Workshop Report

Future Directions of the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

Workshop date: October 29, 2021

Workshop Background

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space ('AnVIL') is a secure, cloud-based environment where researchers can store, share, and analyze unrestricted- and controlled-access genomic datasets and associated phenotypic data or metadata, particularly those generated by NHGRI consortia and initiatives. Since 2018, NHGRI has been funding and managing AnVIL through two cooperative agreements awarded to groups lead at the Broad Institute and Johns Hopkins University.

Given that AnVIL had recently completed its third year of funding, NHGRI convened a to inform future directions of the AnVIL program as it continues to mature. The goal of this workshop was to identify the gaps, challenges, and opportunities related to NHGRI's investments in AnVIL's cloud-based infrastructure, tools and services. Participants from the genomics research community (including basic genomics, clinical genomics and genomic data science) helped NHGRI identify the activities that are needed by the AnVIl resource to expand, diversify, and support genomics researchers and the AnVIL user community.

The workshop agenda included four breakout rooms, focused on the following topics: (1) Data submission and consortia engagement, (2) Analysis tools, (3) Infrastructure, and (4) Outreach and training. In each session NHGRI was particularly interested in hearing feedback on the following cross-cutting topics: (a) how cloud-based platforms can better serve the needs of genomics researchers; (b) what tools and services would better support clinical genomics researchers; and (c) how to improve interoperability with other NIH cloud-based resources in a federated genomic data ecosystem.

Detailed information about the workshop's goals, agenda, and meeting participants can be found in the workshop booklet in **Appendix 1** and on the workshop webpage.

The workshop discussion in each breakout room was conducted using a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis framework. Specifically, participants were asked to discuss the following: Strengths – Where does AnVIL excel?; Weaknesses – Where is AnVIL at a disadvantage?; Opportunities: Where can AnVIL grow and improve?; and Threats – Which factors jeopardize AnVIL?

The summary of the SWOT analyses conducted by the participants in each breakout room are combined and available in **Appendix 2**. Recurring themes, general considerations, and suggestions for improvement are summarized here.

Summary of Workshop's Findings

Strengths

Participants highlighted the excellent progress was made during the first 3 years of the AnVIL program. There was recognition of the numerous accomplishments (see p. 13). AnVIL's commitment to facilitate NHGRI-funded consortia's data sharing and access is evident from the impressive growth rate of the data submitted to AnVIL. AnVIL is in a position to reduce the time to share and access datasets and exponentially increase the rate of discovery by consortium members as more datasets get integrated in the platform. Participants were particularly excited by NHGRI's engagement with other NIH Institutes and Centers (ICs) to encourage the adoption of the GA4GH Data Use Ontology (DUO) and the Broad's Data Use Oversight System (DUOS), which are both potential 'game changers' to speed up the data access process. Participants were impressed by AnVIL's commitment to data security and encouraged AnVIL to maintain a high security standard, as well as to ensure that users are aware of the high security standards that the platform is implementing. Workshop participants expressed their enthusiasm and satisfaction for the ease of use and extensive catalog of available tools, including hundreds of workflows and thousands of tools already available within the AnVIL platform. Participants appreciated AnVIL's extensive documentation, which enables users to help themselves and to help each other. Extremely positive comments were shared about the established outreach efforts, such as AnVIL's ongoing work with researchers at a variety of institutions that enables the expansion of genomic data science education and both synchronous and asynchronous user support of novices through power users. AnVIL was viewed as a potentially effective platform for training students and consortia on the use of the cloud for genomic analyses. Finally, participants praised NHGRI's efforts to facilitate systems interoperability and access to data residing on different NIH platforms through the NIH Cloud Platform Interoperability (NCPI) effort and AnVIL's extensive use of APIs and GA4GH standards. Overall, the workshop participants expressed great enthusiasm and support of the AnVIL project and highlighted the relevance of the achievements made in the first three years.

Suggestions for Improvement

Challenges related to the use of the cloud

Participants noted that transitioning to the cloud for data sharing, storage, and analysis is a big culture change for most investigators who have exclusively worked using on-premises infrastructure. Onboarding users who are inexperienced with cloud resources is a significant challenge. Cloud costs are a barrier for new users, both practical and psychological (e.g., paying for compute is still a new concept). AnVIL could explore ways for researchers' institutions to streamline setting up Anvil workspaces and associated billing accounts. AnVIL could also look into the feasibility of adding a "free tier," in which new users can become familiar with AnVIL's capabilities before having to pay for cloud costs. The AnVIL team could make available estimated cloud costs incurred for popular analysis workflows available on the platform to help investigators budget for future projects. Finally, NHGRI could find new ways to leverage

the NIH STRIDES cloud credits and discounts to support users, especially those from low-resource institutions.

Support for the clinical research community

Participants encouraged AnVIL to add tools for clinical genomics implementation to support outreach to clinical genomics groups and help convince them that "AnVIL isn't just for basic research." To make AnVIL more appealing to clinical researchers, AnVIL should consider embedding investigators with clinical research priorities in the development of AnVIL's infrastructure and services as well as to identify use cases for clinical research and analysis. Given AnVIL's aim to adopt the FHIR standard for clinical data exchanges, it could also identify clinical use cases for FHIR beyond data exchanges. AnVIL could aim to establish trust relationships with clinical sites and hospitals and look for ways to bring translational impact of AnVIL services into the clinical domain. AnVIL could also support curation of clinical data that is consistent with ClinGen curation processes.

User outreach and training

Participants suggested several ways for the AnVIL to broaden its user base, in particular students, postdocs, and other trainees, including those who are less skilled in computer science and informatics. For example, AnVIL could become more accessible to naïve users by generating and sharing popular analysis workflows, developing and advertising an extensive video guide to its features, providing streamlined billing and payment for cloud services, and making it easier to access data in the interactive workspaces. AnVIL could make onboarding easier for all users, in particular for users from low-resource institutions. For example, the NHGRI Diversity Action Plan program could collaborate with AnVIL to offer opportunities for mentored skills development on the platform. AnVIL could invest in re-training faculty in data science and supporting curriculum development and documentation for undergraduates.

AnVIL could also leverage the research communities funded by NHGRI (and embedded in the consortia that are integrating into AnVIL) to build a sense of community around the platform. AnVIL could provide support or examples to assist users with describing their analyses in a way that supports publications. The AnVIL team could identify use cases to demonstrate AnVIL's capabilities, promote them in a variety of venues, and provide information for how to get involved and contribute to the platform.

Interoperability, analysis tools, and other considerations

Participants encouraged NHGRI to continue to engage more broadly with other NIH Institutes and Centers (ICs) to help the AnVIL program achieve its goal of improving systems interoperability across NIH cloud platforms. AnVIL could increase its phenotypic data harmonization efforts across NHGRI-funded consortia and initiatives that share datasets through the platform and could support more data standards and common data models to improve semantic interoperability with other platforms.

Supporting the development and availability of more analysis tools would make AnVIL more attractive and accessible for basic and clinical genomics researchers. AnVIL could increase the search capabilities for datasets, tools, and workflows that are available on the platform and rebrand itself from being a data submission site (i.e., repository only) to a multi-functional discovery platform for consortia as well as other genomics and clinical researchers.

NHGRI is encouraged to clarify the relationship of the AnVIL and the data coordination centers of NHGRI-funded consortia and initiatives that are expected to leverage the AnVIL services and to better delineate and communicate its long-term plans for the platform; this will help address investigators' fear of investing time and resources in a new resource that may be short-lived.

Acknowledgements

NHGRI wishes to thank the members of the AnVIL External Consultant Committee and, in particular, the ECC's moderators of the workshop breakout rooms: Ms. Karen Davis, Dr. Siddharth Pratap, Dr. Adam Resnick, and Dr. Marylyn Ritchie as well as all the meeting participants for their valuable suggestions and feedback.

NHGRI Workshop Report Writers

- Valentina Di Francesco
- Carolyn Hutter
- Ana Stevens
- Chris Wellington
- Ken Wiley

Appendix 1

Workshop's Booklet

Purpose of the workshop

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) is a secure, cloud-based environment where researchers can store, share, and analyze unrestricted- and controlled-access genomic datasets and associated phenotypic data or metadata, particularly those generated with NHGRI support. Since 2018 NHGRI has being funding and managing AnVIL through two cooperative agreements awarded to the Broad Institute and Johns Hopkins University.

The goal of this workshop is to identify gaps, challenges, and future opportunities related to NHGRI's investments in the AnVIL's cloud-based infrastructure, tools, and services.

At this workshop we will discuss the current status of AnVIL's ability to serve both the basic and clinical genomics research communities, and identify the activities that are needed to expand, diversify, and support the AnVIL user community.

The workshop agenda includes two sessions, each with two concurrent virtual breakout rooms, focused on the following topics:

1. Data submission and consortia engagement
2. Analysis tools
3. Infrastructure
4. Outreach and training

In each session NHGRI is particularly interested in discussing the following cross-cutting topics, although discussions won't be limited to these:

(a) how cloud-based platforms can better serve the needs of genomic researchers;

(b) what tools and services would better support clinical genomic researchers;

(c) how to improve interoperability with other NIH cloud-based genomic resources in a federated data ecosystem.

All meeting materials and recordings will be made publicly available on the workshop website soon after the end of the workshop.

Agenda

October 29, 2021 – 12pm-5pm ET

12:00-12:10 Welcome and purpose of the workshop
Ms. Valentina Di Francesco (NHGRI) and Dr. Ken L. Wiley Jr. (NHGRI)

12:10-12:20 Data Science at the Forefront of Enhancing Diversity in Genomics
Mr. Vence L. Bonham Jr., J.D. (NHGRI)

12:20-12:35 Introduction to AnVIL
Dr. Anthony A. Philippakis (Broad) and Dr. Michael C. Schatz (Johns Hopkins)

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

Data submission and consortia engagement *Moderators: Dr. Adam Resnick (Children's Hospital of Philadelphia) and Ms. Valentina Di Francesco (NHGRI)*		Analysis tools *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. Anne O'Donnell-Luria (Broad)*
12:55-1:40	Discussion	12:55-1:40	Discussion
1:40-1:50	Prepare breakout report	1:40-1:50	Prepare breakout report

1:50-2:15 Report back from Session 1
Rapporteurs: Dr. Adam Resnick and Dr. Marylyn Ritchie

2:15-2:30 15 min break

2:30-3:45 Session 2: Breakout rooms

Infrastructure *Moderators: Ms. Karen M. Davis (RTI International) and Dr. Carolyn M. Hutter (NHGRI)*		Outreach and training *Moderators: Dr. Siddharth Pratap (Meharry Medical College) and Mr. Christopher Wellington (NHGRI)*	
2:30-2:35	Moderator introduction	2:30-2:35	Moderator introduction
2:35-2:50	AnVIL presentation: *Dr. Jeremy Goecks (OHSU) and Dr. Benedict Paten (UCSC)*	2:35-2:50	AnVIL presentation: *Dr. Jeffrey Leek (JHU) and Ms. Tiffany Miller (Broad)*
2:50-3:35	Discussion	2:50-3:35	Discussion
3:35-3:45	Prepare breakout report	3:35-3:45	Prepare breakout report

3:45-4:10 Report back from Session 2
Rapporteurs: Ms. Karen M. Davis and Dr. Siddharth Pratap

4:10-5:00 Summary and closing
Ms. Valentina Di Francesco and Dr. Ken L. Wiley Jr.

Discussant assignments to breakout rooms

Session 1

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. Stephen (Steve) Rich
Dr. Kenneth (Ken) Rice

Breakout room: Analysis tools

Moderators: Dr. Marylyn Ritchie and Dr. Ken Wiley, Jr.

Dr. Nadav Ahituv
Dr. Joshua (Josh) Akey
Dr. Mark Craven
Dr. Sean Davis
Dr. Barbara Engelhardt
Dr. James Knight
Dr. Anshul Kundaje
Dr. Karen Miga
Dr. Adam Phillippy
Dr. Timothy (Tim) Reddy
Dr. Chunhua Weng

Session 2

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

Breakout room: Outreach and training

Moderators: Dr. Siddharth Pratap and Mr. Christopher Wellington

Dr. Cinnamon Bloss
Dr. C. Titus Brown
Dr. Carol Bult
Dr. John Kwagyan
Dr. Andrew Lee
Dr. Robert Meller
Dr. Peter Robinson
Dr. Sourav Roy
Dr. William (Bill) Southerland

Invited Participants

Dr. Nadav Ahituv*
University of California, San Francisco
Nadav.Ahituv@ucsf.edu

Dr. Joshua Akey
Princeton University
jakey@princeton.edu

Mr. Samuel (Sandy) Aronson
Partners Personalized Medicine
saronson@partners.org

Dr. Cinnamon Bloss*
University of California, San Diego
cbloss@ucsd.edu

Dr. Elizabeth (Liz) Blue
University of Washington
em27@uw.edu

Dr. Vivien Bonazzi
Deloitte
vbonazzi@deloitte.com

Dr. C. Titus Brown
University of California, Davis
ctbrown@ucdavis.edu

Dr. Carol Bult*
Jackson Laboratory
Carol.Bult@jax.org

Dr. Mark Craven^
University of Wisconsin-Madison
craven@biostat.wisc.edu

Dr. David Crosslin
Tulane University
crosslin@tulane.edu

Ms. Karen L. Davis*
RTI International
kdavis@rti.org

Dr. Sean Davis*
University of Colorado Denver
SEAN.2.DAVIS@CUANSCHUTZ.EDU

Dr. Brandi Davis-Dusenbery
Seven Bridges
brandi@sbgenomics.com

Dr. Barbara Engelhardt
Princeton University
bee@princeton.edu

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. George Hripcsak*
Columbia University
gh13@cumc.columbia.edu

Dr. Eimear Kenny
Icahn School of Medicine at Mount Sinai
eimear.kenny@mssm.edu

Dr. James Knight
Yale School of Medicine
j.knight@yale.edu

Dr. Iftikhar Kullo
Mayo Clinic
kullo.iftikhar@mayo.edu

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

*notes a member of the AnVIL External Consultant Committee (ECC)
^ notes a member of the National Advisory Council for Human Genome Research

Dr. John Kwagyan
Howard University
jkwagyan@howard.edu

Dr. Andrew Lee
Northern Virginia Community College
ajlee@nvcc.edu

Dr. Tara Matise
Rutgers University
matise@dls.rutgers.edu

Dr. Shannon McWeeney
Oregon Health & Science University
mcweeney@ohsu.edu

Dr. Robert Meller
Morehouse School of Medicine
rmeller@msm.edu

Dr. Karen Miga
University of California, Santa Cruz
khmiga@soe.ucsc.edu

Dr. Aleksandar Milosavljevic*
Baylor College of Medicine
amilosav@bcm.edu

Dr. Lucila Ohno-Machado
University of California, San Diego
lohnomachado@health.ucsd.edu

Dr. Minoli Perera
Northwestern University
minoli.perera@northwestern.edu

Dr. Adam Phillippy
National Human Genome Research Institute
adam.phillippy@nih.gov

Dr. Siddharth (Sidd) Pratap*
Meharry Medical College
spratap@mmc.edu

Mr. Luke Rasmussen
Northwestern University
luke.rasmussen@northwestern.edu

Dr. Timothy (Tim) Reddy
Duke University
tim.reddy@duke.edu

Dr. Adam Resnick*
Children's Hospital of Philadelphia
RESNICK@chop.edu

Dr. Kenneth (Ken) Rice
University of Washington
kenrice@uw.edu

Dr. Stephen (Steve) Rich
University of Virginia
ssr4n@virginia.edu

Dr. Marylyn Ritchie*
University of Pennsylvania
marylyn@upenn.edu

Dr. Peter Robinson^
Jackson Laboratory
peter.robinson@jax.org

Dr. Sourav Roy
University of Texas at El Paso
sroy1@utep.edu

Dr. William (Bill) Southerland
Howard University
wsoutherland@howard.edu

Dr. Chunhua Weng
Columbia University
chunhua@columbia.edu

AnVIL Awardees

Broad Institute award

Dr. Anthony Philippakis (PI)
Broad Institute
aphilipp@broadinstitute.org

Ms. Mary (Katie) Banasiewicz
Vanderbilt University Medical Center
mary.k.banasiewicz@vumc.org

Dr. Eric Banks
Broad Institute
ebanks@broadinstitute.org

Mr. David Bernick
Broad Institute
dbernick@broadinstitute.org

Dr. Robert Carroll
Vanderbilt University Medical Center
robert.j.carroll@vumc.org

Dr. Robert Grossman
University of Chicago
robert.grossman@uchicago.edu

Dr. Ira Hall
Yale School of Medicine
ira.hall@yale.edu

Dr. Jennifer Hall
American Heart Association
jennifer.hall@heart.org

Dr. Tim Harris
University of California, Santa Cruz
tijharri@ucsc.edu

Ms. Tiffany Miller
Broad Institute
tiffanym@broadinstitute.org

Dr. Brian O'Connor
Broad Institute
boconnor@broadinstitute.org

Dr. Anne O'Donnell-Luria
Broad Institute
odonnell@broadinstitute.org

Dr. Andre Paredes
University of Chicago
paredes@uchicago.edu

Dr. Benedict Paten
University of California, Santa Cruz
bpaten@ucsc.edu

Ms. Candace Patterson
Broad Institute
candace@broadinstitute.org

Ms. Valerie Reeves
Broad Institute
vreeves@broadinstitute.org

Ms. Radhika Reddy
University of Chicago
reddyr@uchicago.edu

Mr. David Rogers
Clever Canary
dave@clevercanary.com

Mr. Jason Walker
Washington University in St. Louis
Jason.walker@wustl.edu

Dr. Ting Wang
Washington University in St. Louis
twang@genetics.wustl.edu

Dr. Deena Zytnick
American Heart Association
deena.zytnick@heart.org

Johns Hopkins University award

Dr. Michael C. Schatz (PI)
Johns Hopkins University
mschatz@jhu.edu

Dr. Enis Afgan
GalaxyWorks/Johns Hopkins University
afgane@gmail.com

Dr. Vincent J. Carey
Harvard Med. School/Brigham & Women's Hosp.
stvjc@channing.harvard.edu

Dr. Kyle Ellrott
Oregon Health & Science University
ellrott@ohsu.edu

Dr. Jeremy Goecks
Oregon Health & Science University
goecksj@ohsu.edu

Dr. Kasper Hansen
Johns Hopkins University
khansen@jhsph.edu

Dr. Ava Hoffman
Johns Hopkins University
ava.hoffman@jhu.edu

Ms. Natalie Kucher
Johns Hopkins University
nkucher3@jhu.edu

Dr. Jeffrey T. Leek
Johns Hopkins University
jtleek@hey.com

Dr. Martin Morgan
Roswell Park Comprehensive Cancer Center
martin.morgan@roswellpark.org

Dr. Stephen Mosher
Johns Hopkins University
smosher3@jhu.edu

Dr. Anton Nekrutenko
Pennsylvania State University
aun1@psu.edu

Ms. Lori Shepherd
Roswell Park Comprehensive Cancer Center
lori.shepherd@roswellpark.org

Dr. Frederick Tan
Carnegie Institution for Science
tan@carnegiescience.edu

Dr. Casey Overby Taylor
Johns Hopkins Medicine
cot@jhu.edu

Dr. Levi Waldron
CUNY Graduate School of Public Health & Policy
levi.waldron@sph.cuny.edu

Ms. Jennifer Vessio
Johns Hopkins University
jvessio1@jhu.edu

NHGRI AnVIL Staff

Ms. Valentina Di Francesco
valentina.difrancesco@nih.gov

Ms. Elena Ghanaim
elena.ghanaim@nih.gov

Dr. Carolyn M. Hutter
carolyn.hutter@nih.gov

Dr. Teri Manolio
teri.manolio@nih.gov

Dr. Tiffany Rolle
tiffany.rolle@nih.gov

Ms. Ana Stevens
ana.stevens@nih.gov

Ms. Helen Thompson
helen.thompson@nih.gov

Mr. Christopher Wellington
chris.wellington@nih.gov

Dr. Ken L. Wiley, Jr.
ken.wiley@nih.gov

All NHGRI extramural staff is invited to participate in this workshop as listeners only.

Appendix 2

Strengths - Where does AnVIL excel?
Strong commitment to data security.
A versatile platform for training students and consortia on the use of the cloud for genomic analyses.
NHGRI's Leadership of the NCPI efforts and AnVIL's extensive use of APIs and GA4GH standards facilitates interoperability.
Extensive documentation, which enables users to help themselves and help each other.
Ability by third party groups to build on the platform while still prioritizing security.
NHGRI's robust engagement with other NIH Institutes and Centers to encourage the adoption of the GA4GH Data Use Ontology (DUO) and Data Use Oversight System (DUOS) to streamline the data access review process.

Weaknesses - Where is AnVIL at a disadvantage?
Curation of tools and workflows to facilitate searches by users could be improved.
Lack of phenotypic harmonization across programs.
The relationships between AnVIL and data coordination centers of NHGRI funded consortia could be better defined.
Significant hurdles required to access AnVIL just to test the platform.
Lack of video documentation.
Users cannot log in anonymously.
AnVIL lacks embedded personnel (e.g., key personnel and developers) with clinical research priorities.

Opportunities: Where AnVIL can grow and improve?
AnVIL could create a safe space for groups hesitant to host diverse controlled access datasets in public repositories.
AnVIL could transition over to tools development and analyses for discovery, in addition to acting as a data repository.
The AnVIL team could demonstrate that AnVIL can work for the clinical community.
AnVIL could add additional data standards and data models to improve the interoperability model for Terra.
The outreach team could conduct robust research into how AnVIL is being used.
AnVIL has the opportunity to introduce cloud computing for the next generation of scientists.
AnVIL could integrate a diversity of human genetic datasets and reference genomes and make them backwards compatible.

Threats: Which factors jeopardize AnVIL?
Cloud costs are a major barrier for many users.
Challenges in making AnVIL interoperable with other platforms.
Difficulties in shifting scientific culture to the cloud.
Challenges in making tools and resources in a manner that meets users where they are.
AnVIL faces data security threats, both by outsiders and by people with authorized access.
Potential institutional fear of investing in a dead-end utility if NHGRI's long term commitment to the resource is unclear.
Potential lack of users' skill transferability between AnVIL and other cloud platforms.