AnVIL Community Conference 2026

Abstract Book

August 31 – September 01, 2026

anvilproject.org
help.anvilproject.org
AnVIL on LinkedIn

Table of Contents

KEYNOTE TALKS

Five Years on AnVIL: An Early Adopter's Experience with Consortium Data Sharing

Ben Heavner, University of Washington

Ben Heavner, PhD, is a Senior Research Scientist at the University of Washington's Genetic Analysis Center, where he leads the Data Team for the GREGoR Consortium Data Coordinating Center. Over the past five years, he has helped design and implement GREGoR's approach to consortium data sharing, integration, and release, including its use of AnVIL.

Dr. Heavner brings more than a decade of experience working with NIH data and cloud computing initiatives. At the Genetic Analysis Center, he has held leadership roles with the NHLBI TOPMed Program and NHGRI's PRIMED Consortium, and has contributed to efforts spanning reproducible genomic analysis, data stewardship, cloud-based research platforms, and large-scale biomedical data sharing.

He earned his PhD from Cornell University, where he worked on computational reconstruction of the yeast metabolic network, and completed postdoctoral training at the Institute for Systems Biology.

Based in Seattle, Dr. Heavner is particularly interested in lowering barriers to sharing high-quality datasets and enabling collaborative, reproducible, data-intensive biomedical research.

The GTEx project(s) and AnVIL - protecting and sharing data

Kristin Ardlie, Broad Institute

Kristin Ardlie is an institute scientist at the Broad Institute of MIT and Harvard. She directed the Laboratory Data Analysis and Coordination Center for the GTEx project and more recently for the developmental (dGTEx) and non-human primate (NHP-dGTEx) projects. The GTEx project enrolled close to 1000 donors and produced RNA sequence data from up to 20,000 tissues to characterize the relationship between genetic variation and the regulation of gene expression across multiple human tissues. The current projects aim to understand gene regulatory changes across development, increasingly at the single cell level.

Ardlie earned her B.S. from the Australian National University and her Ph.D. from Princeton University. She completed a fellowship with the Harvard Society of Fellows, and worked as a research fellow at the Whitehead Institute/MIT Center for Genome Research. After spending 6 years at Genomics Collaborative Inc., a commercial research sample repository, she returned to join the Broad Institute.

INVITED TALKS

Tools for rare disease diagnosis: seqr genomic analysis platform, automated reanalysis and federated variant level matching
Heidi Rehm, Broad Institute

Heidi Rehm, a human geneticist and genomic medicine researcher, is co-director of the Program in Medical and Population Genetics and an institute member at the Broad Institute. She is the chief genomics officer in the Department of Medicine at Massachusetts General Hospital (MGH), working to integrate genomics into medical practice with standardized approaches. She is also a professor of pathology at Harvard Medical School and faculty member of the Center for Genomic Medicine at MGH.

As a board-certified laboratory geneticist and chief medical officer of the Broad Clinical Laboratories, she is guiding genomic testing for clinical and clinical research use, defines standards for the interpretation of sequence variants through ClinGen, co-leads the Broad Center for Mendelian Genomics focused on discovering novel rare disease genes and the Matchmaker Exchange to aid in gene discovery. She is a strong advocate and pioneer of open science and data sharing, working to extend these approaches through her role as a vice chair of the Global Alliance for Genomics and Health. Rehm is also a principal investigator of the Broad All of Us Genome Center, supporting the sequencing and omics data generation on a cohort of one million individuals in the U.S. and co-leading gnomAD, the Genome Aggregation Database.

From Cohort to Cloud: Building the Multi-omics for Health and Disease Resource on AnVIL
Jill Moore, University of Massachusetts

I am interested in applying machine learning methods to better understand gene regulation and its role in human disease. Our lab studies how the genome encodes regulatory information and how this information is interpreted across different cellular and disease contexts. We focus on three core areas:

1. Mapping and Classifying Cis-Regulatory Elements
   We develop computational approaches to systematically identify and annotate cis-regulatory elements—such as enhancers, promoters, silencers, and insulators—across diverse tissues and cell types.

2. Decoding Regulatory Function
   We investigate how combinations of regulatory elements control gene expression programs, using integrative genomics and machine learning to uncover the logic of transcriptional regulation in both normal development and cellular identity.

3. Understanding Regulatory Disruption in Disease
   We explore how genetic variation and epigenetic alterations disrupt regulatory element function, with a focus on uncovering mechanisms that contribute to human diseases such as metabolic disorders, immune dysregulation, and cancer.

The Human Pangenome Reference Consortium Release 2: Available in AnVIL!
Benedict Paten, University of California Santa Cruz

Dr. Paten is a professor and the Jack Baskin Chair in Biomolecular Engineering within the Department of Biomolecular Engineering at the University of California, Santa Cruz. He received his Ph.D. in computational biology from the University of Cambridge and the European Molecular Biology Laboratory. Dr. Paten's work is broadly focused on the growing field of computational genomics. He leads the Computational Genomics Lab and sister Computational Genomics Platform groups within the Genomics Institute. He is involved in many large-scale genomics efforts. Through these efforts and many smaller collaborations, he is helping to develop methods to further our ability to read and interpret genomes.

ALS Compute: A Central Repository of Case/Control Whole-Genome Sequencing Data for ALS/FTD Gene Discovery
John Landers, University of Massachusetts

John Landers is an international leader in the field of ALS genetics research. He received his PhD in Molecular Biology from the University of Pennsylvania in 1995. His career has mainly focused on using new and high-throughput technologies for the identification of neurodegenerative disease genes with a strong focus on ALS and FTD, such as whole-genome sequencing, RNA-Seq, and exome capture/sequencing to identify novel causative genes for ALS and FTD. Through his efforts, the Landers Lab has developed extensive collaborations with the leaders of ALS and FTD genetic research. Dr. Landers is also the Research Leader of Project MinE USA, part of an international consortium of 16 countries to whole genome sequence 22,500 samples to identify novel genes contributing to ALS. Dr. Landers is the founder of ALS Compute, a central repository of WGS for case control studies.

Principles, Structures and Implementation of Experiential Education in Computational Genomics
Scott H. Harrison, North Carolina Agricultural & Technical State University

Dr. Scott H. Harrison is a faculty member within the Department of Biology at North Carolina Agricultural and Technical State University. Dr. Harrison has undergraduate degrees in Computer Science and Microbiology and a doctoral degree in Microbiology and Molecular Genetics from Michigan State University, as well as postdoctoral experience and training in areas of bacterial genomics, systematics, and cancer proteomics. Dr. Harrison's research

interests are for conducting analyses of microbial genomics, physiology, and health. This work is becoming increasingly transdisciplinary due to the number of growing cross-connections between these areas of study. Much of Dr. Harrison's work in biological data analysis centers upon evaluations of pathway networks and phenotypic associations, with a goal to increase the speed and frequency by which computational analysis interfaces with experimental and biomedical investigation.

Adoption, governance, and democratizing potential of cloud platforms
Sarah Nelson, University of Washington

I am a Senior Research Scientist and Project Manager in the Genetic Analysis Center at the University of Washington, where I focus on coordination of large-scale NIH genomics research consortia. My training and research in ethical, legal, and social implications (ELSI) of genomics has included consumer genetics and third-party interpretation, appropriate use of race and ancestry in research, and data governance in cloud-based platforms.

Governance of Human Data Sharing for Research: Beyond Controlled Access.
Pilar Ossorio, University of Wisconsin

Dr. Ossorio is Professor of Law and Bioethics at the University of Wisconsin, where she is on the faculties of the Law School and the Department of Medical History and Bioethics at the School of Medicine and Public Health. She is also a faculty member in the UW Masters in Biotechnology Studies program, Program Faculty in the Graduate Program in Population Health, and a Trainer in both the Neuroscience Training Program and the Computation in Biology and Medicine program. She is an Investigator at the UW-affiliated, private, non-profit Morgridge Institute for Research. Throughout her career Dr. Ossorio has participated in numerous advisory committees and boards that aid governments in setting science policy and she has served on or chaired numerous committees and working groups that advise large-scale genome research initiatives.

Dr. Ossorio's research interests focus on ethics and governance of emerging technologies and of research involving human participants. She has published on data sharing in scientific research; the use of race in biomedical and social science research; ethical and regulatory issues in human subjects research; and the regulation and ethics of research in digital spaces. She is also quite interested in novel ethical, regulatory, and policy issues raised by research aimed at moving scientific and engineering findings from the laboratory to the product development and medical/therapeutic applications (translational research).

SUBMITTED TALKS

The Common Fund Data Ecosystem (CFDE) Incubator for Pan-NIH Resources: Enhancing Scientific Interoperability

Julie Jurgens, Broad Institute

The NIH has funded many valuable but disconnected resources, thus limiting scientific innovation. A solution is scientific interoperability: enabling the scientific method by supporting cross-resource search and access of data or artifacts through scientific or biological concepts. Scientific interoperability is supported by a data layer (making data technically reusable) and a science layer (making concepts, evidence, and meaning reusable), which can be connected via knowledge graphs (KGs), gene sets, metadata, provenance, and AI. To better define scientific interoperability and formulate steps toward it, the Common Fund Data Ecosystem (CFDE) Knowledge Center (cfdeknowledge.org) held the CFDE Scientific Interoperability Incubator for Pan-NIH Resources in July 2026. The event had 72 participants from NIH, academia, industry, and AI companies, enabling cross-disciplinary discussion across an interoperability stack comprising the data layer (including protected data), science layer, AI workflows and integration strategies, and breakouts focused on actionable pilots within a federated model. Participants converged on three pilot projects: 1) The NIH Resource Gateway, a trusted, NIH-endorsed, and agent-ready interface enabling researchers to query and reason across NIH resources, 2) A multi-resource Gene Set Registry with connected provenance and metadata, and 3) KG Integration through modular, reusable components enabling cross-graph MCP or agentic queries. We are soliciting volunteers to inventory existing and planned MCP servers, APIs, KGs, and gene sets across NIH resources; define priority use cases; pilot KG integration and Gateway integration for 3-5 resources; and publish draft standards and a roadmap to expand pilots. These pilots have the potential to facilitate AnVIL data discovery and co-analysis with other NIH resources through GA4GH standards, and AnVIL access through federated, computable, semi-automated governance with GA4GH DUO and AnVIL's DUOS implementation. By uniting leaders from the NIH and many of its funded resources, these efforts will create a collaborative forum for improved resource connection and scientific discovery.

Cloud-Native Machine Learning Workflows for RNA-Seq Analysis with Galaxy on AnVIL

Isis Y. Narvaez-Bandera, Moffitt Cancer Center

Cloud computing platforms are transforming large-scale genomic analysis by enabling scalable, reproducible workflows without local infrastructure. Here we demonstrate the capacity of Galaxy on AnVIL to execute multi-tissue, multi-tool machine learning analyses at scale, using age prediction from GTEx v8 gene expression data as a benchmark use case.

GTEx v8 RNA-seq data for 7,255 donors across 21 tissues were accessed directly on AnVIL via Data Repository Service (DRS) URIs, eliminating data transfer overhead and enabling cloud-native data access at scale. All preprocessing, model training, and evaluation were executed entirely within Galaxy on AnVIL, requiring no local compute resources or programming expertise. We benchmarked two complementary Galaxy machine learning tools —TabularLearner and ImageLearner — for binary classification of biological age (young: 20–49 vs. old: 60–79) using all ~54,592 expressed genes as features and a stratified 70/10/20 train-validation-test split applied consistently across all 21 tissues.

Galaxy TabularLearner, which internally evaluates 15 classification models and selects the best by cross-validation AUC, achieved a mean holdout test AUC of 0.799 across 21 tissues (range: 0.545–0.954), with Ridge Classifier selected as the optimal model in 15 of 21 tissues. Galaxy ImageLearner (CAFormer S18 architecture) achieved a mean test AUC of 0.701, with TabularLearner outperforming ImageLearner in 19 of 21 tissues. Running both tools in parallel across 21 tissues demonstrated the scalability of the AnVIL Galaxy environment for systematic multi-tissue ML benchmarking.

This work establishes a fully cloud-native, scalable ML pipeline on AnVIL that requires no local compute and is reproducible through shared Galaxy histories. Our results highlight Galaxy on AnVIL as a powerful platform for high-dimensional genomic machine learning at population scale, providing a reusable framework for transcriptome-based phenotype classification applicable across the AnVIL ecosystem.

GA4GH Standards in AnVIL: Current Use, Future Opportunities, and a Path for Community Input

Angela Page, GA4GH / Broad Institute

The Global Alliance for Genomics and Health (GA4GH) develops technical standards designed to make genomic and health data more interoperable, findable, and accessible across systems and borders. AnVIL already relies on several of these standards — from data access and identity management to variant representation and file formats — though the full scope of that implementation isn't always visible to researchers who use the platform day to day.

This presentation offers a practical overview of where GA4GH standards currently show up in AnVIL and the datasets hosted within it, including tools and APIs built on GA4GH specifications. It also looks ahead: where are the gaps, and which emerging or under-adopted standards could meaningfully improve how AnVIL supports genomic research at scale? Areas worth watching include federated analysis, phenotypic data representation, data use conditions, and the growing role of AI in clinical genomics workflows, all of which are active areas of GA4GH product development.

GA4GH is a community-driven organization, and its work program is shaped by the problems that real research communities bring to it. Driver Projects and Communities of Interest propose the use cases that inform new standards work; without that input, standards risk being designed in the abstract rather than for actual practice. AnVIL's user community has exactly the kind of large-scale, multi-dataset, federated research experience that GA4GH needs to hear from. This presentation will cover how to engage, whether that means flagging a gap, providing feedback on existing products, or proposing a new area of work.

From Data Model to AnVIL Upload: Implementing GREGoR Data Submission in the ICTS Dashboard

Charles Hadley King, University of California, Irvine

Background: Large-scale genomics studies require integration of participant metadata and sequencing data across multiple teams. The GREGoR Consortium Data Model harmonizes these multimodal datasets and supports validated data submission to the NHGRI AnVIL platform. To support the UCI GREGoR workflow we developed the Institute for Clinical and Translational Sciences (ICTS) Dashboard as a functional implementation of the GREGoR Data Model.

Methods: The ICTS Dashboard was implemented as an open-source platform using the version-controlled GREGoR Data Model as the source for database structures, application-programming interfaces (APIs), validation rules, and user interface. The system was designed as modular infrastructure in which schema, database models, serializers, APIs, and interface components are updated as the consortium standards evolve. We extended this architecture with an AnVIL upload-preparation module that creates auditable upload records, initializes the expected GREGoR TSV table set, tracks generated artifacts, and records validation runs.

Results: The ICTS Dashboard now supports collaborative management of the GREGoR Data Model defined tables. At the time of writing it hosts the metadata from 2,706 participants, 1,343 families, 2,722 biospecimen sequences, and 303 genetic findings. The upload workflow prepares tables as TSVs, records validation outcomes, and records upload status. During upload preparation, an initial set of validation tests identify conditions that are not implementable using only JSON Schema validation, including whether assay-specific experiment and alignment records are represented in the generic experiment and aligned tables, whether called-variant records resolve to the appropriate alignment-set records, and whether biospecimen and sequencing records remain connected to the correct participant and family context.

Conclusion: The GREGoR Data Model provides a blueprint for harmonizing rare disease genomics data. The ICTS Dashboard demonstrates how that blueprint can be implemented as functional infrastructure. This enables validation during the initial data capture, biospecimen tracking, and reproducible AnVIL uploads from a local system.

POSTER ABSTRACTS

1. From Batch to Bots: Galaxy on AnVIL

Enis Afgan; Johns Hopkins University

Galaxy on AnVIL brings Galaxy's accessible, reproducible analysis environment to AnVIL's secure cloud ecosystem. Through a familiar web interface, researchers can run established bioinformatics tools, assemble reusable workflows, track complete analysis histories, and share methods and results, without managing command-line software or complex infrastructure. The deployment keeps Galaxy's core services on a small, cost-conscious virtual machine while sending computationally intensive jobs to Google Cloud Batch, which provisions appropriately sized resources on demand and releases them when work is complete. This lightweight architecture provides an easy entry point for interactive analysis while retaining the elasticity needed for larger workloads. For AnVIL users, the result is a practical bridge between governed cloud data and Galaxy's strengths: broad tool availability, transparent provenance, reproducibility, collaboration, and approachable analysis for researchers with varied computational experience.

Recent developments aim to bring GalaxyAI to AnVIL. GalaxyAI introduces an AI-assisted way to interact with Galaxy, helping researchers discover tools, design workflows, troubleshoot analyses, and navigate results through natural-language conversation. This includes the availability of Orbit/Loom, an AI research harness for Galaxy bioinformatics, that connects users to GalaxyAI and their analysis environment. Broader rollout on AnVIL is pending NIH policy guidance on the use of large language models. Two technical paths are available: operating a model locally, which protects data boundaries but is expensive and currently relies on less-capable models; or using a third-party hosted model such as Google Gemini for Government, which offers stronger capabilities but is not currently permitted by NIH policy. Resolving this policy and deployment question will determine how GalaxyAI can be offered securely, responsibly, and effectively to the AnVIL community.

2. The GRADS-4C Learning Module Initiative for Experiential, Inquiry-driven Learning of Computational Genomics and Cloud-Computing

Sade A. Davenport, Charles C. Naney, Uchenna J. Okefi, and Scott H. Harrison; North Carolina Agricultural and Technical State University

The Genomic Research and Data Science Center for Computation and Cloud-Computing (GRADS-4C) is developing an innovative set of educational learning modules with exercises in bioinformatic content, tasks and resources to support computational genomics and cloud-computing learning objectives,. Learning modules include assessment questions and advanced exercises that promote deeper synthesis and application of knowledge. To facilitate further adoption, modules are being adapted for use within learning management systems that are in common use within the higher education community. Featured modules introduce learners to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, RNA-Seq analysis using the Galaxy platform, and computational notebooks utilizing Jupyter, Python, and virtual environments. The initiative for generating learning modules within GRADS-4C overall creates experiential, inquiry-driven learning opportunities encompassing essential concepts and techniques for conducting computational genomic and cloud-computing analyses.

3. Behind the Portal: What Happens Between Your DAR and Your AnVIL Workspace?
Ava Dreiband; ASHG-NHGRI

NHGRI's Analysis, Visualization and Informatics Lab-Space, or AnVIL, helps researchers use scientifically valuable controlled access genomic datasets. Researchers wishing to use controlled access data must submit a Data Access Request (DAR) through dbGaP. Many researchers are concerned that this process will be opaque, slow, or unnecessarily complex. In practice, most DARs are approved on first review, with the timeline between submission and workspace access driven far more by upstream preparation and clarification turnaround than by the review itself.

From the perspective of a DAR pre-reviewer for the NHGRI Data Access Committee (DAC), much of that friction is avoidable. DAR submissions are evaluated before they reach the full committee, with issues in consent alignment, data use justification, and institutional documentation flagged so they an be clarified quickly. Well-prepared submissions move through review quickly, with well-constructed requests avoiding the most common delays. This poster maps the full DAR lifecycle from submitting the request to dbGaP to accessing the data in AnVIL, with average timing annotated at each stage. Visualizing the process this way explicitly highlights where investigators can meaningfully compress their own timelines with a well-formulated first submission. Doing this starts with understanding what the DAC actually reviews for.

The safeguards those criteria enforce are more consequential now than ever. The DAC's review is structured to catch consent-scope mismatches before access is granted, ensuring that approved uses align with what participants originally agreed to. The DAR process sustains a chain of trust: from participants, who consent to specific uses of their data; through researchers, who agree to honor those uses; to the DAC, which verifies that both are aligned before access is granted. This poster demystifies the request process while reinforcing why the guardrails exist, and why well-prepared requests benefit participants, researchers, and the broader scientific community alike.

4. Examining the Landscape of User Recruitment, Interaction, and Retention Metrics for the AnVIL Platform

Kathryn (Kate) Isaac; Fred Hutch Cancer Center

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) aims to meet the needs of the human genomics, biomedical, and clinical research and education communities. The platform currently supports a wide range of data analysis as well as the ability to safely store and access data in compliance with NIH policies. Work on the AnVIL platform can be easily shared to promote reproducible science and collaboration. Recent work by the AnVIL Outreach Team has sought to better understand barriers to platform adoption as well as current user groups and their needs. Here we discuss the metrics that are available to the AnVIL Team in order to explore user journeys: recruitment to the AnVIL platform, interaction with the platform and support resources, as well as the retention of long-term users and how initiatives such as the annual Community Poll further reinforce these metrics. We also discuss gaps in our metrics, brainstorming ways that we can collect new or utilize existing data to answer specific questions about user groups and their longitudinal journeys. By examining these data sources together and strategizing ways to fill in gaps, we hope to build a more comprehensive view of the AnVIL platform user base and how their experiences and needs change over time. Providing a detailed understanding of user experiences is in support of the overall goals to inform future platform development, enhance user support strategies, and ultimately accelerate the adoption of cloud-based genomic analysis technologies.

5. From Controlled Data Access to Controlled AI Execution: Yoda Platform for Clinical and Genomic Inference

John Kitonyo; Broad Institute

As AI methods trained on clinical and genomic data become increasingly central to biomedical research, cloud platforms must support not only controlled-access data and scalable compute, but also governed pathways for model execution. This is critical in environments where privacy and security requirements are as stringent for models as they are for the underlying data. We present Yoda Platform, a beta, single-tenant, API-first system for governed AI inference designed to complement AnVIL-style research environments.

Yoda has been used to deploy 10+ clinical models in production, with over 1.2 million inference jobs successfully completed to date. It supports versioned model registration, review and publish gates, per-model access control, asynchronous run orchestration, and immutable audit trails. The inference path accommodates structured and file-based inputs through schema-defined run submission and object-storage-backed workflows. It also features optional pre- and post-inference processing stages, capable of serving unimodal and multimodal models as needed.

For the AnVIL community, Yoda serves as a governed inference layer that sits alongside existing controlled-access workflows. AnVIL and related services provide the environment for secure data access and cloud-native analysis, while Yoda provides model lifecycle controls, run-time authorization, durable submission metadata, and auditable execution over sensitive clinical and genomic data. This ensures the reproducibility of model invocation is treated with the same rigor as the reproducibility of data analysis.

Building on this production experience, we detail current beta capabilities such as schema-driven input validation, role- and model-level access control, batched asynchronous run tracking, and configurable execution environments. We also outline near-term roadmap work integrating DUOS-style managed access workflows for model use authorization, directly aligning data-access governance with model-access governance. Our platform provides a practical, proven path toward controlled AI execution in AnVIL-native cloud research settings.

6. eMERGE Data to Discovery: The i2b2-AnVIL Clinical Bridge

Jeffrey G. Klann; Harvard Medical School / Massachusetts General Hospital

A major goal of the AnVIL Clinical Resource (ACR) is to make clinical and phenotypic data sets available to genomic researchers on the AnVIL platform, where they can be linked to or used alongside genomic data for advanced analytics. This also means bringing new clinical data tools into AnVIL. One important target is i2b2 (Informatics for Integrating Biology and the Bedside), a widely-used open-source clinical data warehousing and analytics platform. i2b2 has a flexible data model that can support any subject-oriented data, creating a pathway for importing study-specific formats—such as REDCap databases.

To demonstrate this capability, we developed a secure, cloud-based clinical data portal built on i2b2 and integrated it with AnVIL for the eMERGE consortium. The portal transforms eMERGE 4 clinical data—including EHR diagnoses, medications, procedures, survey responses, and genomic indicators such as polygenic risk scores—into the i2b2 common data model. Researchers use i2b2's graphical cohort-finding tool to query these data and export tailored cohorts directly to AnVIL workspaces for downstream analysis, all within a secured Google Cloud Platform environment.

This work introduces three innovations: automated ETL from REDCap using the redcapi2b2 R package; LLM-assisted generation of derived data elements (such as ""high risk for asthma""), which shortens development time for new study variables; and multimodal querying that lets researchers combine phenotypic and genomic data (e.g., Invitae z-scores) in a single interface.

The system is deployed and in active use by eMERGE 4 investigators, supporting self-service cohort identification and export to AnVIL/Terra.

We would like to extend the portal beyond eMERGE to the broader AnVIL community, pending FISMA and CADR regulatory requirements. This would give any AnVIL researcher a self-service path from clinical phenotypes to genomic discovery.

7. Scalable Machine Learning on AnVIL Platforms for Genomic Analysis

Qing Li; NHGRI/NIH

Large-scale genomic studies increasingly rely on computational frameworks capable of efficiently processing heterogeneous multi-omics data and supporting reproducible, scalable machine learning workflows. Recent advances, such as REGENIE, have enabled rapid and robust genome-wide association analyses, allowing researchers to adjust for polygenic risk profiles and uncover novel variants associated with complex diseases. These approaches improve the detection of individual genetic effects and enhance risk prediction, but typically focus on additive models and do not directly interrogate gene–gene interactions.

To address this gap, we developed a cloud-native machine learning pipeline integrating Google BigQuery for large-scale data management, Apache Spark for distributed computing, Hail for genomic data processing, and H2O AutoML for model development and evaluation. This modular architecture unifies data preprocessing, feature engineering, genomic annotation, and machine learning into a portable workflow deployable across cloud research platforms.

The pipeline was first implemented on the NHGRI AnVIL platform to investigate genomic and epigenomic factors associated with childhood malnutrition, leveraging genomic variants, DNA methylation profiles, principal component features, and demographic variables. To assess portability and scalability, the workflow was deployed on the NIH All of Us Researcher Workbench to identify epistatic interactions associated with childhood-onset hypertension (COEH). The study included cohorts of ~800 COEH cases, ~4,500 adolescents and young adults with hypertension, and ancestry- and sex-matched controls. Genomic regions previously reported in GWAS catalogues were prioritized, followed by functional annotation, quality control, and conversion to Hail MatrixTables for distributed analysis. Multiple H2O machine learning algorithms were trained to detect gene–gene interactions, which may help illuminate the underlying biological pathways beyond what polygenic risk models reveal.

This work demonstrates that the AnVIL platform empowers reproducible, scalable, and portable genomic machine learning, advancing large-scale multi-omics analyses and the discovery of complex genetic architectures.

8. Characterizing Ancestry-Related Heterogeneity Between Additive and Recessive GWAS Models for Type 2 Diabetes

Christelle Moise; Belmont High School

Genome-wide association studies (GWAS) have identified numerous loci associated with type 2 diabetes (T2D), but genetic effects may vary across ancestral populations. We integrated publicly available multi-ancestry GWAS summary statistics from Suzuki et al. with an internal recessive GWAS meta-analysis to investigate ancestry-related heterogeneity. Variants were evaluated using heterogeneity analyses and priortized based on recessive association significance, recessive-to-additive effect size ratios, and allele frequency differences across ancestries. Five loci showed strong evidence of ancestry-dependent recessive effects, with ancestry-specific analyses revealing variation in both effect sizes and allele frequencies. These findings emphasize the importance of diverse populations and recessive models for understanding T2D genetics.

9. AnVIL: A Federated Cloud Platform for Scalable, Secure, and Reproducible Genomic Data Science

Stephen Mosher; Johns Hopkins University

Recent years have seen explosive growth in human genomics, single-cell, functional, and clinical datasets that promise transformative insights for human health. These opportunities come with substantial challenges as moving protected datasets to local institutional compute is costly, complex, difficult to secure and govern.

NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) addresses these challenges by inverting the data-sharing model, bringing researchers to secure data and compute. AnVIL is federally approved infrastructure for NHGRI and beyond, hosting 5 PB of open and controlled-access datasets from flagship programs such as GREGoR, PRIMED, eMERGE, HPRC, T2T, CCDG, and CMG. Quarterly releases and integration with DUOS and dbGaP streamline discovery while meeting NIH's latest requirements for secure, identity-verified access to controlled data.

AnVIL provides integrated ecosystems for reproducible, scalable analysis. Terra delivers secure compute for collaborative projects. Dockstore shares reusable, containerized workflows. Galaxy offers a zero-code, end-to-end analysis environment, while Bioconductor supports interactive, programmatic exploration and advanced statistical workflows and seqr enables Mendelian variant interpretation. The AnVIL Data Explorer lets users build metadata-driven cohorts across studies. New services include an All of Us + AnVIL Imputation Service powered by 500,000+ genomes and the AnVIL Clinical Resource, which supports clinical-genomic harmonization through REDCap ingestion, MapDragon vocabulary management, and OMOP transformations to accelerate discovery.

AnVIL has matured into a secure platform used across hundreds of institutions, supporting thousands of users, with hundreds active monthly. The platform offers an expanding workflow catalog and continued scale-ups in Galaxy and Bioconductor. These capabilities have produced measurable scientific impact through consortium analyses and method development. Community activities such as annual AnVIL Community Conferences, the AnVIL Champions Program foster broader adoption and training.

By reducing redundant data movement, providing secure and auditable access, and offering shareable, reusable workflows together with harmonized clinical-genomic infrastructure, AnVIL enables reproducible cross-study analyses at consortium scale and accelerates translation of genomic data into discovery and clinical benefit.

10. Relative Progress of the FAIR Principles in the Growth of Publications in Genomic Studies

Uchenna J. Okefi, Kristen L. Rhinehardt, and Scott H. Harrison; North Carolina Agricultural and Technical State University

Recent advances in genomic studies and publications have been driven by the rapid evolution of cloud computing, metadata, and FAIR (Findable, Accessible, Interoperable, Reusable) principles. FAIR-related requirements, mandates, and policies, together with major NIH-led initiatives including projects such as AnVIL, All of Us, the Cancer Moonshot, STRIDES, and SCHARE, are accelerating some of this overall FAIR progress through cloud-based platforms. These cloud-based platforms, together with FAIR principles, enable scalable data storage, high-performance analysis, improved reproducibility, and global collaboration, creating a powerful ecosystem for genomic research. Despite the documented impact and progress of FAIR principles in genomic studies, evaluations of FAIR maturity consistently show uneven implementation. While findability and accessibility are advancing more rapidly due to policy mandates and the widespread adoption of cloud-based repositories and identifiers, interoperability and reusability lag due to challenges in metadata standardization, inconsistent documentation, and the complexity of integrating diverse datasets. We analyzed the relative progress of the FAIR principles in the growth of publications in genomic studies by retrieving and reviewing publications from PubMed Central (PMC) across the pre-FAIR Era (2000-2015) and the post-FAIR Era (2016-Present). 40 peer-reviewed cancer genomics articles from the 2000–2015 era were compared with 40 peer-reviewed articles from the 2016–Present. Using a Wilcoxon rank-sum test for indicators of the four FAIR principles across pre-FAIR versus post-FAIR articles, our findings showed significant change for the findability, accessibility, and reusability attributes in the post-FAIR era ($P < 0.05$), although the magnitudes of changes varied. Future developments should prioritize improving interoperability and reusability, which continue to lag behind findability and accessibility. Achieving this goal will require adopting standardized metadata schemas, controlled vocabularies, ontologies, and persistent identifier systems across genomic repositories and cloud platforms.

11. Multi-omics Studies of Liver Disease in Hispanic/Latino Populations at Risk for Metabolic-Associated Fatty Liver Disease

Rashedeh Roshani; Vanderbilt University

Metabolic-associated fatty liver disease (MAFLD) is a heterogeneous condition with varying severity and progression, with particularly high prevalence among Hispanic/Latino populations. To better understand the molecular mechanisms underlying liver disease, it is essential to investigate the multi-omics profiles associated with disease progression. This study examined gene, protein, and metabolite expression changes associated with liver phenotypes in the Border Health Research Cohort (BHRC), a cohort of Mexican Americans residing along the U.S.-Mexico border who exhibit high rates of MAFLD. The longitudinal study focused on a subset of 1,000 BHRC participants characterized using transient elastography phenotyping. BHRC participants characterized using transient elastography. Liver stiffness was measured using FibroScan, and hepatic fat was quantified using the controlled attenuation parameter (CAP). Steatosis cases were defined as individuals with CAP values ≥ 288 dB/m, while fibrosis cases were defined as individuals with liver stiffness measurements (LSM) ≥ 8 kPa; controls had LSM <7.5 kPa. We then employed linear regression models, adjusting for covariates such as age, sex, and three principal components (PCs), to identify differentially expressed genes, proteins and metabolites associated with case-control status. To account for multiple testing, we applied a false discovery rate (FDR)-corrected p-value threshold. We identified up to 62 differentially expressed genes, 991 proteins and 120 metabolites, including several signals previously associated with liver and liver disease. Notably, the results included ALDH3A1(FDRfibrosis = $7.1 \times 10^{-}14$) and KRT8(FDRfibrosis = $7.8 \times 10^{-}17$). Pathway enrichment analysis of the metabolomics data identified alanine, aspartate and glutamate metabolism and glycerophospholipid metabolism as the most significantly enriched pathways. These findings provide insight into the functional mechanisms underlying liver disease and advance our understanding of effector gene mapping and the molecular basis of MAFLD.

12. RawMod: A Deep Learning Approach to Modification Agnostic Classification in Nanopore Sequencing

Bhargav Srinivasan; University of Maryland

Nanopore sequencing offers a powerful platform to represent DNA beyond canonical base pairs; including identification of epigenetic modifications. Epigenetic modifications such as 5-methylcytosine (5mC), 5-hydroxymethylcystosine (5hmC), and N6-methyladenine (6mA) play crucial roles in gene regulation, biological signaling, and other downstream biological processes. Profiling and detecting these modifications at the single-base resolution is required to understand a full scope of genomic function across contexts beyond just the canonical base pair space. Existing epigenetic modification tools are limited to a small set of modifications and rely on specialized algorithms for each set of modifications or the use of an unmodified control dataset.

Our goal is to detect any non-standard base modification without the use of a negative control sample. To this end, we present RawMod, a framework for detecting epigenetic modifications in an agnostic way from raw nanopore signals in two key steps. First, our tool constructs a signal pileup from overlapping read segments to identify deviations from expected signal at single-base resolution relative to a reference genome. Second, we apply a machine learning model consisting of a 1D convolutional trunk per-read and two transformer layers between reads to the signal pileup to determine if a queried site is modified or unmodified, regardless of modification type. On a mixed dataset of 4mC, 5mC, 5hmC, 5hmU, and 6mA modifications, our tool differentiates modified and unmodified sites with an AUROC of 0.95. When testing leave one modification out experiments, our model can generalize on an unseen modification type with an AUROC of 0.80 averaged over all five modification types. These results demonstrate that our approach achieves accuracy comparable to state-of-the-art modification-specific tools while being modification agnostic without a control dataset.

13. Functional Interpretation of Intronic Variants: Integrated Empirical and Digital Approaches

Marianna Weener; Ocular Genomics Institute

Background: Intronic variants represent a critical bottleneck in human genetics and Mendelian disease diagnosis. While computational tools like SpliceAI and AlphaGenome have advanced in silico splice disruption predictions, significant limitations remain—particularly regarding deep intronic variants, tissue-specific regulation, and cryptic exon activations. Bridging the gap between digital prediction and biological validation is a translational priority.

Methods & Objectives: We propose an integrated, bidirectional combination of experimental high-throughput empirical splicing assay with iterative in silico model refinement using AlphaGenome and SpliceAI. We systematically scale functional testing using stable cell line reporter assay to generate quantitative splice outcome data for >2,000 intronic variants across diverse cellular contexts. Also, we benchmark existing computational tools against this empirical ground truth to identify systemic algorithmic biases. Discordant data is used to check biology process underneath this difference with further source code update of in silico models, testing if empirical inputs improve predictive accuracy in independent datasets.

This project leverages the AnVIL ecosystem to manage the substantial computational overhead required for bulk in silico variant scoring and machine learning model training. By utilizing AnVIL's cloud-based infrastructure (such as Terra), we seamlessly integrate our high-throughput functional datasets with large-scale, controlled-access genomic and transcriptomic reference data (GTEx retinal data). Furthermore, our refined prediction workflow is containerized and shared via Dockstore/AnVIL, delivering reproducible, cloud-accessible tools to the broader genomic community.

Significance: This work establishes a scalable paradigm for integrating empirical biology with advanced machine learning, driving intronic variant interpretation away from probabilistic estimation and toward experimentally grounded precision medicine.

14. Through.bio: Tracing the Clinical Impact of Basic Genomic Science

Eric Weitz; Broad Institute of MIT and Harvard

As the AnVIL community continues to accelerate genomics research and basic science, a critical challenge remains: how can we systematically track and demonstrate the translational impact of our foundational work? Traditional bibliometrics fail to capture real-world clinical applications, leaving researchers struggling to trace how early discoveries contribute to improved health outcomes.

We introduce Through.bio, a platform designed to solve this problem at scale by visualizing how basic science becomes medicine. Through.bio constructs robust provenance networks that directly link scientific publications to clinical trials and FDA approvals. By processing over 40 million publications and 550,000 clinical studies, the system uncovers previously undocumented pathways between bench research and modern therapeutics or diagnostics.

For AnVIL researchers utilizing shared data and cloud-based analysis, Through.bio offers a powerful method to move beyond basic citation metrics. It empowers scientists to build auditable, evidence-based narratives of their work's translational footprint. Tracing these connections validates the importance of rigorous genomic analysis and provides critical data needed to strengthen grant proposals, inform research evaluation, and justify continued investment in fundamental biomedical science. The platform is available at https://through.bio.

15. A Year of Teaching Genomic Data Science miniCUREs on AnVIL: Triumphs and Challenges on the Cloud

Sayumi York; Notre Dame of Maryland University

Cloud computing platforms like AnVIL have the potential to greatly increase equity in genomic education by allowing users to borrow computational resources beyond their personal devices and access high quality datasets from public databases anywhere. Conversely, we anticipate challenges in monitoring and controlling cloud compute costs in classrooms and difficulties onboarding new instructors onto an unfamiliar platform. Can educators take advantage of AnVIL's features while avoiding financial pitfalls, especially if they have little background in bioinformatics or cloud computing?

Over the past year we have begun to roll out C-MOOR RNA-seq and 16S Human Gut Microbiome miniCUREs (course-based undergraduate experiences) using interactive environments on AnVIL in classrooms across the country. Across all offerings we have taught 190 students to date, and anticipate another 760 students over the next academic year. Our 31 instructors range from highly-experienced genomic data scientists to instructors with no prior experience in R or cloud computing.

We invite the AnVIL community to hear about how the educator experience on AnVIL has evolved over the past year as new features have been rolled out and what work remains to be done. We will share survey data and cost reports that demonstrate the affordability of the platform, what students have created using AnVIL, and how we aim to provide instructors with foundational skills in computational genomics. Come hear what educators have to say in their own words about working on AnVIL!