

# Infrastructure

29OCT2021



Benedict Paten and Jeremy Goecks



# Overview



- Data Infrastructure
  - Portal, Data Dashboard
  - Security
  - DUOS
- Compute Infrastructure
  - Terra, multicloud
  - AnVIL APIs
  - Component interactions
- Interoperability Infrastructure
  - NCPI
  - RAS, DRS, FHIR
- Future directions

2:30-3:45

Session 2: Breakout rooms

## Infrastructure

Moderators: Ms. Karen L. Davis (RTI International) and Dr. Carolyn M. Hutter (NHGRI)

2:30-2:35

**Moderator introduction**

2:35-2:50

**AnVIL presentation:**

*Dr. Jeremy Goecks (OHSU) and Dr. Benedict Paten (UCSC)*

2:50-3:35

Discussion

3:35-3:45

Prepare breakout report

## Outreach and training

Moderators: Dr. Siddharth Pratap (Meharry Medical College) and Mr. Christopher Wellington (NHGRI)

2:30-2:35

**Moderator introduction**

2:35-2:50

**AnVIL presentation:**

*Dr. Jeffrey Leek (JHU) and Ms. Tiffany Miller (Broad)*

2:50-3:35

Discussion

3:35-3:45

Prepare breakout report

## Breakout room: Infrastructure

*Moderators: Ms. Karen Davis and Dr. Carolyn Hutter*

Mr. Samuel (Sandy) Aronson

Dr. Vivien Bonazzi

Dr. Brandi Davis-Dusenbery

Dr. Richard Gibbs

Dr. George Hripcsak

Dr. Eimear Kenny

Dr. Lucila Ohno-Machado

Dr. Shannon McWeeney

Mr. Luke Rasmussen

# AnVIL Portal: Front-end of the AnVIL Project

The screenshot displays the AnVIL Portal website. At the top left is the AnVIL logo and the text "NHGRI Analysis Visualization and Informatics Lab-space". A search bar is located at the top right. Below the search bar is a navigation menu with links for Overview, Learn, Datasets, News, Events, Team, FAQ, and Help. The main content area features a large heading "Migrate Your Genomic Research to the Cloud" and a sub-heading "Secure, cost-effective genomic analysis at scale." Below this are two buttons: "Get Started" and "Learn More". To the right is a featured article titled "A complete reference genome improves analysis of human genetic variation" with a "Paper" and "Workspace" link. Below the article is a carousel of dots. The bottom section is a grid of tool integrations, each with a logo, name, description, and "Learn More" link.

**AnVIL** NHGRI Analysis Visualization and Informatics Lab-space

Search

Overview Learn Datasets News Events Team FAQ Help

## Migrate Your Genomic Research to the Cloud

Secure, cost-effective genomic analysis at scale.

[Get Started](#) [Learn More](#)

**A complete reference genome improves analysis of human genetic variation**

[Paper](#) [Workspace](#)

Using the AnVIL, researchers find the T2T-CHM13 reference genome universally improves the analysis of human genetic...

**Terra**  
Collaborate in Terra, AnVIL's secure, scalable, cloud compute environment.  
[Launch](#) [Learn More](#)

**Gen3**  
Manage, harmonize, and share large datasets.  
[Launch](#) [Learn More](#)

**Dockstore**  
Create and share Docker-based workflows.  
[Launch](#) [Learn More](#)

**NCPI**  
Interoperate with other NIH data commons.  
[Learn More](#)

**Bioconductor**  
Analyze genomic data in the R statistical language.  
[Learn More](#)

**Galaxy**  
Run batch analysis workflows and interactive visualizations.  
[Learn More](#)

**Jupyter**  
Run interactive analysis with python or R.  
[Learn More](#)

**Seqr**  
Identify disease-causing variants.  
[Launch](#) [Learn More](#)

<https://anvilproject.org>

# Nearly 4pb of data ingested by AnVIL



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
<b>Totals</b>	<b>254</b>	<b>314,038</b>	<b>291,301</b>	<b>3,873.01</b>

Much more to come!

# Egress free GTEx via Gen3/Cleversafe

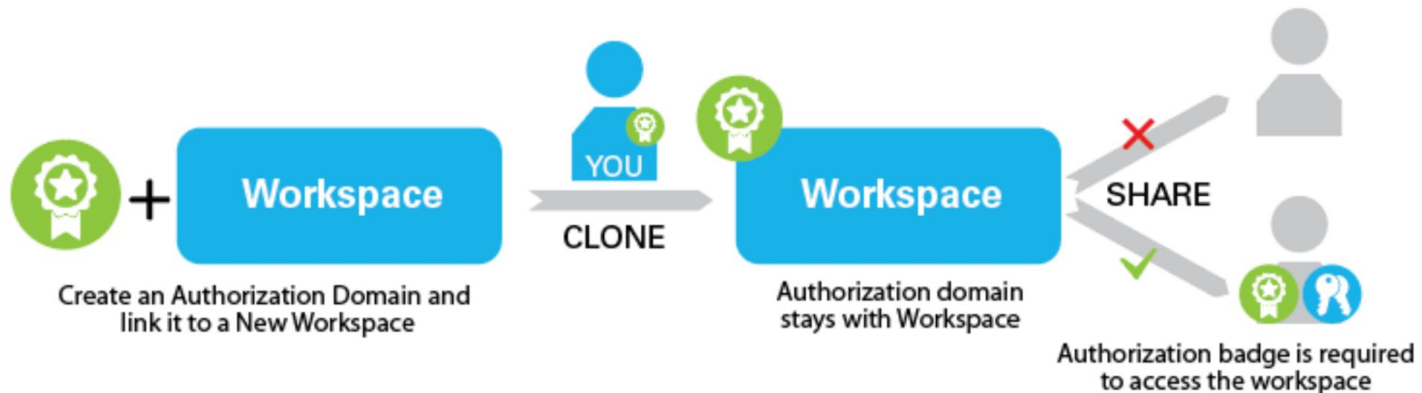
The screenshot shows the Genomeweb website with a navigation bar and a search box. The main headline is "AnVIL Platform Makes Popular NHGRI GTEx Database Free to Download" dated Dec 23, 2020. The article text states: "CHICAGO – The National Human Genome Research Institute's (NHGRI) popular Genotype-Tissue Expression (GTEx) dataset is now available for free download through a cloud-based platform, potentially saving researchers as much as \$14,000 in access and storage costs per download. The move, according to its backers, promises to democratize use of the largest existing compendium of human gene expression and corresponding trait loci, bringing even the smallest institutions into the fold." It also mentions that NHGRI issued version 8 of GTEx, the first "free" release on the Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) platform.

The screenshot shows the AnVIL website with a navigation bar and a search box. The main headline is "GTEx v8 - Free Egress Instructions". The "Overview" section states: "The Genotype-Tissue Expression (GTEx) Program is a widely used data resource and tissue bank to study the relationship between genetic variants (inherited changes in DNA sequence) and gene expression (how genes are turned on and off) in multiple human tissues and across individuals. Previously, large genetic studies identified variants that are associated with human diseases. However, it is less clear how these variants affect gene expression and thereby contribute to human diseases." It also provides information on how to request tissue samples and access the data through the GTEx portal (https://gtexportal.org/).

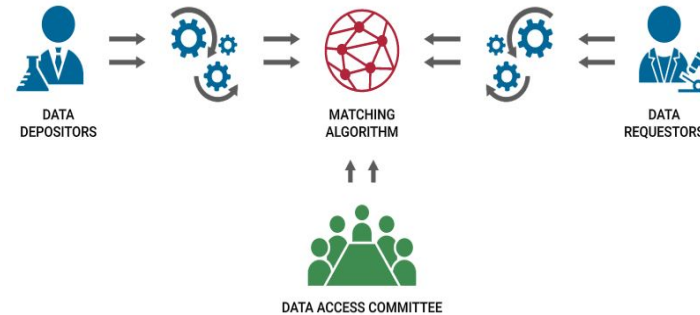
<https://www.genomeweb.com/informatics/anvil-platform-makes-popular-nhgri-gtex-database-free-download#.YEri9JNKigw>  
<https://anvilproject.org/learn/reference/gtex-v8-free-egress-instructions>

# Data access and security on AnVIL

- *Strong emphasis on security*
  - Terra and tools on AnVIL are FedRAMP compliant
  - Operate in a FISMA-Moderate environment and are compliant with NIST-800-53
- *Data access*
  - Workspaces are broken out by study registration and consent group mapping
  - AnVIL workspaces containing controlled data have an authorization domain to limit access to only those with the appropriate permissions to work said data
  - Authorization domains follow a workspace when copied/cloned



# DUOS: Data Use Oversight System

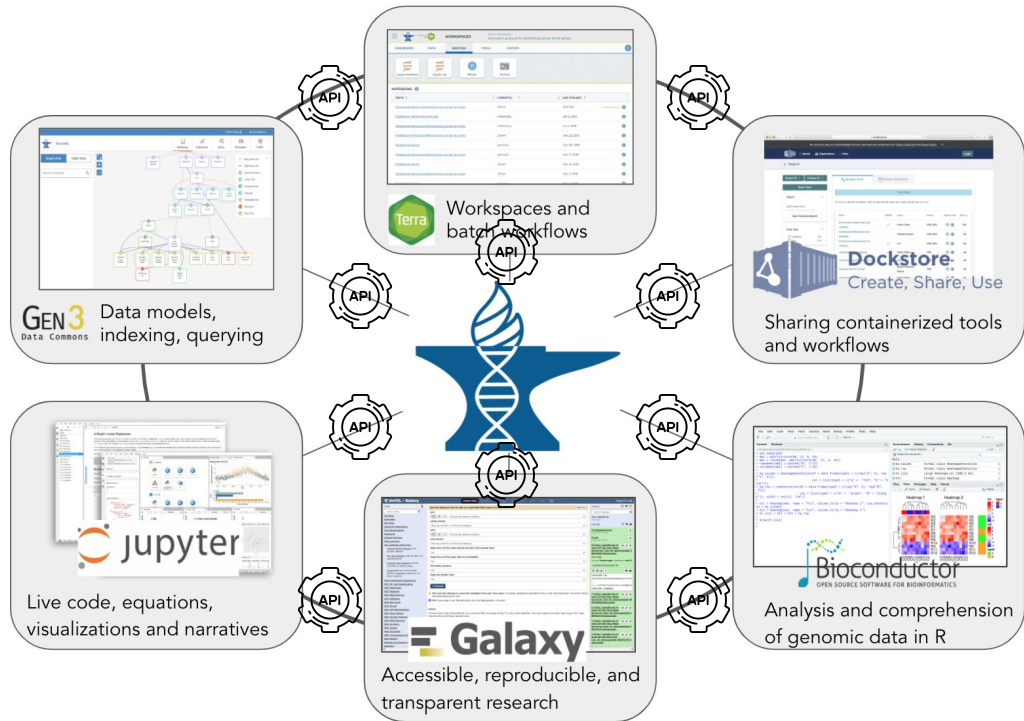


## What is DUOS?

- Interfaces to transform data use restrictions and data access requests to machine-readable code (ADA-M & Consent Codes)
- A matching algorithm that checks if data access requests are compatible with data use restrictions
- Interfaces for the Data Access Committee to adjudicate whether structuring and matching has been done appropriately

Initial testing with NHGRI, NHLBI, NIAID, and JAAMH has been very positive  
(95% concordant with manual review)

# APIs



An *Application Programming Interface (API)* provides a standard way to access underlying system functionality

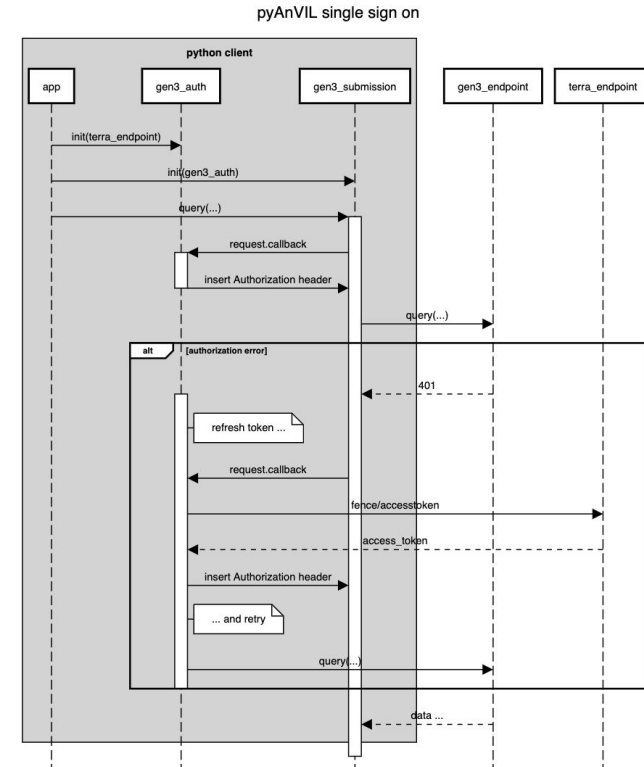
## Mission

- Connect AnVIL software components and data
- Enable external applications and developers to use AnVIL software and data



# pyAnVIL: a Python Library for AnVIL

- A Python library for using AnVIL system components
  - Single sign-on (SSO)
  - Query AnVIL components for available data using various APIs
- Works both in and outside of the AnVIL
- Applications include AnVIL data dashboard, Galaxy data browser, and FHIR prototypes



# Galaxy <-> Terra in Production!

The image displays three overlapping screenshots of web interfaces related to Galaxy and Terra in production.

**Top Screenshot: Terra Cloud Environment**  
Title: Cloud environment  
Text: Environment will consist of an application and cloud compute.  
Section: Environment Settings  
List: Galaxy version 20.09

**Middle Screenshot: AnVIL Welcome Page**  
Title: Welcome to Galaxy on AnVIL  
Text: This project is a collaboration between the Galaxy project, the Broad Institute and NHGRI. Galaxy is a hands-on analysis platform providing a graphical user interface for computing and data analysis. Galaxy is a hands-on analysis platform providing a graphical user interface for computing and data analysis.  
Sidebar: ANVIL Tools, Get Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED.

**Bottom Screenshot: FastQC Report**  
Title: FastQC Report  
Date: Thu 22 Oct 2020 19:17:11  
File: frag180\_1.fq  
Summary: Basic Statistics  
Warning: Per base sequence quality  
Table: Basic Statistics

Measure	Value
Filename	frag180_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35198
Sequences flagged as poor quality	0
Sequence length	100
tQC	54

Per base sequence quality: Quality scores across all bases (Sanger / Illumina 1.9 encoding)

<https://anvilproject.org/learn/getting-started/getting-started-with-galaxy>

# Connecting Galaxy Workbench with Terra Workspaces



**Global Alliance**  
for Genomics & Health  
Collaborate. Innovate. Accelerate.

WORKSPACES | Data | Galaxy Running | Cloud Environment None

DASHBOARD | DATA | NOTEBOOKS | WORKFLOWS | JOB HISTORY

TABLES

- BigQuery\_table (2)
- DRS\_URI (1)
- cohort (1)
- donor (1)
- donor\_set (1)
- file (6)
- partic\_100x8 (100)**
- participant (1)
- participant\_900gs (979)
- sequencing2 (2)

REFERENCE DATA

- hg38
- b37Human

OTHER DATA

- Workspace Data

partic_100x8	age	ase_chrX_raw_counts	ase_counts
<input type="checkbox"/> GTEX-1117F	66	<a href="#">GTEX-1117F.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1117F.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-111CU	57	<a href="#">GTEX-111CU.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-111CU.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-111FC	61	<a href="#">GTEX-111FC.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-111FC.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-111VG	63	<a href="#">GTEX-111VG.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-111VG.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-111YS	62	<a href="#">GTEX-111YS.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-111YS.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1122O	64	<a href="#">GTEX-1122O.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1122O.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1128S	66	<a href="#">GTEX-1128S.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1128S.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1131C	66	<a href="#">GTEX-1131C.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1131C.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1131C	53	<a href="#">GTEX-1131C.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1131C.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-117XS	64	<a href="#">GTEX-117XS.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-117XS.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-117YW	58	<a href="#">GTEX-117YW.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-117YW.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-117YX	55	<a href="#">GTEX-117YX.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-117YX.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1192W	68	<a href="#">GTEX-1192W.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1192W.v8.ase_table.tsv.gz</a>
<input type="checkbox"/> GTEX-1192X	55	<a href="#">GTEX-1192X.readcounts.chrX.txt.gz</a>	<a href="#">GTEX-1192X.v8.ase_table.tsv.gz</a>



Download from

Type to Search

Regular

Label	Time
<input type="checkbox"/> GTEX-1117F-0003-SM-58Q7G.bai	-
<input type="checkbox"/> GTEX-1117F-0003-SM-58Q7G.bam	-
<input type="checkbox"/> GTEX-1117F-0003-SM-6WBT7.crai	-
<input type="checkbox"/> GTEX-1117F-0003-SM-6WBT7.cram	-
<input type="checkbox"/> GTEX-1117F.readcounts.chrX.txt.gz	-
<input type="checkbox"/> GTEX-1117F.v8.ase_table.tsv.gz	-
<input type="checkbox"/> GTEX-1117F.v8.readcounts.chrX.txt.gz	-
<input type="checkbox"/> GTEX-1117F.v8.wasp_corrected.ase_table.tsv.gz	-
<input type="checkbox"/> GTEX-111CU-0003-SM-58Q95.bai	-
<input type="checkbox"/> GTEX-111CU-0003-SM-58Q95.bam	-
<input type="checkbox"/> GTEX-111CU-0003-SM-6WBUD.crai	-

Back | Cancel | OK

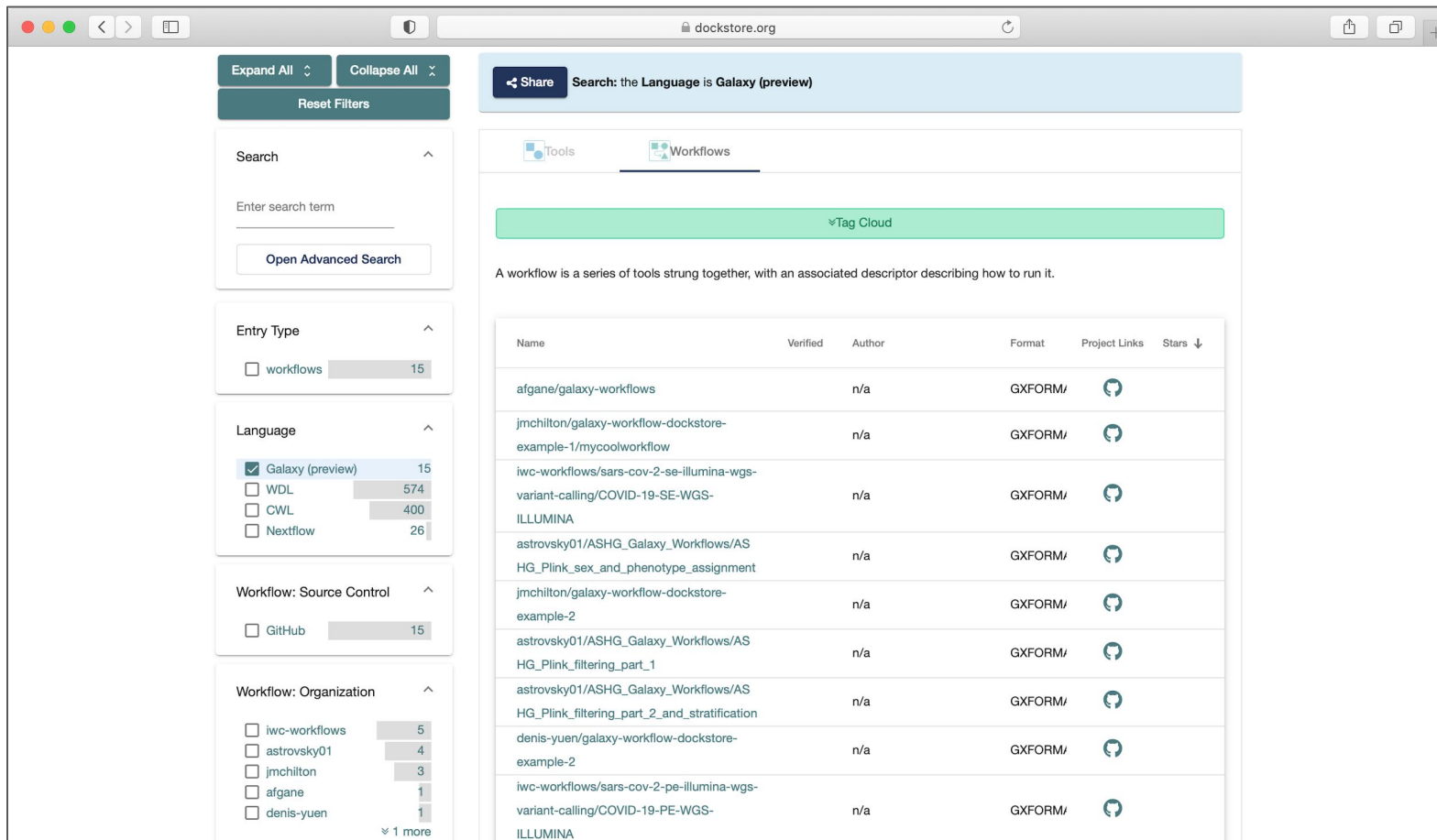
Choose local files | Choose remote files | Paste/Fetch data | Start | Pause | Reset | Close

Convert Formats | Lift-Over | MISCELLANEOUS TOOLS | Virology | COMMON GENOMICS TOOLS

If this is your first time using Galaxy, visit our training site for an introduction to the system and existing workflows in various areas of analysis. Further, if you need help, please click the help section on the masthead and visit one of our several help resources.

Galaxy Training Material

# Galaxy <-> Dockstore in Production!



The screenshot shows the Dockstore search results page for Galaxy workflows. The search term is "the Language is Galaxy (preview)". The results are filtered by Language (Galaxy (preview) is selected) and Workflow: Source Control (GitHub). The results table lists various workflows with their names, verified status, authors, formats, project links, and star counts.

Name	Verified	Author	Format	Project Links	Stars ↓
afgane/galaxy-workflows		n/a	GXF <sup>o</sup> RM/		
jmchilton/galaxy-workflow-dockstore-example-1/mycoolworkflow		n/a	GXF <sup>o</sup> RM/		
iwc-workflows/sars-cov-2-se-illumina-wgs-variant-calling/COVID-19-SE-WGS-ILLUMINA		n/a	GXF <sup>o</sup> RM/		
astrovsky01/ASHG_Galaxy_Workflows/AS_HG_Plink_sex_and_phenotype_assignment		n/a	GXF <sup>o</sup> RM/		
jmchilton/galaxy-workflow-dockstore-example-2		n/a	GXF <sup>o</sup> RM/		
astrovsky01/ASHG_Galaxy_Workflows/AS_HG_Plink_filtering_part_1		n/a	GXF <sup>o</sup> RM/		
astrovsky01/ASHG_Galaxy_Workflows/AS_HG_Plink_filtering_part_2_and_stratification		n/a	GXF <sup>o</sup> RM/		
denis-yuen/galaxy-workflow-dockstore-example-2		n/a	GXF <sup>o</sup> RM/		
iwc-workflows/sars-cov-2-pe-illumina-wgs-variant-calling/COVID-19-PE-WGS-ILLUMINA		n/a	GXF <sup>o</sup> RM/		

<https://dockstore.org/search?descriptorType=gxformat2>

# NIH Cloud Platforms Interoperability (NCPI)

*The NIH Cloud Platform Interoperability Effort (NCPI) will establish and implement guidelines and technical standards to empower end-user analyses across participating platforms and facilitate the realization of a trans-NIH, federated data ecosystem.*

- [NCPI Website](#)
- Meetings
  - October 3rd 2019 - Initial Meeting
  - April 16th 2020 - 2nd NCPI Meeting
  - October 30th 2020 - 3rd NCPI Meeting
  - May 3rd 2021 - 4th NCPI Meeting
  - October 5th 2021 - 5th NCPI Meeting
- Working Groups
  - Community Governance Working Group
  - Coordination Working Group
  - FHIR Working Group
  - Outreach and Training Working Group
  - Systems Interoperation Working Group



<https://anvilproject.org/ncpi>

# Technologies supporting interoperability

## RAS: Researcher Auth Service

Goal: Unified identity/authentication

RAS is an effort by the NIH's Center for Information Technology (CIT) to provide a common mechanism by which researchers can establish their identity and access data they are authorized to use.



## DRS: Data Repository Service

Goal: Unified data access across storage infrastructures

[Data Repository Service API](#), are a standardized set of cloud data access methods. The primary functionality is to map a logical ID to a means for physically retrieving the data represented by the `drs://URI` scheme.



## FHIR: Fast Healthcare Interoperability Resources

Goal: data harmonization and an API for exchange of electronic medical records.

FHIR facilitates interoperation between health care systems, to make it easy to provide health care information on a wide variety of devices.



# RAS and DRS in Production!



The AnVIL homepage features a DNA double helix graphic on the left. The main heading is "NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL)". Below this is the sub-heading "EXPLORE, ANALYZE, AND SHARE DATA". A paragraph of text describes the site's purpose: "This website supports the management, analysis and sharing of human disease data research community and aims to advance basic understanding of the genetic basis of disease and accelerate discovery and development of therapies, diagnostic tests, and other tools for diseases like cancer." There are two login buttons: "NIH Login" and "Login from Google". At the bottom, it says "If you have any questions about access or the registration process, please contact support@datacommons.io." The footer includes "Dictionary v2.10", "Submission v2021.02", and "Portal v2021.02".

Login with RAS

The NIH Researcher Auth Service (RAS) Sign In page has a header with the NIH logo and "National Institutes of Health Turning Discovery Into Health". The main heading is "NIH Researcher Auth Service (RAS) Sign In". It contains a form with fields for "Username" (containing "anvil\_user") and "Password" (masked with dots). There is a "Forgot Password?" link and a checkbox for "View consent options upon login". A blue "Sign in" button is at the bottom. A light blue banner below the button says "Smart Card Holder? Sign in with PIV Card." and a link for "Trouble signing in?" is at the bottom.

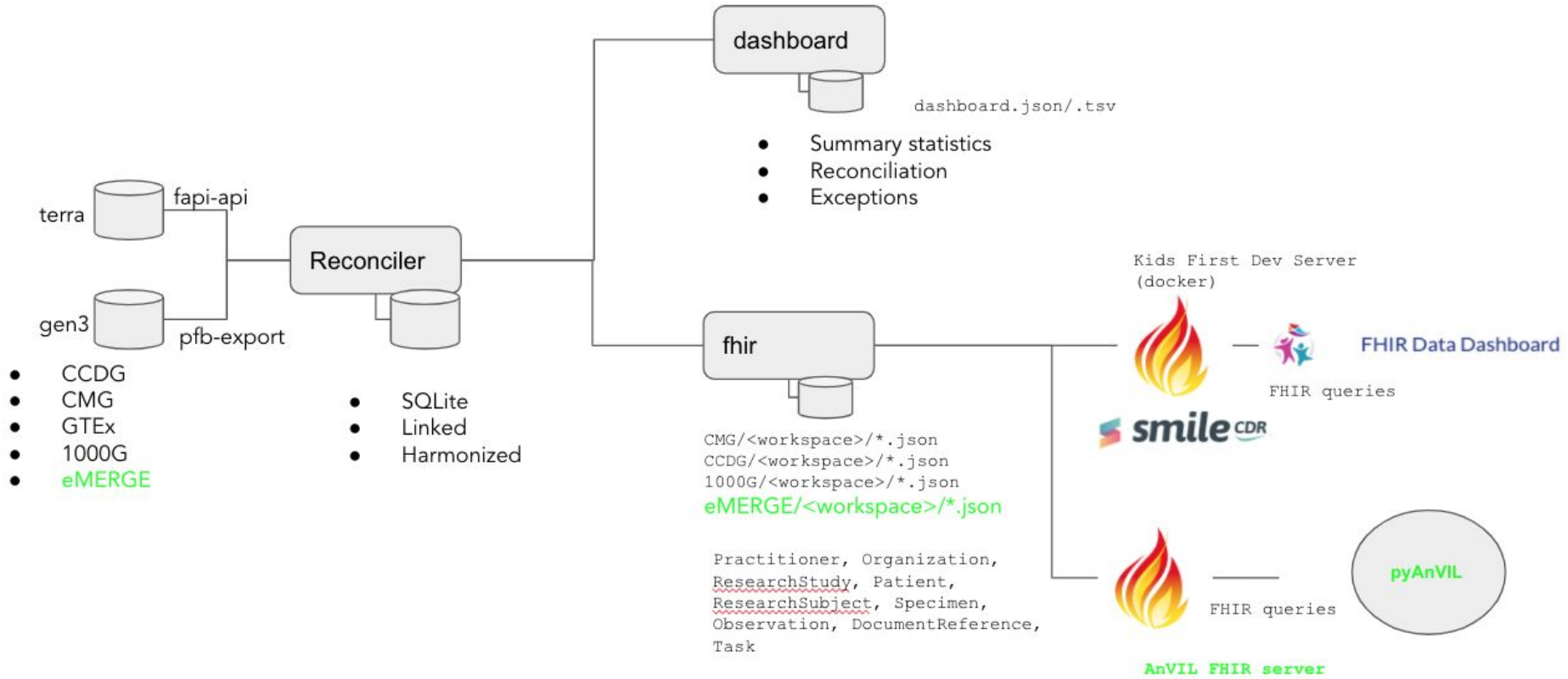
Cohort Search

The AnVIL Data Explorer interface shows a sidebar with "Data Access" options: "Data with Access" (selected), "Data without Access", and "All Data". Below are "Filters" for "Sequencing" (Projects, Subject, Sample) and "Collaborator". A table shows counts for "Projects" (2) and "Subjects" (5,706). A "Sex" chart shows 2,542 Females (44.5%) and 2,466 Males (43.2%). An "Ancestry" chart shows 698 Black or African American (12.2%) and 5,706 White (99.8%). A table below lists "Project ID" and "Ancestry".

Export to Terra (DRS)

The Terra Data Explorer interface shows a "WORKSPACES" header with "Data (read only)". Below are tabs for "DASHBOARD", "DATA", "NOTEBOOKS", "WORKFLOWS", and "JOB HISTORY". A table lists data items with columns for "TABLES", "read\_group", "read\_group\_set", "sample", and "sample\_set". The table includes rows for "read\_group" (24 items), "read\_group\_set" (1 item), "sample" (66 items), and "sample\_set" (1 item). Below the table are sections for "REFERENCE DATA" (hg38) and "OTHER DATA" (Workspace Data, Files). The footer shows "1 - 24 of 24" and "Items per page: 25".

# Initial AnVIL <-> Kids First FHIR Success!







# Future directions

---



- **More robust AnVIL APIs**
  - Provide a unified, stable API endpoint for the AnVIL and its components with OpenAPI documentation
  - Extend API wrapper libraries in Python and R
  - Incorporate additional community APIs and features from GA4GH and FHIR
- **Data organization and additional outputs**
  - Automate dataset ingestion to processed results
  - Increase number and kinds of processed results available, including aggregated results
  - Catalogues of curated datasets with genotype + phenotype
- **Analysis support**
  - Make it easier to navigate between AnVIL applications and NCPI platforms + move data/results and workflows between applications, platforms, & clouds
  - For machine learning: model zoos/integration with external zoos with ready-to-use model
  - Improve clinical data ingestion in applications by increasing use of FHIR