Data Submission and Consortia Engagement

29 OCT 2021

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

<table>
<thead>
<tr>
<th>Time</th>
<th>Data submission and consortia engagement</th>
<th>Analysis tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:35-1:50</td>
<td>Moderator introductions</td>
<td>Moderator introductions</td>
</tr>
<tr>
<td>12:35-12:40</td>
<td>AnVIL presentation:</td>
<td>AnVIL presentation:</td>
</tr>
<tr>
<td>12:40-12:55</td>
<td>Dr. Brian O’Connor (Broad) and</td>
<td>Dr. Vincent Carey (HMS)</td>
</tr>
<tr>
<td></td>
<td>Dr. Frederick Tan (Carnegie)</td>
<td>and Dr. Ira Hall (Yale)</td>
</tr>
<tr>
<td>12:55-1:40</td>
<td>Discussion</td>
<td>Discussion</td>
</tr>
<tr>
<td>1:40-1:50</td>
<td>Prepare breakout report</td>
<td>Prepare breakout report</td>
</tr>
</tbody>
</table>

**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
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 Collectors**
- Define
- Create DD
- Collect
- Review
- Transform

**Data Submitters**
- Submit
- Review
- Transform
- Release
- Search
- Export

**Data Ingesters**
- Data Collectors
- Data Submitters
- Data Ingesters
- Data Analysts

Feedback loops to iteratively refine process

You are here
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

<table>
<thead>
<tr>
<th>Primary Column</th>
<th>Q4</th>
<th>Q3</th>
<th>Q2</th>
<th>Q1</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oct</td>
<td>Nov</td>
<td>Dec</td>
<td>Jan</td>
<td>Feb</td>
</tr>
<tr>
<td>1000G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPRC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTEx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>open access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bi-sulfite sequencing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recount3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome in a Bottle</td>
<td>Genome in a Bottle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCDG</td>
<td>CCDG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIA</td>
<td>NIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WGSPD &amp; ConvergentNeuro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dementia Long-Read</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identification of risk factors for ALS &amp; FTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMG</td>
<td>CMG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAPK</td>
<td>GAPK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TARN</td>
<td>TARN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical WGS</td>
<td>Clinical WGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREGor (MRGC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMDG</td>
<td>PMDG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eMERGE II &amp; III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eMERGE PRS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSER</td>
<td>CSER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGNITE II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIME</td>
<td>PRIME</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COVID19 Prospective genomic study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENCODE</td>
<td>ENCODE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic testing standard of care vs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clinical WGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIMH - National Institute of Mental Health</td>
<td>NIMH - National Institute of Mental Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

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

- GDC: 20.6%
- Kids First: 19.4%
- AnVIL: 15.1%
- BDCat: 44.9%

**Now!**

- GDC: 19.5%
- Kids First: 13.3%
- AnVIL: 35.7%
- BDCat: 31.5%
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

Learn

Overview

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

- **dbGaP**
- **Data Model**
- **Submission & QC**

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

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! :)

anvilproject.org/data

anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T
Re-analyzing Diversity in 3,202 Samples

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, Kirsch, Zarate, et al. (2021)

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

Aganezov, et al., 2021 bioRxiv // github.com/schatzlab/t2t-variants
Clinical Genomics Engagements

American Heart Association

eMERGE (and more…)

Social Determinants of Health

pharmcat.org
GSP // MaGIC Jamboree
June 2020

GDSCN Kickoff
March 2021

RCMI // VADSTI
April 2021

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

AnVIL Office Hours
November 2021

AnVIL Office Hours Pilot
September 2021
"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[not] 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."

**Attendees**
Day 1: 106, Day 2: 91

<table>
<thead>
<tr>
<th>Day 1 (June 10th, 2020)</th>
<th>Day 2 (June 11th, 2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00 AM NHGRI welcome and introduction (lecture)</td>
<td>11:00 AM Data analysis on AnVIL - use cases (lecture)</td>
</tr>
<tr>
<td>11:05 AM AnVIL introduction (lecture)</td>
<td>Slides, Video</td>
</tr>
<tr>
<td>11:30 AM Terra - data access and discovery (lecture)</td>
<td>Slides, Video</td>
</tr>
<tr>
<td>12:00 PM AnVIL data catalog and exploration (lecture)</td>
<td>Slides, Video</td>
</tr>
<tr>
<td>12:30 PM Break</td>
<td>11:15 AM Dockstore and WDL (lecture)</td>
</tr>
<tr>
<td>1:30 PM Linking billing accounts, eRA commons account on Terra (hands-on)</td>
<td>Slides, Video</td>
</tr>
<tr>
<td>2:00 PM Using AnVIL to access, browse, and share data (hand-on)</td>
<td>Slides, Video</td>
</tr>
<tr>
<td>2:30 PM Breakout sessions</td>
<td>11:45 AM Terra - for data analysis (lecture)</td>
</tr>
<tr>
<td>3:00 PM Closing</td>
<td>12:00 PM Concept of Workspaces (hands-on*)</td>
</tr>
<tr>
<td></td>
<td>12:30 PM Break</td>
</tr>
<tr>
<td></td>
<td>1:30 PM Batch processing of data with Workflows (hands-on*)</td>
</tr>
<tr>
<td></td>
<td>2:00 PM Exploratory analysis with Notebooks &amp; R Studio (hands-on*)</td>
</tr>
<tr>
<td></td>
<td>2:25 PM Jamboree closing</td>
</tr>
<tr>
<td></td>
<td>2:30 PM Breakout sessions</td>
</tr>
</tbody>
</table>

**Registrants**
- 53 CMG
- 12 CCDG
- 10 NHGRI
- 10 GSP - AC
- 4 GSP - CC
- 24 unidentified

**Attendees**

<table>
<thead>
<tr>
<th>NHGRI GSP -- MaGIC Jamboree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendees</td>
</tr>
<tr>
<td>Day 1: 106, Day 2: 91</td>
</tr>
</tbody>
</table>

**Low cost of training**

$1.35/ participant
RCMI -- VADSTI

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

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)

### Outreach Working Group
- **Chairs:**
  - Jeffrey Leek (JHU)
  - Frederick Tan (Carnegie)

### Portal Working Group
- **Chairs:**
  - Michael Schatz (JHU)
  - Benedict Paten (UCSC)

### Data Ingestion Committee
- **Members:**
  - Michael Schatz (JHU)
  - Anthony Philippakis (Broad) et al

### Data Access Working Group
- **Chairs:**
  - Stacey Donnelly (Broad)
  - Carolyn Hutter (NHGRI)

### Data Processing Working Group
- **Chairs:**
  - Eric Banks (Broad),
  - Ira Hall (WashU/Yale)

### Phenotype Working Group
- **Chairs:**
  - David Crosslin (eMERGE - UW)
  - Robert Carroll (VUMC)

### AHA/AnVIL Working Group
- **Members:**
  - Michael Schatz (JHU)
  - Anthony Philippakis (Broad) et al
Vision for Future Engagement

01 Loyalty

02 Conversion

03 Intent

04 Evaluation

05 Awareness

- MOOCs, Media Conference, etc.
- MaGIC Jamboree Getting Started
- AC2 Credits DeepPilots
- Community Discourse Forums

Forums

AC2 Credits

DeepPilots

MaGIC Jamboree

Getting Started

MOOCs, Media Conference, etc.

Community Discourse Forums

Adapted from poweredtemplate.com