCMI -- VADSTI

HOWARD UNIVERSITY | Research Centers in Minority Institutions | RESEARCH FORWARD

About | Agenda | Curriculum | Register

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»

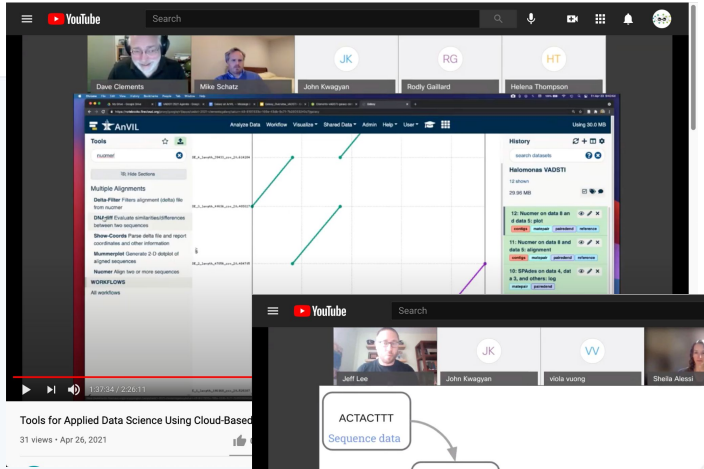
read more»

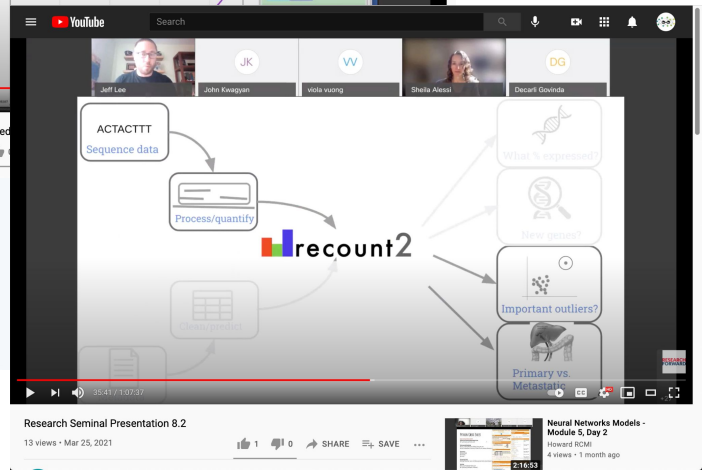
Module 8
Current Research Topics Seminar: Biomedical, Clinical & Genomic Application

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

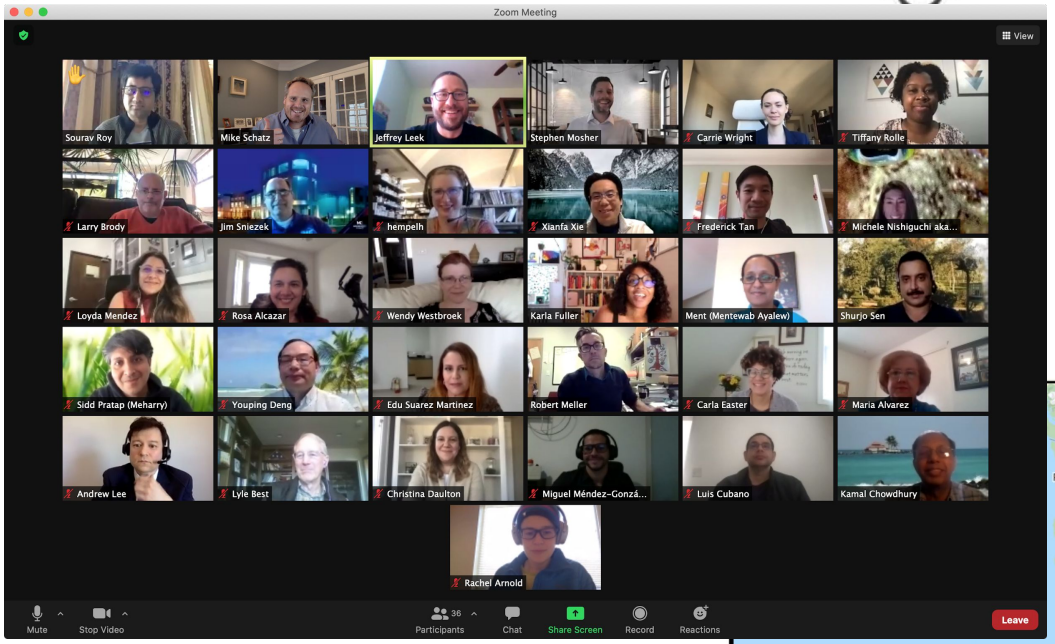
Download 8.2 Calendar Notice»



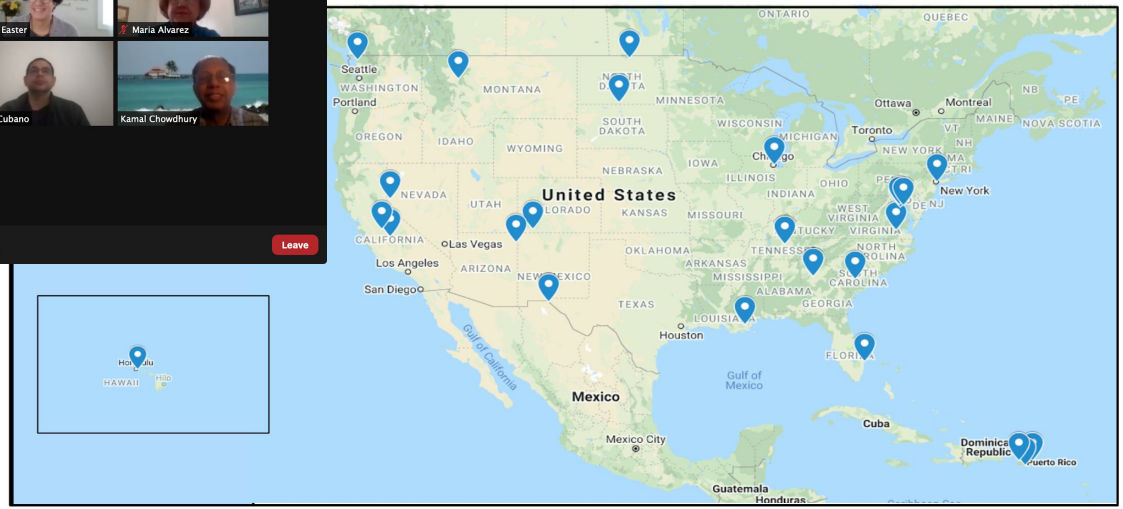


anvilproject.org/events/vadsti-2021

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)

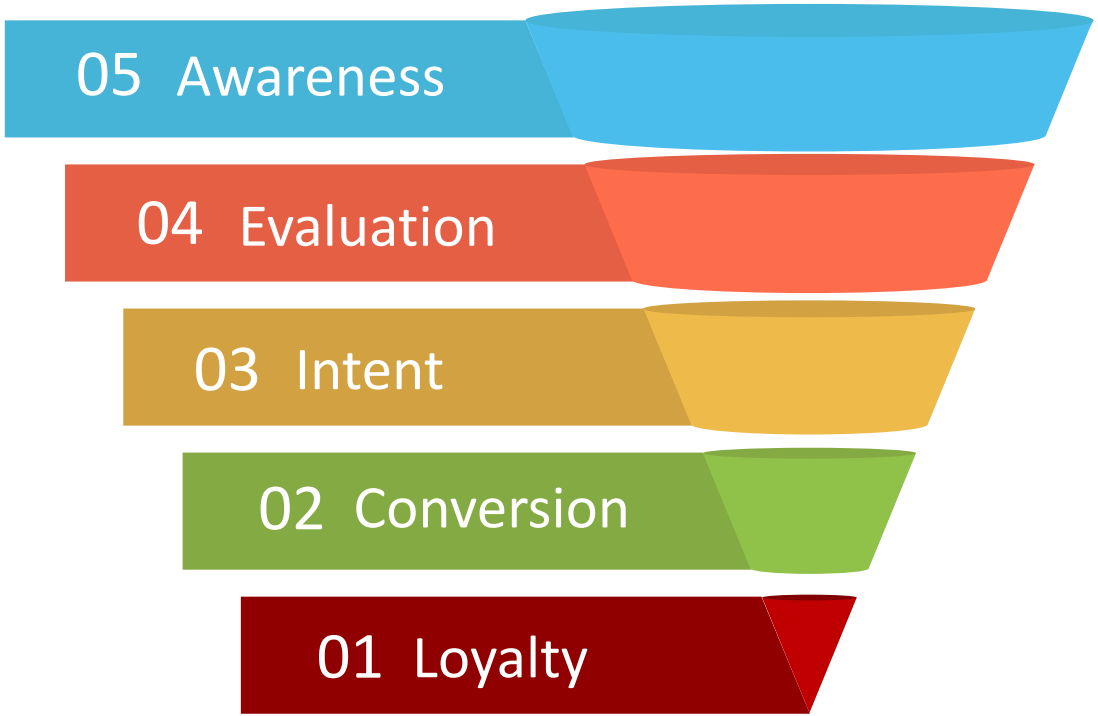
Data Ingestion Committee

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



- MOOCs, Media
- Conference, etc.

- MaGIC Jamboree
- Getting Started

- AC2 Credits
- DeepPilots

- Community Discourse Forums

Adapted from poweredtemplate.com