RESEARCH ARTICLE

Insights into the Datasets, Tools, and Training Needs of the AnVIL Community: 2024

[version 1; peer review: 1 approved with reservations]

Kathryn J. Isaac[1], Katherine E. L. Cox[2], Kai Yin Ho[3], Elizabeth M. Humphries[1], Natalie Kucher[2], Jeffrey T. Leek[1], Stephen Mosher[2], Michael C. Schatz[2,4], Frederick J. Tan[2], Ava M. Hoffman[1,5]

[1]Biostatistics Program, Fred Hutchinson Cancer Center, Seattle, Washington, USA
[2]Department of Biology, Johns Hopkins University, Baltimore, Maryland, USA
[3]Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA
[5]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA

v1 **First published:** 22 Dec 2025, **14**:1424
https://doi.org/10.12688/f1000research.173844.1
Latest published: 22 Dec 2025, **14**:1424
https://doi.org/10.12688/f1000research.173844.1

Abstract

Background

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) provides a secure cloud-based environment where research and education communities can analyze genomic and biomedical data. The platform supports a wide range of data analysis as well as the ability to safely store and access data in compliance with NIH policies. The purpose of this study is to better understand the current user base of the AnVIL platform.

Methods

We conducted the AnVIL Community Poll to collect baseline information, identify development opportunities, guide the prioritization of user support strategies, and succinctly but comprehensively describe the current AnVIL Community. The poll was shared through social media and relevant mailing lists. Respondents were categorized as potential or returning users depending on their usage description.

Results

Our sample of the AnVIL community found opportunities for platform

Open Peer Review

Approval Status ?

	1
version 1 22 Dec 2025	? view

1. **Hassan Rehan**, Purdue University, West Lafayette, USA

Any reports and responses or comments on the article can be found at the end of the article.

adoption beyond the current user base and identified areas where training should be enhanced, training preferences, and user computational needs. Specifically, while most respondents were involved in human genomics research, there may be potential for growth in adoption of the platform by prioritizing materials to support clinical researchers. All respondents felt availability of specific tools or datasets was a key feature of the platform. The broader community may also benefit from further development or showcasing of resources to facilitate cost management, finding and incorporating analysis tools, and data import. Our sample greatly preferred virtual training opportunities and returning users of the platform foresaw needing large amounts of storage.

Conclusion

This poll provided an insightful snapshot of the current state of the AnVIL and demonstrated areas where the AnVIL Team can take specific steps to address barriers related to platform adoption and further support the existing and varied AnVIL Community. This work can be built upon through user interviews, community discussion, and coordinating a recurring poll.

Keywords
cloud computing, genomics, user research, precision medicine

This article is included in the Genomics and Genetics gateway.

This article is included in the Bioinformatics gateway.

Corresponding author: Kathryn J. Isaac (kisaac@fredhutch.org)

Author roles: Isaac KJ: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Cox KEL**: Methodology, Writing – Review & Editing; **Ho KY**: Methodology, Writing – Review & Editing; **Humphries EM**: Data Curation, Methodology, Writing – Review & Editing; **Kucher N**: Methodology, Project Administration, Writing – Review & Editing; **Leek JT**: Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **Mosher S**: Methodology, Project Administration, Writing – Review & Editing; **Schatz MC**: Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **Tan FJ**: Conceptualization, Funding Acquisition, Methodology, Visualization, Writing – Review & Editing; **Hoffman AM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was carried out under National Institutes of Health, National Human Genome Research Institute Project Number 5U24HG010263.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Isaac KJ, Cox KEL, Ho KY *et al.* **Insights into the Datasets, Tools, and Training Needs of the AnVIL Community: 2024 [version 1; peer review: 1 approved with reservations]** F1000Research 2025, **14**:1424
https://doi.org/10.12688/f1000research.173844.1

First published: 22 Dec 2025, **14**:1424 https://doi.org/10.12688/f1000research.173844.1

Introduction

The exponential growth in biomedical data generation has created unprecedented challenges in data storage, analysis, and sharing.[1,2] Despite its potential for research and teaching applications,[3] cloud computing platforms in biomedicine have historically faced adoption challenges. Would-be users struggle with technical complexity around data transfer and user interface,[4] familiarity and onboarding,[3] cost management,[5] and data security concerns.[4,6] Understanding the user's activity, their backgrounds, and perspective can help address these concerns and guide platform development, adoption, and impact.[7–12] For example, usage metrics collected by the Cancer Genomics Cloud have revealed time savings available to researchers in cloud-based genomic analysis environments.[13] Direct polling of users can also inform potential areas of improvement. The Galaxy Project team routinely gauges satisfaction with their hands-on activities.[14,15] Similarly, Nextflow community leaders have surveyed users to understand their transition to cloud computing and other preferences.[16] However, less is publicly known about users and their preferences on other platforms.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)[17] has emerged as a response to data storage, analysis, and sharing challenges, offering a federated cloud platform that integrates data management, analysis capabilities, and access controls.[1] While AnVIL and other cloud platforms are enabling the shift from traditional institutional computing clusters to cloud-based genomic analysis,[18] understanding user experiences and potential challenges will be central to AnVIL's ability to meet the needs of cutting-edge research.[19,20] Little is formally known about the types of users that leverage AnVIL, including their experiences, preferences, and awareness of documentation and resources.

As a comprehensive environment that includes Terra, Galaxy, RStudio/Bioconductor, Dockstore, workflows and Jupyter Notebooks, AnVIL offers multiple paths for data analysis and collaboration. However, flexibility also introduces complexity for the researcher choosing their analytical approach. AnVIL's highly customizable environment also makes it challenging for platform developers to identify specific challenges users are having that can be met with actionable improvements. For example, developers might not know whether users are unfamiliar with the AnVIL implementation of the tool (such as "Bioconductor on AnVIL") or are simply unfamiliar with the tool more generally. AnVIL's innovative approach to data sharing, where researchers "move to the data" rather than downloading datasets locally, leads to more unknowns, such as which datasets are most in demand and what kinds of data (genomic, metabolomic, etc.) developers should keep in mind. The most important features of the AnVIL platform might also differ between returning and potential new users.

To systematically investigate user perspectives, we conducted a community poll targeting returning and potential AnVIL users. Our study aimed to:

1. Examine the background and current work of users to develop appropriate personas
2. Understand barriers to platform adoption and user preferences for training and support
3. Assess researchers' technological comfort with cloud-based genomic analysis tools
4. Identify computational and data analysis resource needs

By providing a detailed understanding of user experiences, this user poll seeks to inform future platform development, enhance user support strategies, and ultimately accelerate the adoption of cloud-based genomic analysis technologies.

Methods

Design of the poll

Given that AnVIL is a newer platform that has never polled users before, this community poll was designed to collect baseline information specifically with regards to the aforementioned study aims (Table 1). Questions were sourced from the AnVIL team as a whole – a combined effort across Johns Hopkins University, the Broad Institute, Fred Hutch Cancer Center, Vanderbilt University Medical Center, and other institutions within the AnVIL team.

The AnVIL Community Poll was constructed using Google Forms. The poll had 6 parts (Figure 1, Table 1). In the first part, using a multiple choice question, respondents were asked to select the best description of their current usage of the AnVIL platform. We used the answers to categorize respondents as returning or potential users of the AnVIL platform. Going forward, we discuss usertypes as either returning or potential users of the AnVIL. We define returning users as those who selected any of the following: "For ongoing projects (e.g., consistent project development and/or work)", "For short-term projects (e.g., short, intense bursts separated by a few months)", or "For completed/long-term projects (e.g.,

Table 1. Study aims related to the design of the State of the AnVIL 2024 Community Poll. Sections of the State of the AnVIL 2024 Community Poll (Part 1, Part 2, etc.) are listed in the first column as the row names. Column names are the enumerated study aims. X's are added at the intersection of a poll section and a study aim if questions within that section are relevant to the study aim. Supplemental material (see **Data and Software Availability**) provides a further breakdown of how each question relates to the study aims.

	Examine the background and current work of users to develop appropriate personas	Understand barriers to platform adoption and user preferences for training and support	Assess researchers' technological comfort with cloud-based genomic analysis tools	Identify computational and data analysis resource needs
Part 1: Self-identify	*x*			
Part 2: Feature importance + Returning user specific Q's	*x*	*x*	*x*	*x*
Part 3: Demographics	*x*			
Part 4: Experience	*x*		*x*	*x*
Part 5: Awareness		*x*		
Part 6: Preferences		*x*	*x*	*x*

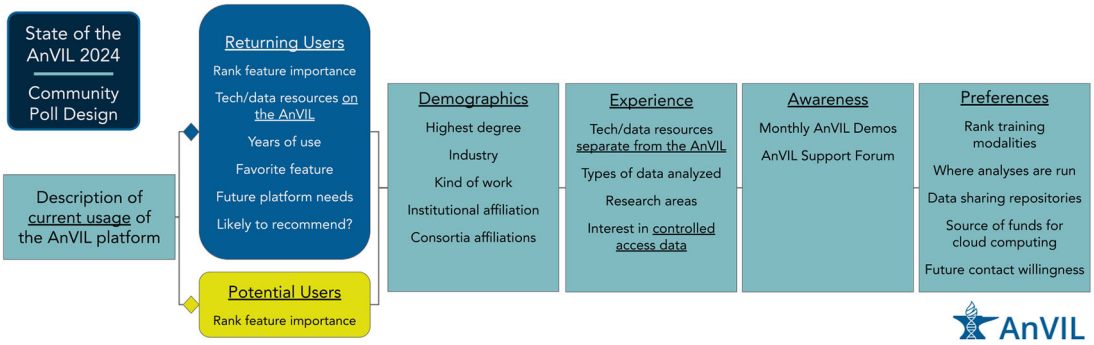

Figure 1. State of the AnVIL 2024 Community Poll Design. The State of the AnVIL 2024 Community Poll was designed such that there were 6 parts. Every respondent was asked the same questions for Parts 1, 3, 4, 5, and 6 while respondents were provided with user type specific questions in Part 2, depending upon their answer to the first question of the poll. Part 1 was used to classify a respondent as a returning or potential user of the AnVIL; Part 2 asked questions specific to each user type (returning or potential); Parts 3-6 contained demographics, experience, awareness, and preferences questions respectively. Brief descriptions of questions within each part are provided.

occasional updates/maintenance as needed)"; and potential users as those who selected any of the following: "I do not currently use the AnVIL, but have in the past", "I have never used the AnVIL, but have heard of it", or "I have never heard of the AnVIL". Depending on the answer to this question, respondents were directed to the second part as either a returning or potential user of the AnVIL platform. Part 2 of the poll had a single question for potential users or 7 questions for returning users. The 6 additional questions asked about their experience with, needs, and recommendation likelihood for using the AnVIL while both categories of respondents were asked to rank features of the AnVIL according to their importance for their continued (returning user) or potential (potential user) use of the platform. Every respondent was given the same questions for the remaining parts of the poll (parts 3-6), with 5, 5, 2, and 6 questions respectively for each section. If respondents assented that they were willing to be contacted again to give input in the future, then they were directed to provide an email address. Finally, all users were provided with a link to a separate google form in case they wanted to be entered in a potential prize raffle. Due to there being a total of 20-26 questions, we estimated the poll would take 10-15 minutes to complete (~30 seconds/question).

The outline of the poll is as follows (Figure 1):

Part 1: Self-identify as a returning or potential user

Part 2: Rank existing or potential features according to their importance (everyone)

+ returning AnVIL user specific questions about experience with the platform

Part 3: Demographics (e.g., degree level, type of work, institution, consortia affiliations, etc.)

Part 4: Experience (e.g., general bioinformatics experience, relevant datasets, etc.)

Part 5: Awareness and utilization of available AnVIL support

Part 6: Preferences (e.g., training modality, analyses platforms, etc.)

For a pdf version of the full poll, see the **Data and Software Availability**.

Recruitment for/Dissemination of the poll

The AnVIL Community Poll was disseminated through multiple channels including posting on social media platforms like X (formerly Twitter) and LinkedIn, emailing applicable mailing lists (registered AnVIL users through Terra, AnVIL Demo attendees, AnVIL mailing list, etc.), and posting within Slack workspaces (AnVIL, nf-co.re, Community Bioconductor, Fred Hutch Cancer Center, and the Johns Hopkins University Genomics and Biostat communities). We also posted an advertisement for the survey as a news item on the anvilproject.org portal and the AnVIL Support Forum (help.anvilproject.org). Finally, the poll was sent directly to some select individuals and consortia who were asked to further circulate it. All advertisements mentioned the possibility of a prize raffle. The poll was conducted in Spring 2024. It was administered using Google Forms and open for responses from February 15th to March 25th.

Because of the nature of dissemination (through social media), we cannot confidently calculate a response rate. However, we consider this to be a data gathering exercise where we can report back the findings to the AnVIL Community and see if and how they resonate.

Ethical considerations

This work has been reviewed by The Johns Hopkins University Homewood IRB and was determined to not-qualify as human subject research (#HIRB00020632). Participating in the poll was completely voluntary, and the results were de-identified.

Analysis of the poll

Following the closing of the poll, the Google sheet with form responses was locked so that edits could not be made and the results were imported from there for analysis using R. Respondent email (if provided) was used to de-duplicate responses. Identifying information or information that was highly specific and potentially identifiable like specific institution names was removed (email addresses) or annotated and grouped into categories (type of institution, e.g., an R1 University, R2 University, Community College, etc.) before removal.

Respondents were asked about their level of experience with human genomic, non-human genomic, and clinical research. Possible responses included “Not at all experienced”, “Slightly experienced”, “Somewhat experienced”, “Moderately experienced”, and “Extremely experienced” (a 5 point Likert scale). If a respondent selected “Moderately” or “Extremely” experienced for any of these categories, they were assigned as someone with experience in that research category. This method of assigning experience is an example of a Top 2 Box simplification.[21]

Poll takers were also asked to select the type(s) of work they perform regularly Possible choices for type(s) of work were Computational work, Engineering work, Wet lab work, Clinical work, Computational education, Wet lab education, Project leadership, Project management, Program administration, or Other (with free text entry if Other was selected). For analysis purposes, we grouped users into 5 possible personas: Clinician, Analyst, Educator, Leadership (PI), or Admin. These groups were created based on participant answers to the “type of work” question. A clinician persona was assigned if “Clinical work” was selected; An analyst persona was assigned if only “Computational work” was selected; An educator persona was assigned if “education” appeared in a response; A leadership persona was assigned if project leadership, management, or administration co-occurred with “Computational work”; and an Admin persona was assigned

if any of project leadership, management or administration was selected without any other kinds of work selected. With this assignment procedure, some responses were not assigned personas and two were assigned multiple personas. To simplify those assigned multiple personas, the selections were individually inspected and categorized. We also performed principal components analysis, and did not observe any conspicuous clustering of respondents (see https://github.com/fhdsl/SOTA2024_ReportOut).

When respondents were asked to rank preferences or level of importance (Figure 5), responses were translated into rank values. If n choices were given, "Most preferred/important" was assigned a value of 1, with each subsequent value denoting decreasing preference or importance. The "Least important/preferred" was assigned the largest possible value. An overall or average rank choice was assigned by summing the rank values and dividing by the total number of respondents. For an average knowledge or comfort score (Figure 7A), a similar procedure was followed; however, for these questions the higher value was associated with greater knowledge or comfort and a score of 0 was used for responses of not knowing at all. Rank and comfort scores were averaged within user cohorts (e.g., persona or user type) rather than considering all respondents together.

In Part 5 of the poll, respondents were asked about their experiences with user support options. Respondents could select "No, didn't know of", "No, but aware of", or some specific description related to how they used the support option (e.g., attending a demo or answering someone's post on the support forum). We broadly translated these responses into two binary reporters: one for awareness and the other for active use. To understand awareness, we combined responses in the following way: Anything other than "No, didn't know of" was combined to represent awareness. To understand use, we combined responses such that anything other than "No, didn't know of" and "No, but aware of" were examples of use or future use. If looking at past use (ignoring future use), "Not yet, but registered to" was also not included in answers representing use. (See **Data and Software Availability** for a table representing these categorizations.)

All code for reproducing the stats and figures in this analysis are available on GitHub: https://github.com/fhdsl/SOTA2024_ReportOut. A de-identified dataset is also available.

Results

Responses

This AnVIL user poll was conducted in Spring 2024, receiving 52 total responses. Two responses were determined to be submitted by duplicate users, leaving a total of 50 user responses used in the analysis. Given how the AnVIL Community Poll was advertised through general, often public channels rather than enumerated closed lists, it was not possible to compute an exact response rate. Considering the distribution of daily responses compared to recruitment efforts for the poll, responses were received on days without advertisement, but days with advertisement (at least 7 unique days) aligned with or directly preceded receiving responses. The two days with the highest response count (7 and 8 responses) aligned with a user listserv reminder email (2024-March-13) and an X (formerly Twitter) post (2024-March-14). The average number of responses in a day was about 1, though the most common number of responses was 0 (21 different days), followed by 1 (8 different days) if restricted to non-zero responses. When considering the institutional affiliation of respondents, there were no higher than 2 responses from users from the same institution on any day. Both potential and returning users responded throughout the time period with no discernible pattern related to recruitment method.

User backgrounds and current work

Of the 50 de-duplicated responses, 44% (n = 22) were from returning users and 56% (n = 28) were from potential users. The majority of returning user responses (73%, n = 16) belonged to the group who use the AnVIL for ongoing projects, with consistent work on the platform. The majority of potential users were evenly split between two groups (46%, n = 13 each), those who have never used the AnVIL (but have heard of it) and those who have used the AnVIL previously, but don't currently (Figure 2A).

Most respondents in our sample had obtained or were in the process of obtaining a PhD (62%, n = 31) though a range of career stages were represented (Figure 2B) including those with little to no post-secondary education and those with multiple advanced degrees. Clinical advanced degrees (MDs) (6%, n = 3) were less represented than either research degrees (PhD) or Master's degrees (18%, n = 9).

Most of the respondents using the AnVIL (n = 21) reported being affiliated with a research intensive institution (e.g., an R1 University, a medical campus, a research center, or the NIH), while one reported being affiliated with an education focused institution (e.g., R2 University or community college). Only potential users of the platform reported being affiliated with an industry-based institution (Figure 2C). In our sample, we observed very few industry partners.

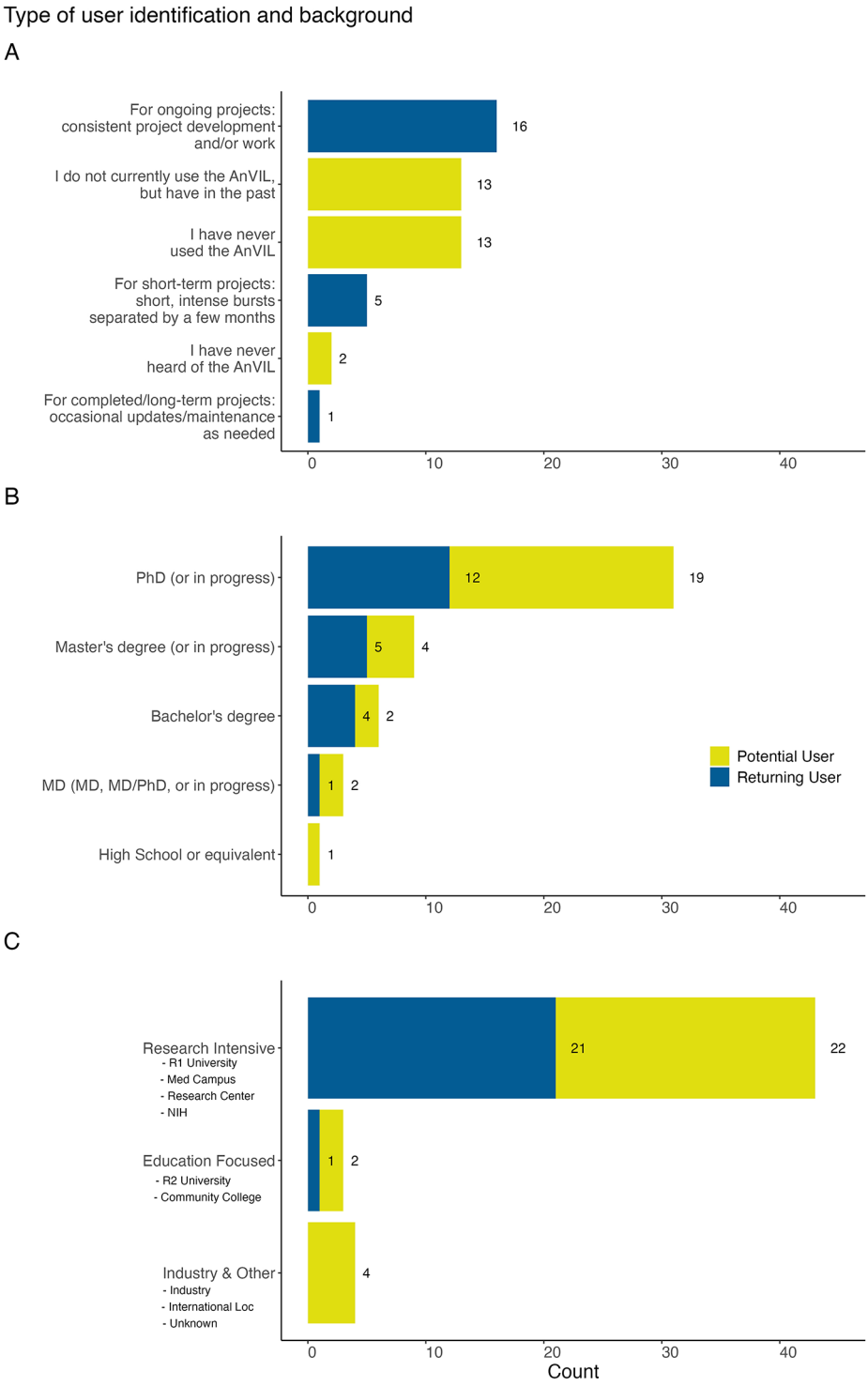

Figure 2. Background of users in our sample. A: We asked all respondents "How would you describe your current usage of the AnVIL platform?" Respondents were identified as returning or potential AnVIL users based on their responses. Most of the 22 returning users leverage AnVIL for ongoing projects. The 28 potential users were evenly split between those who have never used the AnVIL (but have heard of it) and those who have used the AnVIL previously, but don't currently. B: We asked "What is the highest degree you have attained?" Most of the respondents have a PhD or are currently working on a PhD, though a range of career stages were represented. C: We asked all respondents "What institution are you affiliated with?" Most of the respondents, but also the majority of returning individuals using the AnVIL, reported being affiliated with a research intensive institution.

Of the 50 responses, 21 provided consortia affiliations across 23 unique affiliations (respondents could report more than one consortium). Of the 21 responses providing consortia affiliations, 13 were from returning users. Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR),[22,23] Polygenic Risk Methods in Diverse Populations (PRIMED),[24,25] Electronic Medical Records and Genomics (eMERGE),[26] Centers for Common Disease Genomics (CCDG)[27] and The Genotype-Tissue Expression Project (GTEx)[28] were the top 5 most represented with n=3 each for GREGoR, PRIMED, and eMERGE and n=2 each for CCDG and GTEx. Additionally, all 5 of these consortia were represented in the subset of consortia affiliations reported by returning users.

When asked about experience with analyzing human genomic, human clinical, and non-human genomic data, 21 respondents report that they are extremely experienced in analyzing human genomic data, while only 6 respondents report that they are not at all experienced in analyzing human genomic data. However, for human clinical data and non-human genomic data, more respondents report being not at all experienced in analyzing those data than report being extremely experienced (Figure 3A). Returning and potential users showed a similar distribution on experience levels across these 3 research categories. If we consider the Top 2 Box simplification, lumping together "Moderately" and "Extremely" experienced responses (the highest 2 of the possible rankings), returning users show slightly more experience across all 3 research categories.

When limiting the responses we consider to include only respondents who reported being "moderately" or "extremely" experienced for at least one of the categories, we are left with 37 responses (17 from returning users). Considering the overlap among experience within these categories for these responses, 32% (n = 12) reported these levels of experience for all 3 research categories and 27% (n = 10) reported this level of experience for only human genomic research (but no other research categories) (Figure 3B). 16% (n = 6) reported this level of experience for both human clinical and human genomic research (but not non-human genomic research) and another 16% (n = 6) reported this level of experience in analyzing both non-human and human genomic research.

When asked to select the kind of work that they performed, 44% (n = 22) selected only 1 description, 36% (n = 18) selected 2 descriptions, 16% (n = 8) selected 3 descriptions, and 4% (n=2) selected 5 descriptions. "Computational work" was the most frequently selected response, reported by 68% of respondents (n = 34). "Project management" and "Project leadership" followed at 36% (n = 18) and 24% (n = 12) respectively. Of the 22 respondents who selected only 1 description, "Computational work" was the most frequently selected (n = 13, 59%). Of the responses who selected more than one description, "Computational work" was paired most frequently with "Project management" or "Project leadership". (n = 11 and 7 respectively). Two respondents utilized the "Other" option, each supplying one work description. These included "Cloud Architect" and "Software Development". The distribution of work descriptions selected were similar when comparing between potential and returning users (Figure 4).

All of the results summarized so far describe the varied backgrounds and current work of respondents in our sample. Despite this variety, we expect several overarching user groups: Admins, Analysts, Clinicians, PIs, and Educators. Given that the kind of work question focused on work activities and that these expected user groups have differential work activities, we assigned personas based on the responses to the kind of work question (Figure 4). 10 respondents were left uncategorized, but the majority of respondents were assigned to either the Leadership (PI) persona (n = 13, showing evidence of computational work together with some form of project management, leadership, or administration) or the Analyst persona (n = 13, only reporting computational work). 7 respondents were categorized as an Admin persona. The Leadership (PI), Analyst, and Admin personas showed a similar split of assignment across potential and returning users. The Clinician and Educator personas were differentially represented in this sample: 2 potential users were assigned a Clinician persona because of selecting clinical work (this option was not selected by returning users) and 4 potential users were assigned an Educator persona while only 1 returning user was.

Barriers & user preferences

The kind of work that researchers perform or different hats that respondents wear are likely to influence their preferences and barriers to adoption. Respondents were asked about features that are most important for their continued or potential use of the AnVIL. Of the features listed, currently the AnVIL utilizes use-based billing and does not offer a free version. When looking at responses specific to the assigned personas, we observe that the average rank choice for easy billing setup is highest for Admins compared to other personas; Educators and Clinicians rank both a free version with limited compute or storage and greater adoption by the scientific community more highly than other personas; and PIs rank having specific tools or datasets available/supported as their most important feature of AnVIL. Specific tools or datasets being available and supported ranks as the most important feature for both returning and potential users of the AnVIL (Figure 5A). Potential users rank having a free version of the AnVIL with limited compute or storage more highly than

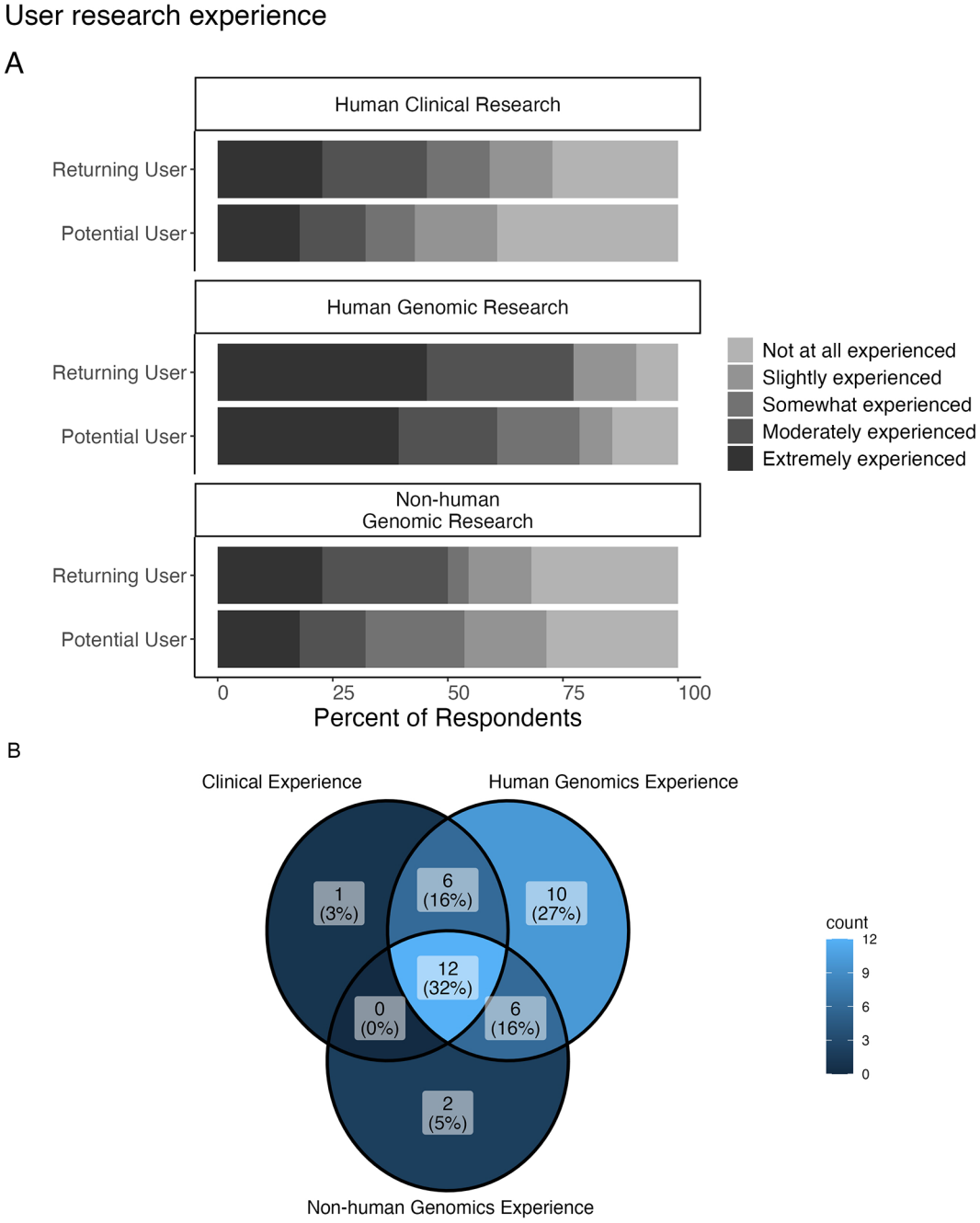

Figure 3. Background of users and researcher's technological comfort. A: We asked respondents "How much experience do you have analyzing the following data categories?" 21 respondents report that they are "extremely experienced" in analyzing human genomic data, while only 6 respondents report that they are "not at all experienced" in analyzing human genomic data. However, more respondents report being "not at all experienced" in analyzing human clinical data and non-human genomic data. B: Venn diagram showing the overlap for respondents who reported being "moderately" or "extremely experienced" for these various research categories (n = 37). 32% reported such levels of experience for all 3 research categories; the next highest percentage (27%) reported such levels of experience for human genomic data only with no overlap with other research categories.

returning users do. Easy billing setup and on demand support and documentation are ranked in the middle with similar preferences observed across user types. Flat-rate billing rather than use-based billing is ranked the lowest across this set of features within this sample.

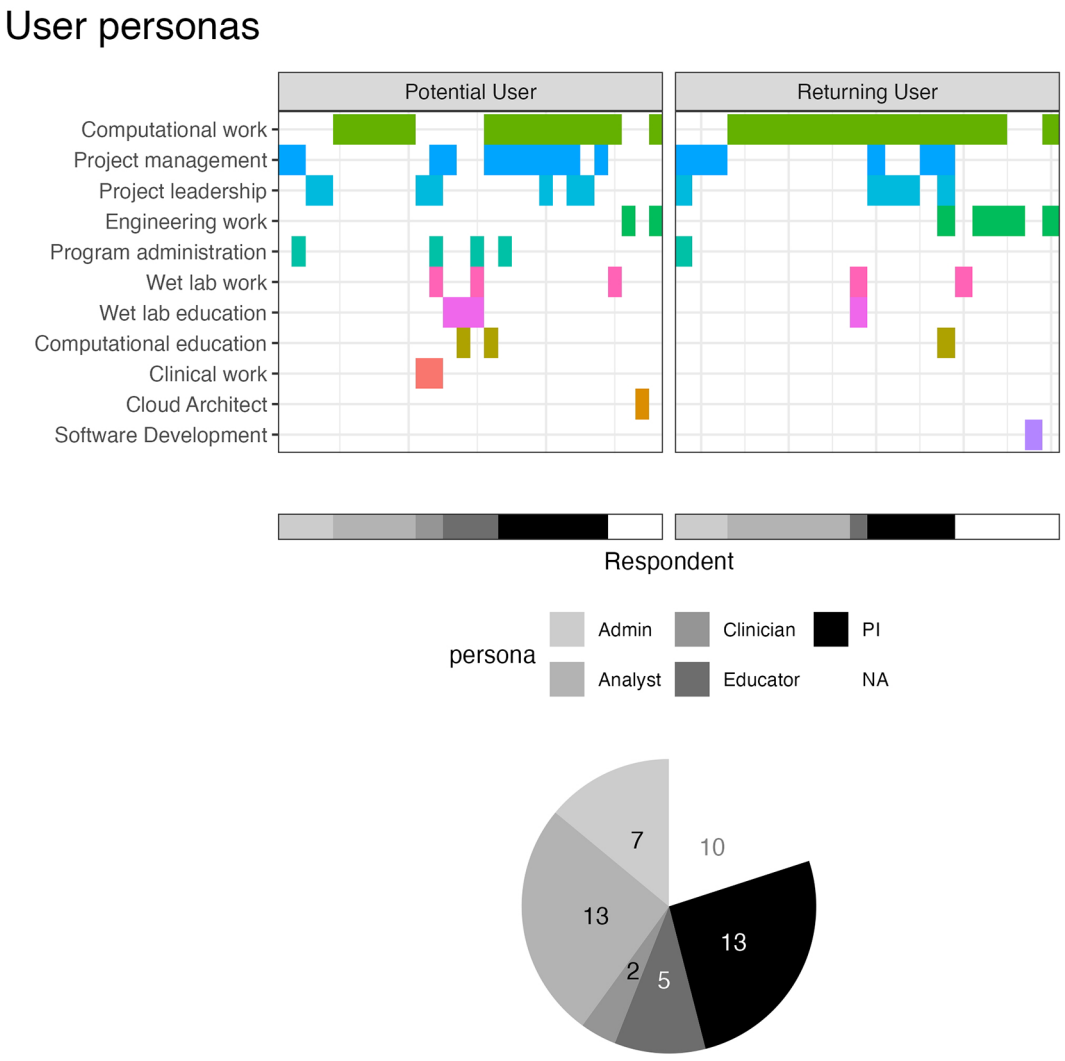

Figure 4. Current work of users related to appropriate personas. We asked all respondents "What kind of work do you do?" Possible selections (computational work, computational education, project management, etc.) are shown on the y-axis and cells are colored for each choice a respondent (x-axis) selected. Based on selections, respondents were clustered into Admin, Analyst, Clinician, Educator, PI, and not assigned (NA) personas. These assignments are shown in grayscale along the x-axis with a corresponding pie chart showing the relative abundance of these assignments. 2 potential users were assigned "Clinician" personas and 4 were assigned "Educator" personas, compared to 0 and 1 respectively for returning users. The other personas show similar abundances between potential and returning users.

As for respondent preferences regarding training modalities/locations, we see that the vast majority of responses in our sample rank virtual options above all other modalities (Figure 5B). Respondents viewed conferences, on-site at their institution, or AnVIL specific events with similar preference ranks. These observations were consistent across user types. At the time of this poll, the AnVIL had offered training through every represented modality option except for an AnVIL-specific event (e.g, the AnVIL Community Conference whose inaugural event occurred after the poll closed).

The AnVIL already supports several virtual training opportunities including monthly AnVIL Demos and a 24/7 online support forum (help.anvilproject.org). The monthly AnVIL Demos are virtual meetings hosted over zoom where typically the first 30 minutes are used for a demonstration highlighting what is possible on the AnVIL and the last 30 minutes are reserved for questions and community discussion. The AnVIL Demos are advertised through the AnVIL mailing list, on the events tab of the AnVIL portal, and described on the support forum. Recordings are posted to YouTube and listed as part of the AnVIL Collection. The 24/7 online support forum provides a place for users to submit questions or read through/reply to threads of questions regarding the platform. The AnVIL outreach team responds to questions and

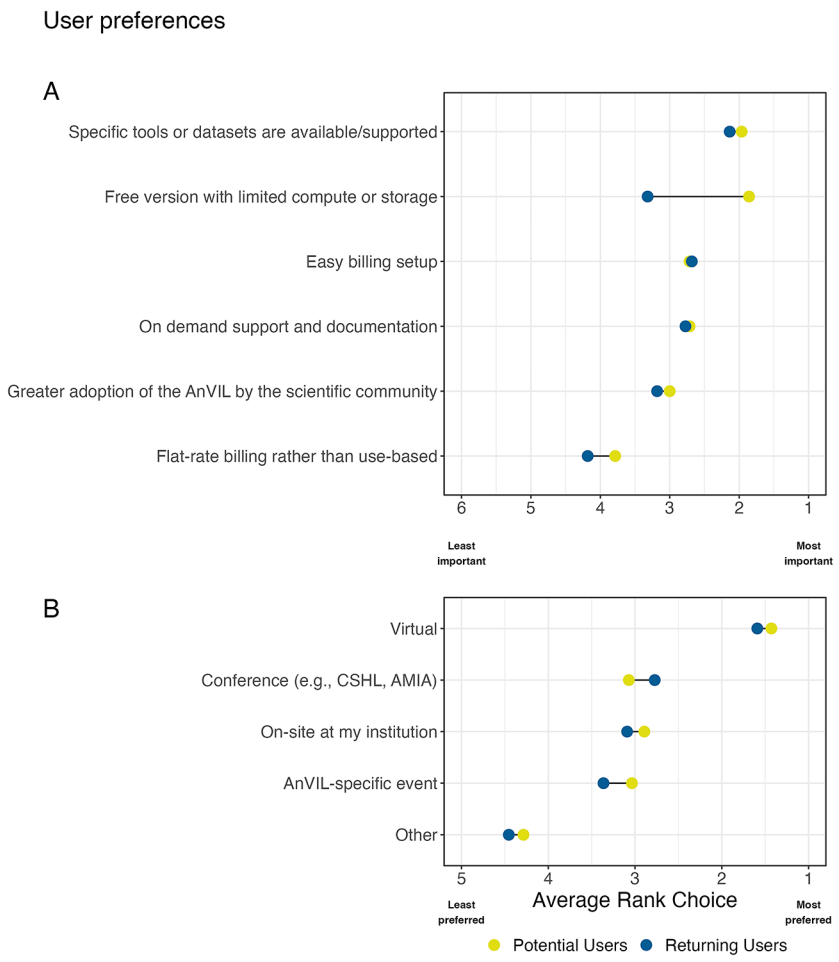

Figure 5. Barriers to platform adoption and user preferences for training and support in our sample. A: We asked all users to "Rank the following features according to their importance to you as a potential user or for your continued use of the AnVIL." Responses were averaged within potential and returning user cohorts to find an average rank. All respondents rated having specific tools or datasets supported/available as an important feature for using AnVIL. Compared to returning users, potential users rated having a free-version with limited compute or storage as the most important feature for their potential use of the AnVIL. B: We asked all respondents "Rank how/where you would prefer to attend AnVIL training workshops." Responses were averaged within potential and returning user cohorts to find an average rank. Both returning and potential users preferred virtual training workshops over other modalities.

sorts threads into relevant categories (e.g., data access, feature requests, etc.). When asked about their utilization of these virtual training opportunities, we found that more respondents were aware of these offerings than were not aware. Potential users were split at 46% (n = 13) knowing about monthly AnVIL Demos or the support forum and 54% (n = 15) not being aware of the support opportunities (Figure 6B,D). Interestingly, only 71% (n = 20) of potential users had matching aware/not aware responses for both AnVIL Demos and the support forum. Of the 15 not aware respondents, only 73% (n = 11) were not aware of both AnVIL Demos and the support forum. Similarly, 86% (n = 19) of returning users had matching aware/not aware responses for both AnVIL Demos and the support forum. Only 4 returning user responses were not aware of both the AnVIL Demos or the support forum. Returning users showed a higher percentage of respondents who were aware of the support opportunities (n = 16, 73% and n = 17, 77%) than potential users (Figure 6B, D). Returning users are also more represented in the group of respondents who have attended AnVIL Demos (n = 12 or 71% of attenders) (Figure 6A) or utilized the support forum (n = 8 or 67% of those who have read through, posted, or replied) in this sample. In general, most poll respondents have not attended AnVIL Demos with only 34% (n = 17) reporting attendance (Figure 6A). For those who have attended an AnVIL Demo within this sample, attending more than one demo or just one demo occurs at similar rates (Figure 6A). Most respondents have not utilized the support forum with only 24% (n = 12) in some way reporting use. Reading through others' posts is the most common form of utilization (Figure 6C).

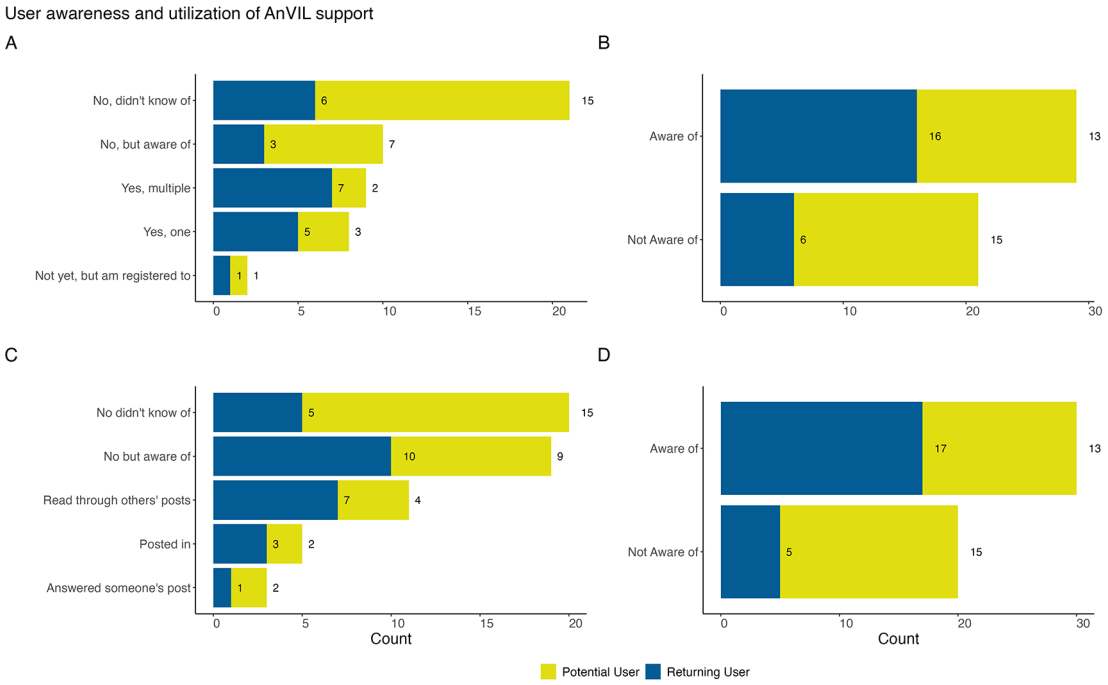

Figure 6. Awareness and utilization of training and support. A & B: We asked all respondents "Have you attended a monthly AnVIL Demo?" A: Most respondents had not attended an AnVIL Demo. However, returning users were more represented among AnVIL Demo attendees. B: All responses to (A) except "No, did not know of" were aggregated, showing that the majority of respondents are aware of AnVIL Demos. C & D: We asked all respondents "Have you ever read or posted in our AnVIL Support Forum?" C: Raw responses are shown (users could select more than one). Most respondents have not used the AnVIL support forum, but utilization in some form is reported by 24% of respondents; reading through others' posts is the most common way of utilizing the support forum within this sample. D: Each set of user responses are recoded and aggregated to examine whether users are or are not aware of the AnVIL Support Forum. We observe that there is awareness of the support forum across potential and returning users.

User technological comfort

The poll asked about comfort with technology in two different sections: within Part 2 (the returning user exclusive questions) and within Part 4 (Experience). Within Part 4, all respondents were asked to rank their knowledge or comfort with certain technologies separate from the AnVIL. Only respondents identified as returning users were asked about these same technologies (excluding specific programming languages) on the AnVIL. Returning AnVIL users were the only respondents asked about knowledge of or comfort with AnVIL specific features such as Data Use Oversight System (DUOS), Terra Data Repository (TDS), Terra Workspaces, or generally accessing controlled access data. Returning users reported higher comfort levels for the various tools and technologies when compared to potential users. The one deviation from this trend is a slight increase in the comfort with Galaxy reported by potential users. Overall, there is less comfort with containers or workflows than various programming languages and integrated development environments (IDEs). While respondents report fairly high comfort with Jupyter Notebooks and RStudio for interactive analysis on the AnVIL, Galaxy on AnVIL is among the technologies with the lowest average knowledge or comfort. Returning users report low levels of knowledge or comfort with TDR and DUOS as well. They report intermediate levels of comfort with accessing controlled access data in general and the highest level of comfort with workspaces.

All respondents in our sample were asked where they currently run analyses. Respondents could select multiple choices. Institutional High Performance Computers (HPCs) (70%, n = 35; 64% of returning users and 75% of potential users) and locally on personal computers (68%, n = 34; 55% of returning users and 79% of potential users) were reported at the highest rates in this sample. For provided cloud options, Google Cloud Platform (GCP)[29] was reported as used the most (n = 21) split across user types (55% of returning users and 32% of potential users) of the AnVIL, compared to 22 returning AnVIL users in our sample). We also observed that potential users report using Galaxy (a free option)[30] (18%, n = 5), more than returning users do (5%, n = 1). Potential users also submitted All of Us,[31] sciserver.org, and UKBB RAP[32] as other platforms they use to perform analyses.

Figure 7. Technological comfort with cloud-based genomic analysis tools. A: We asked respondents "How would you rate your knowledge of or comfort with these technologies or data features?" Except for Galaxy, potential users tended to report lower comfort levels for the various tools and technologies when compared to returning users. Overall, there was less comfort with containers or workflows than using various programming languages and integrated development environments (IDEs). B: We asked all respondents "Where do you currently run analyses?" Institutional HPC and locally run (personal computers) were the most common responses. Google Cloud Platform (GCP) was reported as used more than other cloud providers within this sample. We also saw that potential users reported using Galaxy (a free option) more than returning users do.

User resource and data needs

Returning users (n = 22) were asked about their foreseeable computational needs. The most common response was needing large amounts of storage (e.g., Terabytes), selected by 50% (n = 11); 32% (n = 7) only selected this option. Many nodes, GPUs, and large memory (>192 GB RAM) were also reported as foreseeable needs, but to a lesser extent (selected by 27%, 27%, and 18% respectively). Only 1 response selected all 4 computational needs. The most common combination of selections was GPUs and many nodes (n = 2). All other combinations were only selected once or not at all (though this could be an artifact of a small sample size and 11 possible combinations of multiple options).

46% (n = 23) of all respondents flagged the AnVIL as a repository that they use or are considering use of to share data to comply with the NIH DMS policy. Of those 23, 65% (n = 15) were returning users of the AnVIL (68% of the returning user sample of this community poll).

Over half of all respondents (n = 29) reported that they are "extremely interested" in working with controlled access datasets. When provided choices of specific controlled access datasets that respondents have accessed or are particularly interested in accessing with AnVIL, All of Us[33] and UK Biobank[34] were the most selected (n = 34 each). GTEx was selected by 64% (n = 32) of respondents and CCDG was selected by 40% (n = 20). All of Us, UK Biobank, and GTEx remained the top three when we subset the data to consider (1) only responses from those moderately or extremely experienced with human clinical data, (2) only responses from those moderately or extremely experienced with human genomic data, and (3) only responses from those moderately or extremely experienced with non-human genomic data. CCDG remained the 4th rank selection for those moderately or extremely experienced with non-human genomic data, but was surpassed by eMERGE and HPRC[35] when considering the subsets of researchers experienced with human clinical and human genomic data. Note that poll respondents were informed that All of Us and UK Biobank were not currently available on the AnVIL due to policy restrictions. (Note: Since the results of this poll were analyzed, All of Us is now more accessible to researchers.[36])

When asked about interest in different types of data, respondents selected Genomes/exomes (88%, n = 44) and Transcriptomes (62%, n = 31) as the major types of data they would be interested in analyzing with the AnVIL. A variety of other options (e.g., Phenotypic, Single Cell, Electronic Health Record, Epigenomes, Metabolomes, Proteomes, Imaging, etc.) were selected at a rate of at least 10% (n = 5), but no more than 40% (n = 20). Survey, structural, and variant calling were the only options selected or mentioned by fewer than 10% of respondents.

Discussion

These results together show that while the broad majority of AnVIL users in our sample share commonality in their research interests and needs from the AnVIL platform (e.g., high storage volumes, use as a data repository, ability to access controlled access datasets, etc.), the AnVIL Community maintains a wide range of expertise and research interests. The study found opportunities for platform adoption as well as areas where training should be enhanced to better support the growing community (Table 2, Figure 8).

Our first aim with this poll was to examine the background and current work of users. Given that AnVIL was developed to integrate and analyze large human genomics datasets and other biomedical data,[17] we would expect that users have backgrounds and current interests aligned with human genomics data and research. In general there is a close alignment between what we observe and this expectation. Returning users in our sample did show slightly more experience with human genomic research when compared to non-human genomic and clinical research categories. This could be because AnVIL was designed in partnership with NHGRI, which focuses on human genomics, so respondents are likely to have more experience in that area generally. We also observed that returning users felt more experienced in research overall, which could be a potential source of response bias.

Table 2. Takeaways and next steps for the AnVIL Team. Takeaways from analysis of the poll responses are listed in the first column as the row names. The second column proposes specific examples of next steps that the AnVIL Team can take to address these points.

Takeaways & AnVIL Team Next Steps	
Takeaway	***Proposed Next Step(s)***
Clinical community potential	Feature use cases of working with clinical data within AnVIL
	Host an AnVIL Demo on exporting data from REDCap to AnVIL
Cost hesitancy	Host an AnVIL Demo on estimating, reviewing, and controlling costs
	Add appendix on estimating and reviewing costs to outreach materials
Importance of tools	Highlight existing and create new resources on finding and wrapping tools
Lack of comfort with data resources	Advertise and further develop training on working with the data resources to request and import data
The AnVIL Community is complex	Continue this poll on a recurring basis
	Conduct in-depth interviews with users
	Specifically understand our industry partners

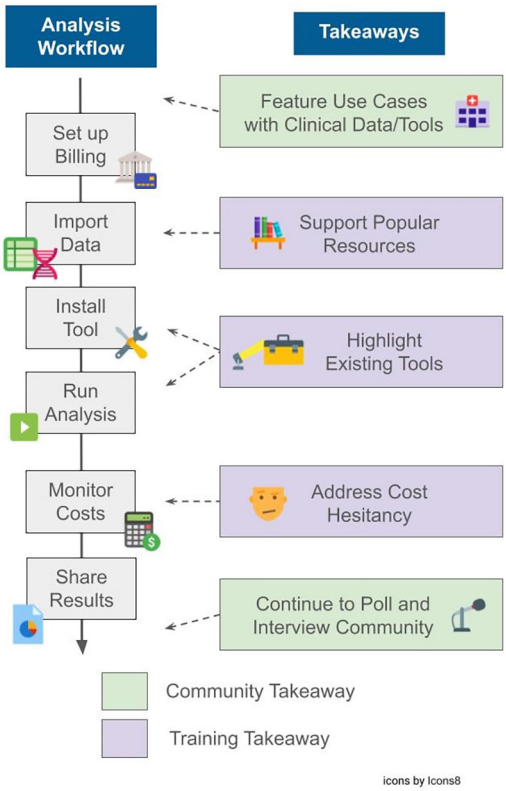

Figure 8. Relating poll takeaways to steps in a typical analysis workflow using the AnVIL platform. Poll takeaways from Table 2 are condensed to combine the takeaway with the proposed step(s) to address it and categorized as training or community takeaways. Training takeaways relate to those with steps the AnVIL Team can take to create new and enhance or highlight existing training materials to address. Community takeaways relate to those meant to converse with, learn from, or grow the AnVIL Community. These takeaways are aligned along a typical analysis workflow that may be performed on the AnVIL platform.

In addition to supporting human genomics research, there is also an interest in the AnVIL Community to increase the usage of the platform by clinicians and clinical researchers for biomedical data and analysis in general. Expanding the AnVIL Community to include clinicians could facilitate interdisciplinary breakthroughs in fields like precision medicine.[37–39] While our sample seems to be predominantly represented by human genomics researchers, we observe some respondents (both returning and potential users) with experience or interest in clinical data and research. For example, we observed (1) respondents with professional degrees, (2) those who selected clinical work as a kind of work they perform, (3) the representation of those who report they are moderately or extremely experienced with human clinical data analysis, (4) the increase in interest among that clinical experience cohort for the clinical dataset Electronic and MEdical Records and Genomics Project (eMERGE), (5) eMERGE consortia affiliations, (6) Electronic Health Record and phenotypic data being data types respondents were moderately interested in analyzing with the AnVIL. This suggests an outreach opportunity, where the AnVIL team could feature use cases of working with clinical data within AnVIL or host a demo of exporting data from REDCap (a clinical data capture and management platform)[40,41] to AnVIL. In addition, the AnVIL Team could partner with relevant consortia to try to train and bring in more of the clinical research community and continue development of features useful to the clinical community like the recent REDCap integration.[42]

Moreover, we aimed to consider backgrounds and research interests in order to develop broad personas for users of the AnVIL. Given the small sample size of the poll, utilizing multiple questions to categorize or cluster respondents was infeasible and we opted for a simplified approach of parsing the responses to the question about kinds of work respondents performed. Personas help the AnVIL Team understand the projects that people are doing and therefore ways in which they may want to utilize the platform. Different types of research activities require different resources and features. In addition, personas and roles that researchers may be identified with are fluid. For example, a specific user may use AnVIL for a research project but also use the platform in order to run a workshop or as a learning environment supporting a class project. Therefore, we utilize personas within this analysis to confirm user groups that we would expect to see and identify potential areas where the AnVIL Team can prioritize outreach, conversations, or feature development.

We were limited both by the small sample size and confirmation bias given the approach taken here[43,44]; therefore we recommend working closely with the community to listen to and understand their roles and needs.

Our second aim with this poll was to identify barriers to platform adoption and user preferences for training and support. Cost seems to be a barrier to platform adoption in our sample as evidenced by potential users ranking offering a free tier as an important feature. The AnVIL provides computing and storage billing in line with the amount of use rather than charging a flat rate for access. No free tier with limited compute or storage is offered. It is possible that potential users are unaware of the billing logistics utilized by the AnVIL platform while returning users would not have selected this feature as important knowing that it is not offered. The AnVIL is working to improve cost transparency and has provided documentation on controlling costs[45] as well as discussions on estimating the costs for interactive analyses[46] and a resource benchmarking how long common workflows take with specific datasets and the associated costs.[47] In addition, the AnVIL is developing a tool to help researchers estimate costs of a workflow.[48] The desire for a free tier might also reflect frustrations with administrative overhead associated with paid platforms, as "easy billing setup" was also ranked highly for returning and potential users.

Both returning and potential users ranked specific tools and datasets being supported as very important for using the AnVIL. AnVIL already supports a variety of analysis solutions and tools and hosts a large amount of relevant data, with more being submitted and released on a regular basis. Galaxy on AnVIL and its Toolshed allow for users to utilize even more tools with specific versions. Due to the lower knowledge/comfort score associated with Galaxy on AnVIL, this could be a potential area to prioritize in developing training materials so that users are empowered to bring their tools to the AnVIL if they are not already/directly supported.

The vast majority of users in our sample preferred virtual training opportunities over the other suggestions. The preference for virtual training could be due to the geographic and institutional spread of the AnVIL Community, funding constraints for travel, or any of the benefits that virtual events offer such as flexibility or increased accessibility for those with disabilities.[49,50] This poll was conducted before the inaugural AnVIL Community Conference and future polls and conversation may show this to be a preferred option as well. The AnVIL Outreach Team supports virtual training opportunities through several modalities (workshops, demonstrations, a 24/7 community support forum, webbooks, YouTube shorts, etc.). Users were largely aware of these support opportunities (namely AnVIL Demos and the support forum), even if utilization was lower. Training opportunities like workshops typically cover associated cloud computing costs to reduce the cost barrier. Handling this barrier for truly asynchronous training opportunities may be another area to prioritize. A solution could include providing estimates for the cloud computing cost of completing the training and an appendix instructing learners how to review incurred costs.

The third aim of this poll was to assess researchers' technological comfort with cloud-based genomic analysis tools. Users appeared to be fairly comfortable with the cloud and running interactive analyses on the cloud, suggesting barriers to adoption could be platform specific rather than reluctance to use cloud tools. Less comfort or knowledge was observed with topics such as containers, workflows, and interactive analysis with Galaxy on AnVIL. In addition the TDR and DUOS systems had a low level of reported comfort. These are all areas where documentation or training could be advertised or further developed.

The final aim of this poll was to identify computational and data analysis resource needs. Several of the findings related to this aim have been discussed already – perhaps the most consequential being AnVIL's support for analysis tools and acting as a data repository where researchers can store and access data. An additional conclusion relates to the future computational needs of AnVIL users. Most returning users reported that they foresee needing large amounts of storage in the future. This is unsurprising due to the vast amounts of data generated in the biomedical research field.[39] The responses and observations within this aim suggest that AnVIL is providing the computational and data analysis resource needs of users whether it be enabling data transfer from high performance computers, providing a variety of data analysis tools, or hosting popular datasets and enabling access as appropriate.

The findings of this poll are a snapshot from a limited set of current and potential AnVIL users. Responses may be biased by a variety of known phenomena, such as central tendency bias[51] when rating their knowledge or comfort with various technologies, or a non-response bias where respondents feel more strongly than those who did not respond.[52–55] In constructing the poll, care was taken to avoid common pitfalls in question design and poll administration that have been found to introduce or intensify bias.[56,57] We do not make statistical comparisons between groups due to the small sample size. While the number of responses we received was lower than expected or preferred, we view this poll and analysis as a worthwhile exercise in data gathering and a starting point where we make broad observations that the AnVIL Team then brings into conversations going forward. The data collection and analysis have helped the AnVIL Team to learn more

about the AnVIL user base, previewing their needs and preferences. This in turn benefits researchers and clinicians who are interested in coming to the AnVIL so that they can see what the community currently looks like, perhaps even enabling them to find like-minded collaborators.

While we did not observe returning AnVIL users working within industry-based institutions in our sample, by considering other sources of user information (e.g., publications referencing the use of AnVIL or username domains), we observe that some users of the AnVIL work with industry-based institutions/companies. It is possible our sample did not capture that user base because of how the poll was disseminated. More work may be needed to understand the work and needs of AnVIL's industry-based partners.

We would like to repeat this poll in the future, tracking how responses change over time and using the responses as a starting point for community discussion and more in-depth user interviews.[58] For future iterations of the poll, some adaptations may be helpful in order to target different user bases (such as data submitters or administrators that are supporting projects but not necessarily running analyses) with specialized questions/poll subparts rather than using a single set of questions. Future iterations or user interviews should additionally explore why users are motivated to use the AnVIL. With regards to user interviews, because the AnVIL Community is relatively new and there is little established knowledge about the user base, it is possible that in-depth qualitative data from user interviews will lead to more novel and intricate insights.[59] Though conclusions from user interviews will be even less generalizable than this small-sample poll, and researcher bias may impact how the conversation is steered or what parts of the conversation are highlighted later.

We are committed to transparency in sharing the poll, its analysis, and our findings. Users may be more comfortable or prioritize making requests or voicing a need/opinion if they see others have the same thoughts.[60,61] The AnVIL portal points to places where users can get help, ask questions, or provide feedback: https://anvilproject.org/help. Terra additionally displays active feature requests that can be explored: https://support.terra.bio/hc/en-us/community/topics/360000500452-Active-Feature-Requests. The findings of the poll will be useful in soliciting additional/future feedback from the AnVIL Community and will inform future platform development as well as enhance user support strategies. The AnVIL Team hopes to ultimately accelerate the adoption of cloud-based genomic analysis technologies and grow the AnVIL Community.

Data and software availability

Source code available from: https://github.com/fhdsl/SOTA2024_ReportOut

Archived software available from: https://doi.org/10.5281/zenodo.17611423

License: MIT License

The complete poll, deidentified data, supplementary Table 1 (relation of study aims and poll questions), and supplementary Table 2 (raw responses translated to awareness and use) are available at https://github.com/fhdsl/SOTA2024_ReportOut.

Acknowledgements

The authors thank the members of the AnVIL Team for their contributions to and their review of this work.

References

1. Dahlquist JM, Nelson SC, Fullerton SM: **Cloud-based biomedical data storage and analysis for genomic research: Landscape analysis of data governance in emerging NIH-supported platforms.** *HGG Adv.* 2023; **4**: 100196.
Publisher Full Text

2. Sriram V, Conard AM, Rosenberg I, *et al.*: **Addressing biomedical data challenges and opportunities to inform a large-scale data lifecycle for enhanced data sharing, interoperability, analysis, and collaboration across stakeholders.** *Sci. Rep.* 2025; **15**: 6291.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Khan MA, Salah K: **Cloud adoption for e-learning: Survey and future challenges.** *Educ. Inf. Technol.* 2020; **25**: 1417–1438.
Publisher Full Text

4. Krampis K, Wultsch C: **A review of cloud computing bioinformatics solutions for next-gen sequencing data analysis and research.** *Meth Next Gener. Seq.* 2015; **2**: 2.
Publisher Full Text

5. Cole BS, Moore JH: **Eleven quick tips for architecting biomedical informatics workflows with cloud computing.** *PLoS Comput. Biol.* 2018; **14**: e1005994.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Banimfreg BH: **A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics.** *Healthcare Analytics.* 2023; **3**: 100190.
Publisher Full Text

7. Howison J, Deelman E, McLennan MJ, *et al.*: **Understanding the scientific software ecosystem and its impact: Current and future measures.** *Res. Eval.* 2015; **24**: 454–470.
Publisher Full Text

8. Barone L, Williams J, Micklos D: **Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators.** *PLoS Comput. Biol.* 2017; **13**: e1005755.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Treloar A, Crott E: **ARDC survey on cloud services usage in research.** [cited 10 Sept 2025].
Publisher Full Text

10. Afiaz A, Ivanov AA, Chamberlin J, *et al.*: **Best practices to evaluate the impact of biomedical research software-metric collection beyond citations.** *Bioinformatics.* 2024; **40**: 40.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Jankowski H, Wernert J, Hancock DY, *et al.*: **Jetstream Annual User Survey - 2020 Summary Report.** 2020 [cited 10 Sept 2025].
Reference Source

12. Wernert JA, Miles TA, DeStefano L: **2022 annual user satisfaction survey evaluation report.** 2022 [cited 10 Sept 2025].
Reference Source

13. Lau JW, Lehnert E, Sethi A, *et al.*: **The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research.** *Cancer Res.* 2017; **77**: e3–e6.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Hiltemann S, Rasche H, Gallardo-Alba C, *et al.*: **Galaxy Training: a Powerful Framework for Teaching!.** 2023.

15. Galaxy Community Conferences (GCCs): [cited 26 Aug 2025].
Reference Source

16. **The State of the Workflow 2023: Community Survey Results.** [cited 26 Aug 2025].
Reference Source

17. Schatz MC, Philippakis AA, Afgan E, *et al.*: **Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space.** *Cell Genom.* 2022; **2**: 100085.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Dai L, Gao X, Guo Y, *et al.*: **Bioinformatics clouds for big data manipulation.** *Biol. Direct.* 2012; **7**: 43; discussion 43.
Publisher Full Text

19. Fischer J, Beck BW, Sudarshan S, *et al.*: **Methodologies and practices for adoption of a novel national research environment.** *Proceedings of the Practice and Experience on Advanced Research Computing* New York, NY, USA: ACM; 2018; pp. 1–7.

20. Hancock DY, Fischer J, Lowe JM, *et al.*: **Jetstream2: Accelerating cloud computing via Jetstream.** *Practice and Experience in Advanced Research Computing.* New York, NY, USA: ACM; 2021.
Publisher Full Text

21. **How To Use A Top 2 Box Score In Your Survey Analysis.** *SurveyMonkey.* [cited 25 Aug 2025].
Reference Source

22. GREGoR Consortium: [cited 26 Aug 2025].
Reference Source

23. Dawood M, Heavner B, Wheeler MM, *et al.*: **GREGoR: accelerating genomics for rare diseases.** *Nature.* 2025; **647**: 331–342.
PubMed Abstract | Publisher Full Text

24. Kullo IJ, Conomos MP, Nelson SC, *et al.*: **The PRIMED Consortium: Reducing disparities in polygenic risk assessment.** *Am. J. Hum. Genet.* 2024; **111**: 2594–2606.
PubMed Abstract | Publisher Full Text | Free Full Text

25. PRIMED Consortium: [cited 26 Aug 2025].
Reference Source

26. welcome to eMerge > Collaborate: [cited 26 Aug 2025].
Reference Source

27. **Centers for Common Disease Genomics.** [cited 26 Aug 2025].
Reference Source

28. GTEx Portal: [cited 26 Aug 2025].
Reference Source

29. **Cloud Computing Services.** *Google Cloud.* [cited 26 Aug 2025].
Reference Source

30. Galaxy: [cited 26 Aug 2025].
Reference Source

31. Researcher Workbench: [cited 26 Aug 2025].
Reference Source

32. Website: [cited 26 Aug 2025].
Reference Source

33. All of Us Research Hub: [cited 26 Aug 2025].
Reference Source

34. Website: [cited 26 Aug 2025].
Reference Source

35. **Human Pangenome Reference Consortium.** *Human Pangenome Reference Consortium.* [cited 26 Aug 2025].
Reference Source

36. Sheets B: **Introducing the All of Us + AnVIL Imputation Service.** *Terra.* 25 Aug 2025 [cited 26 Aug 2025].
Reference Source

37. Goetz LH, Schork NJ: **Personalized medicine: motivation, challenges, and progress.** *Fertil. Steril.* 2018; **109**: 952–963.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Kessler C: **Genomics and Precision Medicine: Implications for Critical Care.** *AACN Adv. Crit. Care.* 2018; **29**: 28–35.
Publisher Full Text

39. Cremin CJ, Dash S, Huang X: **Big data: Historic advances and emerging trends in biomedical research.** *Curr. Res. Biotechnol.* 2022; **4**: 138–151.
Publisher Full Text

40. Harris PA, Taylor R, Thielke R, *et al.*: **Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support.** *J. Biomed. Inform.* 2009; **42**: 377–381.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Harris PA, Taylor R, Minor BL, *et al.*: **The REDCap consortium: Building an international community of software platform partners.** *J. Biomed. Inform.* 2019; **95**: 103208.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Website: [cited 26 Aug 2025].
Reference Source

43. Pohl R: *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory.* Psychology Press; 2004.

44. Confirmation bias: [cited 11 Sept 2025].
Reference Source

45. Understanding Cloud Costs: [cited 26 Aug 2025].
Reference Source

46. Website: [cited 26 Aug 2025].
Reference Source

47. Website: [cited 26 Aug 2025].
Reference Source

48. Whitaker-Allen N: **Biological Data Science 2024.** [cited 26 Aug 2025].
Reference Source

49. Massengale LR, Vasquez E: **Assessing accessibility: Are online courses better than face-to-face instruction at providing access to course content for students with disabilities?** *J. Scholarsh. Teach. Learn.* 2016; **16**: 69–79.
Publisher Full Text

50. Guetter CR, Altieri MS, Henry MCW, *et al.*: **In-person vs. virtual conferences: Lessons learned and how to take advantage of the best of both worlds.** *Am. J. Surg.* 2022; **224**: 1334–1336.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Douven I: **A Bayesian perspective on Likert scales and central tendency.** *Psychon. Bull. Rev.* 2018; **25**: 1203–1211.
PubMed Abstract | Publisher Full Text

52. Scheaf DJ, Loignon AC, Webb JW, *et al.*: **Nonresponse bias in survey-based entrepreneurship research: A review, investigation, and recommendations.** *Strateg. Entrep. J.* 2023; **17**: 291–321.
Publisher Full Text

53. Sedgwick P: **Non-response bias versus response bias.** *BMJ.* 2014; **348**: g2573–g2573.
Publisher Full Text

54. **Nonresponse Bias: What It Is And How To Avoid It.** *SurveyMonkey.* [cited 26 Aug 2025].
Reference Source

55. Self-Selection Bias: *Encyclopedia of Survey Research Methods. 2455 Teller Road, Thousand Oaks California 91320 United States of America.* Sage Publications, Inc.; 2008.
Publisher Full Text

56. Choi BCK, Pak AWP: **A catalog of biases in questionnaires.** *Prev. Chronic Dis.* 2005; **2**: A13.
PubMed Abstract

57. Hendra R, Hill A: **Rethinking Response Rates: New Evidence of Little Relationship Between Survey Response Rates and Nonresponse Bias.** *Eval. Rev.* 2019; **43**: 307–330.
PubMed Abstract | Publisher Full Text

58. Cresswell K, Domínguez Hernández A, Williams R, *et al.*: **Key Challenges and Opportunities for Cloud Technology in Health Care: Semistructured Interview Study.** *JMIR Hum. Factors.* 2022; **9**: e31246.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Lim WM: **What is qualitative research? An overview and guidelines.** *Australas. Mark. J. AMJ.* 2024; **33**: 199–229.
Publisher Full Text

60. School Directory: **The gentle science of persuasion, part three: Social proof.** [cited 26 Aug 2025].
Reference Source

61. Venema TAG, Kroese FM, Benjamins JS, *et al.*: **When in Doubt, Follow the Crowd? Responsiveness to Social Proof Nudges in the Absence of Clear Preferences.** *Front. Psychol.* 2020; **11**: 1385.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 07 January 2026

https://doi.org/10.5256/f1000research.191696.r444898

?

Hassan Rehan
Purdue University, West Lafayette, Indiana, USA

This paper presents findings from the 2024 AnVIL Community Poll, which was conducted to better understand who is currently using the NHGRI AnVIL platform, who potential users might be, and what barriers, training needs, and resource gaps exist. The authors analyze responses from 50 participants and use the results to describe user experience levels, awareness of platform features, perceived challenges, and preferred modes of support and training. The work is framed as an exploratory snapshot rather than a definitive assessment, with an emphasis on openness, reproducibility, and informing future platform development and outreach efforts.

Strengths

The study is clearly motivated and addresses a relevant problem for the genomics and cloud research community, particularly as shared platforms like AnVIL continue to grow. The survey itself is thoughtfully designed and closely aligned with the stated goals of the study. Methods are described transparently, and the public availability of both the data and analysis code is a strong point that aligns well with open science practices.

The authors appropriately rely on descriptive statistics and avoid overinterpreting the data, which is especially important given the sample size. The discussion is generally balanced, with clear acknowledgment of limitations and potential sources of bias. Overall, the paper succeeds in providing useful, practical insight rather than attempting to make claims beyond what the data can support.

Weaknesses and Areas for Improvement

Sample size and representativeness

The relatively small and self-selected sample limits how broadly the findings can be generalized. While this is acknowledged in the manuscript, the authors could emphasize more strongly that the results should be viewed as exploratory signals rather than representative trends across the full AnVIL user base.

Persona assignment methodology
The construction of user personas is reasonable and useful for interpretation, but it involves some subjective grouping decisions. A brief discussion of how sensitive the results might be to alternative grouping approaches—or a short justification of why the chosen approach was preferred—would strengthen this section.

Scope of conclusions
Some of the recommendations and takeaways read as platform-level implications. These would be more appropriately framed as hypotheses, observations, or starting points for future investigation rather than direct conclusions drawn from the data.

Industry representation
The limited participation from industry users is noted, but the manuscript could more clearly distinguish between a lack of responses and a lack of interest or engagement from industry. This distinction would help avoid unintended overinterpretation.

Points That Should Be Addressed to Ensure Scientific Soundness

- More clearly reinforce the exploratory nature of the findings throughout the Discussion
- Slightly temper language around broader platform or community-wide implications
- Expand briefly on the limitations and subjectivity involved in persona construction

These points represent clarifications and framing improvements rather than fundamental methodological concerns.

Overall, this is a well-executed and transparent community study that provides a useful baseline view of AnVIL users and their needs. While the results are necessarily limited in scope, the paper offers valuable insight and sets the stage for more comprehensive future surveys and evaluations.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cloud-based biomedical research platforms, applied data analytics in genomics, research infrastructure evaluation, and AI-enabled systems for large-scale scientific computing. I have experience reviewing studies focused on user-centered platform design, reproducibility, and adoption of computational tools in biomedical research environments.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.