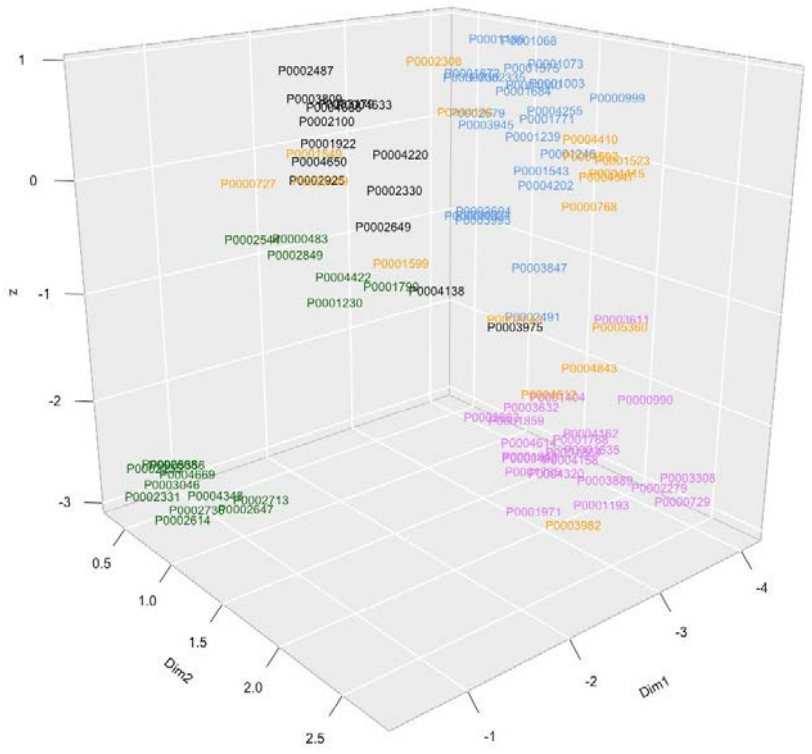
Breakout Session 5:

PII-secured AWS Computing Environment (PACE)- Some lessons learned working with PII in the STRIDES AWS Environment

Dr. Daniel Veltri
Health Scientist (Data Science), NIH/NIAID

PII-secured AWS Computing Environment (PACE)

Some lessons learned working with PII in the STRIDES AWS Environment

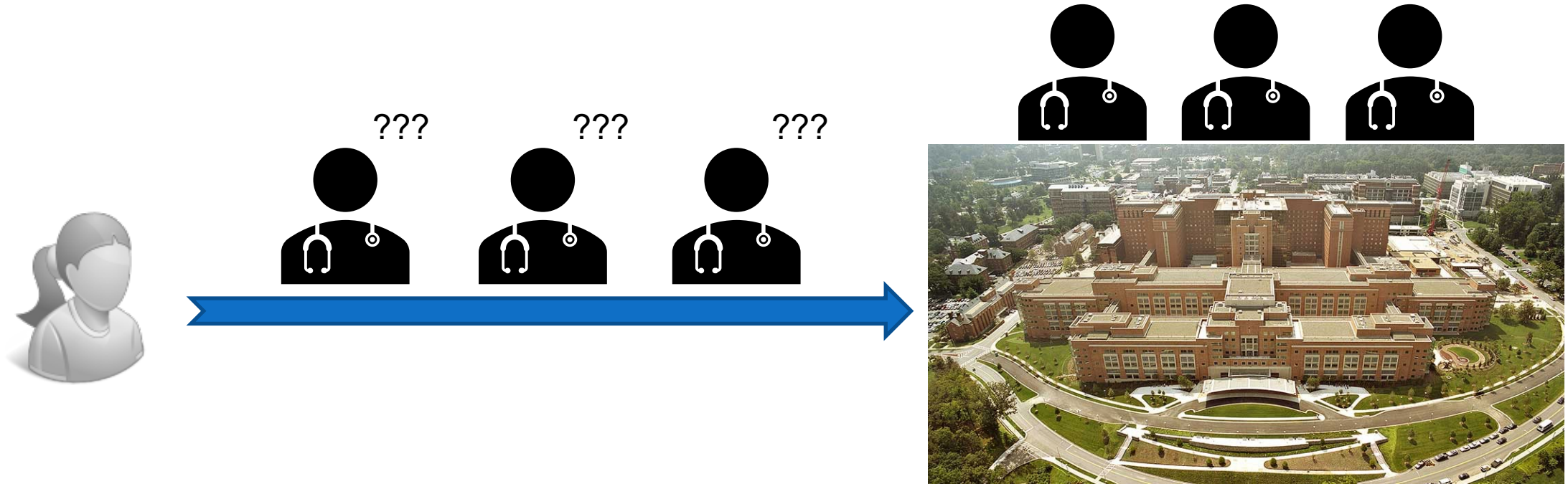


Leiden Clustering of 95 Juvenile Rare Disease Patients

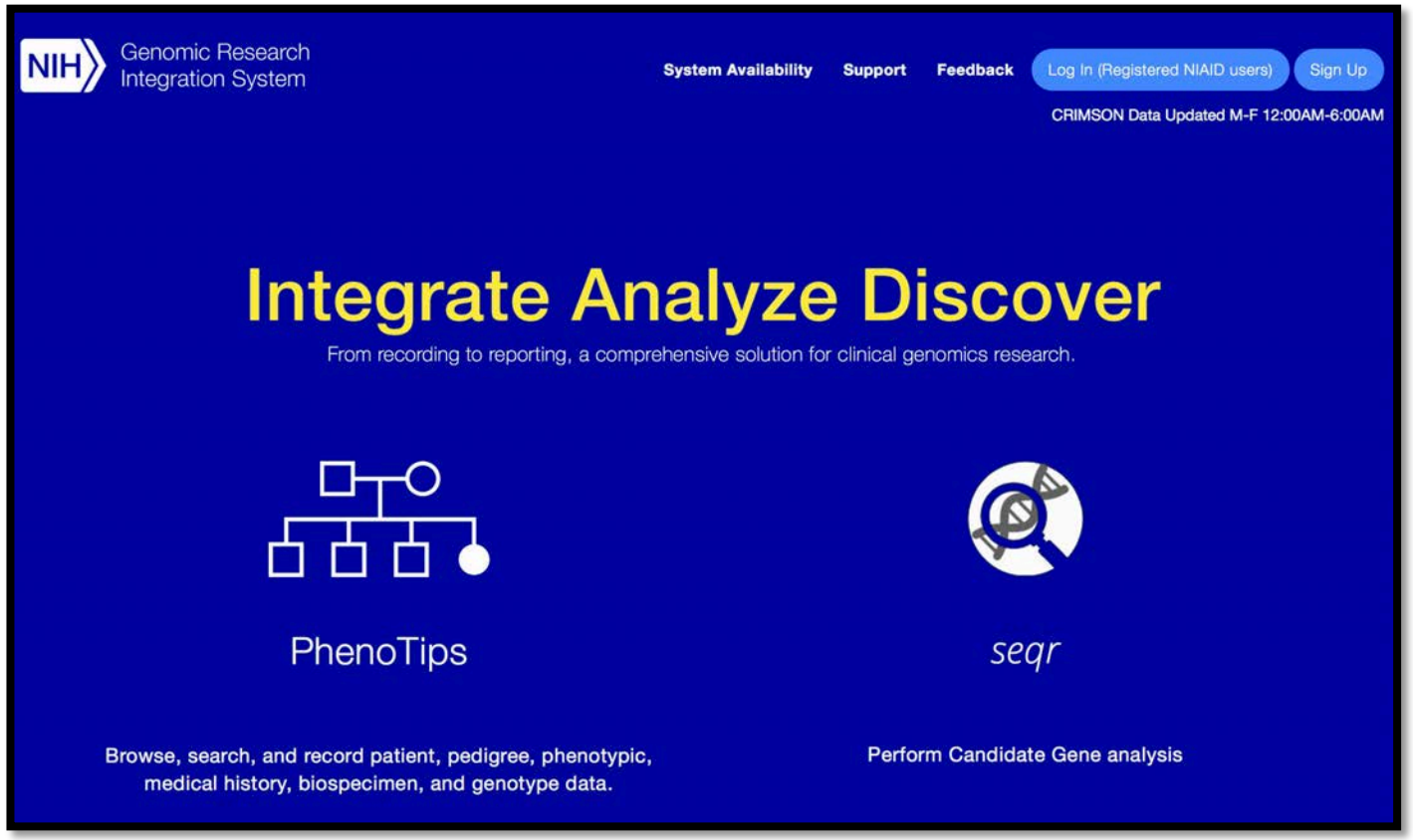
Daniel Veltri, Ph.D.
Bioinformatics and Computational Biosciences Branch (BCBB)
OCICB/OSMO/OD/NIAID/NIH

NIAID

Motivation: Diagnosing patients with rare diseases



NIAID Genomic Research Integration System (GRIS)



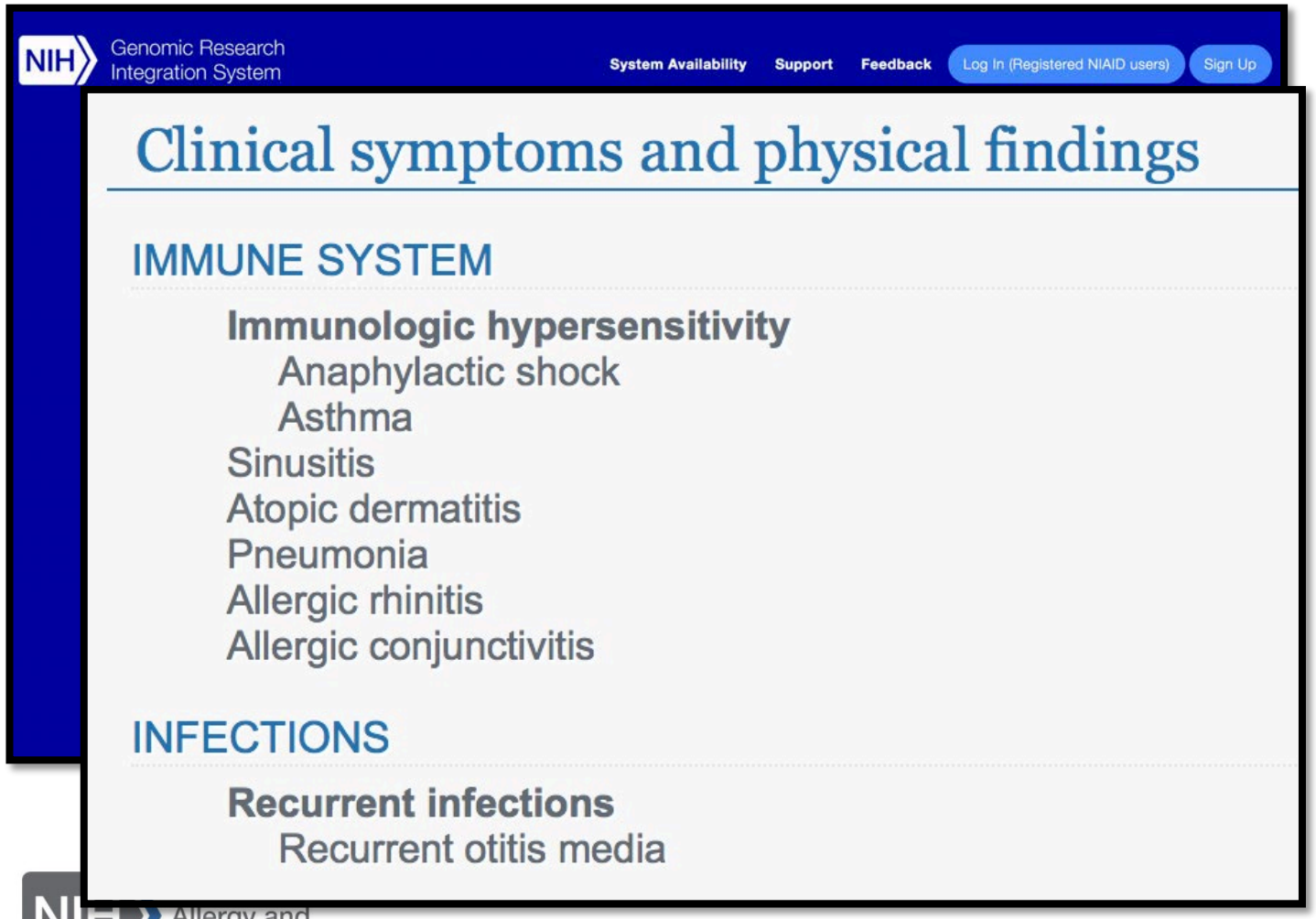
NIAID GRIS Team

NIAID Central Sequencing Program

NIH National Institute of Allergy and Infectious Diseases

NIAID

NIAID Genomic Research Integration System (GRIS)



NIAID GRIS Team

NIAID Central Sequencing Program

NIH Allergy and Infectious Diseases

NIAID

How could Machine Learning (ML) help us?

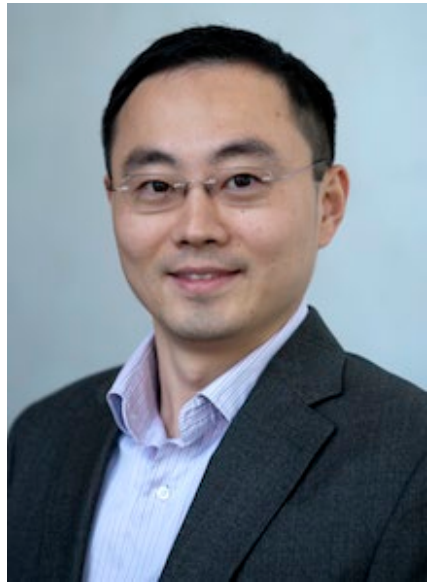
- Promote faster diagnoses for patients
- Reduce time burden on researchers to manually curate symptoms (phenotypes)
- More consistent phenotype curation between labs extracted from a large collection of clinical notes

Working with the Lu Lab at NIH/NCI

- We assisted the Lu Lab in evaluating two of their new natural language processing deep learning models on GRIS patient clinical notes: PhenoTagger[1] and PhenoRerank[2]

- We would like a space to continue such collaborations and development of new tools, but this is complicated by the presence of PII - *major restrictions on most systems!*

- We wanted to leverage ODSS funding and STRIDES to create a secure but collaborative environment for clinical note evaluation
  - *We did it, but it took a long time to get authorized!*



Zhiyong Lu, Ph.D. FACMI FIAHSI
Deputy Director for Literature Search, NCBI
Senior Investigator, NLM

National Institute of Allergy and Infectious Diseases

[1] PhenoTagger: *Bioinformatics* 37(13):1884-1890, 2021
[2] PhenoRerank: *J. of Biomed. Informatics* 129:104059, 2022

NIAID

Challenges obtaining an Authority to Operate (ATO)

- At the time we started, we were the first NIAID project to try out STRIDES, and the first NIAID program to consider PII in the cloud
- STRIDES was still finalizing their AWS environment, security controls, and awaiting their own ATO
- Throughout, we have had to juggle cybersecurity requirements put on us from both CIT/STRIDES and NIAID
- While we inherit a majority of security controls from STRIDES, they do not allow themselves to be considered a “parent system”
  - This was a major conundrum for NIAID Cybersecurity
  - We ended up creating our own PACE security boundary
  - The privacy impact of our system as “Moderate”

ATO challenges continued …

- NIAID required us to use specific software for security logging – the STRIDES team had to come up with a novel way to support it
- Our project also needed role-specific accounts to limit PII access – the STRIDES team had to come up with a novel implementation for this
- Midway, NIAID underwent major changes in cybersecurity leadership
- NIST released a major new cybersecurity controls revision after we had started on the old one

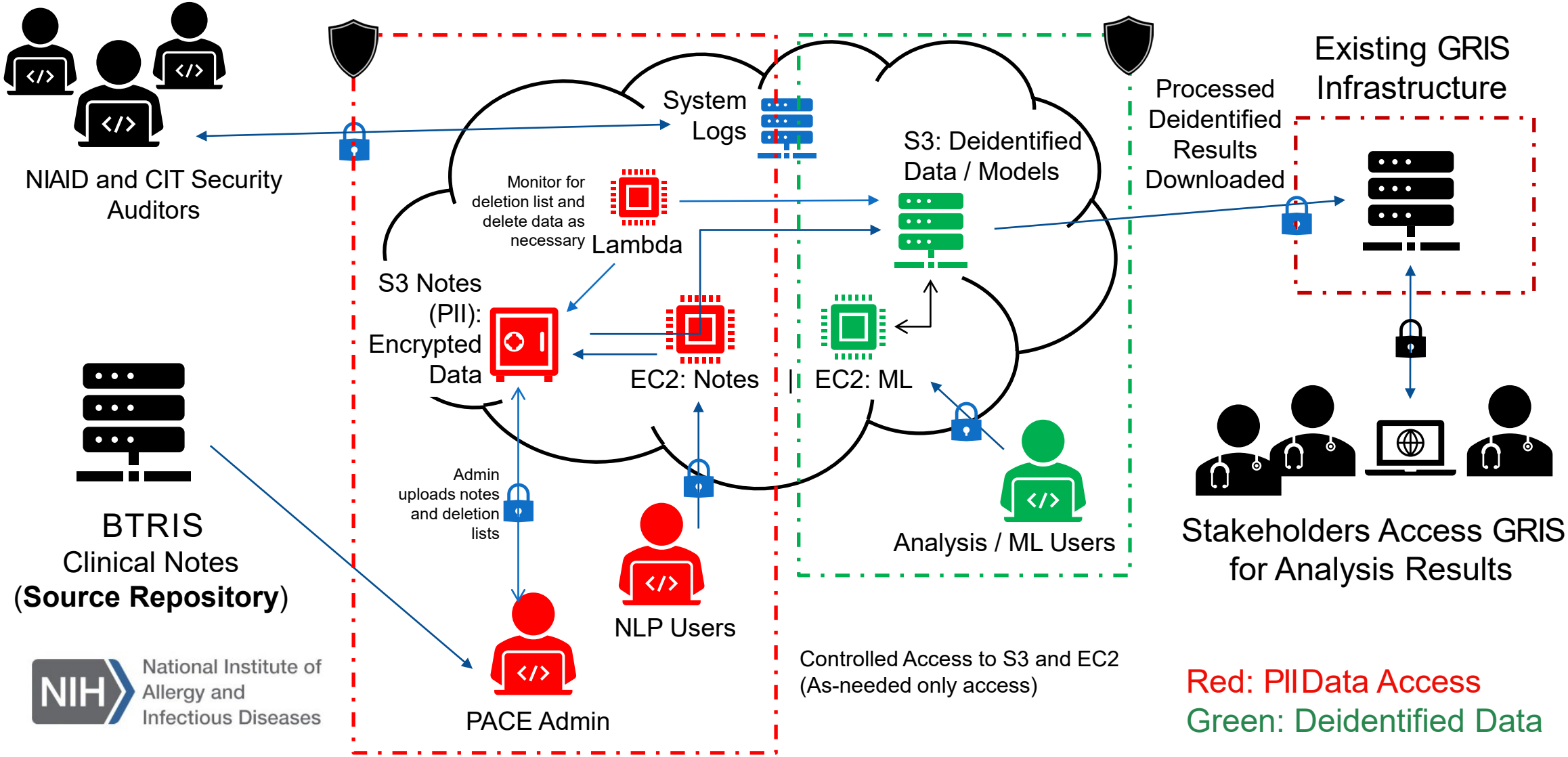
ATO challenges still continued …

- STRIDES does not have an “official” recommended security scanning solution for AWS – NIAID’s current solutions were also not a great fit for us (we ended up getting approved to use AWS Inspector)

- In terms of setting up the environment roles to access resources – both *STRIDES and the AWS Enterprise Support teams were amazing help!*

We did manage to finally get ATO in 2023!

High level look at PACE architecture



National Institute of Allergy and Infectious Diseases

NIAID

Some lessons learned …

- Dealing with PII? Avoid being a first-mover, wait for environments to be established and tested first. *It is complicated enough without PII!*
- Prior to starting – have a rock solid understanding of the *parent system* and the security boundaries you will use. *I should have asked more questions up front*
- Have a cloud expert in house - even for a simple system like this one, the STRIDES and AWS support teams can only do so much for you. *Ultimately, the design, implementation, and maintenance is on you*
- Does your group have other cloud environments or vendors? Think carefully how you can keep synchronized
  - Can you standardize the implementation, security scanning, and reporting procedures?
  - This could be a challenge for PACE moving forward in STRIDES as most of NIAID uses the Monarch Platform

Major Thanks!

Sandhya Xirasagar, Ph.D.
Initial PM for GRIS and huge help to the start of this project

Giovanni Borjas and Metasebia Gizaw
Two incredible ISSOs that helped me navigate the difficult ATO process

Andrew Kulak
Incredible support with AWS environment configuration and general questions for STRIDES

NIAID/OCICB: Dr. Darrell Hurt, Dr. Andrew Oler, Paul Suh (NIAID CISO), and Mike Tartakovsky
CIT/STRIDES: Nick Weber and James Davis
AWS: Tom Fonseca and Gargi Singh
NIAID Central Sequencing Team: Morgan Similuk, ScM and Dr. Rajarshi Ghosh
NLM Lu Lab: Drs. Zhiyong Lu, Ling Luo, and Shankai Yan, Lai Po-Ting, Chih-Hsuan Wei, and Robert Leaman
NIH BTRIS Team: Dr. Michael Ring and Andrea Beri

A big thanks to NIH/ODSS and STRIDES for supporting this project!

Any questions? Email: daniel.veltri@nih.gov

NIH National Institute of Allergy and Infectious Diseases

NIAID

Developing a Scalable and Reusable Framework for State-of-the-Art Structural Variant Calling of Whole Genome Sequencing Data

Andrew Oler, Ph.D.; Eric Karlins, M.S.; and Daniel Veltri, Ph.D.

Bioinformatics and Computational Biosciences Branch (BCBB)

OCICB/OSMO/OD/NIAID/NIH

NIAID

Motivation: Looking for Structural Variants (SV) in genomes

- SVs comprise a range of large genomic rearrangements (> 50bp) and imbalances that can play significant roles in a variety of diseases and important for diagnosing patients

- This project uses NIH/ODSS funding and STRIDES to help run the Broad Institute's state-of-the-art GATK-SV pipeline on the Terra Platform to look for SVs in whole genome samples

- This was pilot initially started for Covid-19 patients but has now expanded to include general GRIS patients whenever whole genome sequencing data is available

Broad Institute GATK-SV Pipeline on Terra

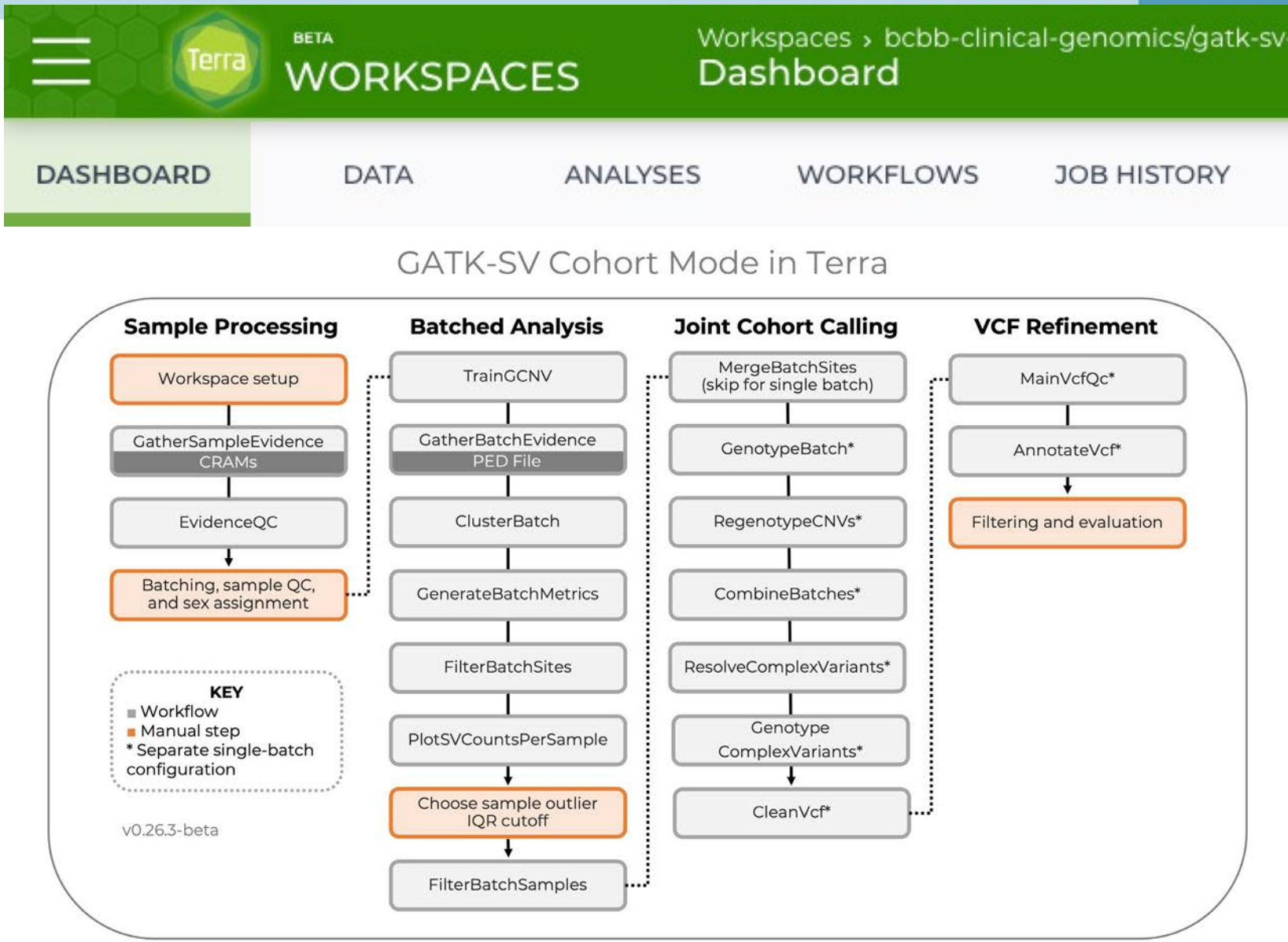
Article

A structural variation reference for medical and population genetics

https://doi.org/10.1038/s41586-020-2287-8
Received: 2 March 2019
Accepted: 31 March 2020
Published online: 27 May 2020
Open access
Check for updates

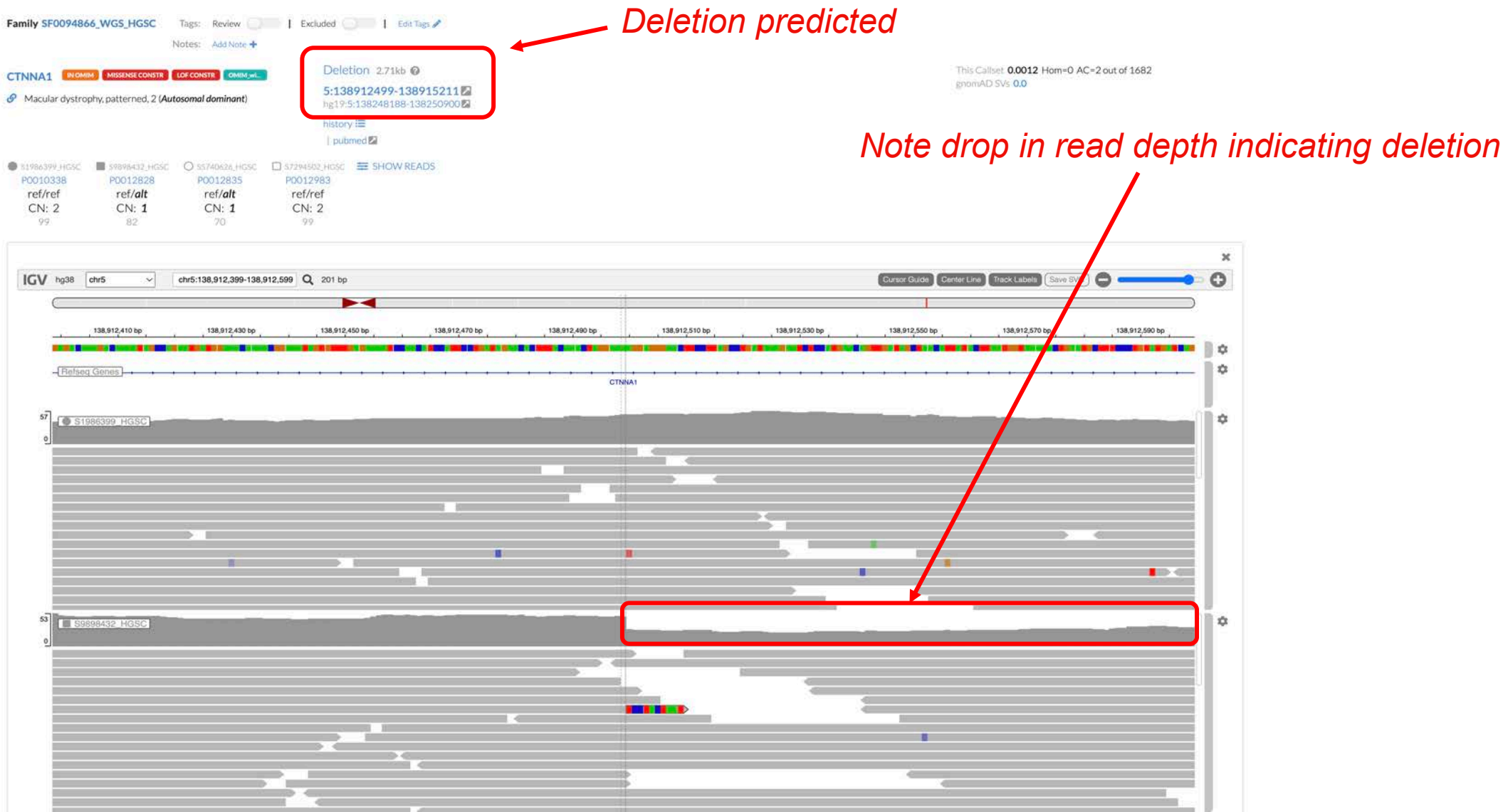
Ryan L. Collins[1,2,3,158], Harrison Brand[1,2,4,158], Konrad J. Karczewski[1,5], Xuefang Zhao[1,2,4], Jessica Alföldi[1,5], Laurent C. Francioli[1,5,6], Amit V. Khera[1,2], Chelsea Lowther[1,2,4], Laura D. Gauthier[1,7], Harold Wang[1,2], Nicholas A. Watts[1,5], Matthew Solomonson[1,5], Anne O'Donnell-Luria[1,5], Alexander Baumann[7], Ruchi Munshi[7], Mark Walker[1,7], Christopher W. Whelan[7], Yongqing Huang[7], Ted Brookings[7], Ted Sharpe[7], Matthew R. Stone[1,2], Elise Valkanas[1,2,3], Jack Fu[1,2,4], Grace Tiao[1,5], Kristen M. Laricchia[1,5], Valentin Ruano-Rubio[7], Christine Stevens[1], Namrata Gupta[1], Caroline Cusick[1], Lauren Margolin[1], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Kent D. Taylor[9], Henry J. Lin[9], Stephen S. Rich[9], Wendy S. Post[10], Yii-Der Ida Chen[9], Jerome I. Rotter[9], Chad Nusbaum[1,155], Anthony Philippakis[7], Eric Lander[1,11,12], Stacey Gabriel[1], Benjamin M. Neale[1,2,5,13], Sekar Kathiresan[1,2,6,14], Mark J. Daly[1,2,5,13], Eric Banks[7], Daniel G. MacArthur[1,2,5,6,156,157] & Michael E. Talkowski[1,2,4,13]

The same pipeline used for gnomAD-SV
https://www.nature.com/articles/s41586-020-2287-8



National Institute of Allergy and Infectious Diseases

Example of an SV result in the Broad *seqr* program



NIAID

Summary

- We have run the pipeline on >550 Covid-19 samples and >5000 GRIS patient samples
- Compute costs are in the $3-5 range per sample
- Customized filtering of the results has been important to adjust sensitivity and specificity for the analysis – finalized protocols are still being developed
- This pipeline has assisted with the successful diagnosis of multiple patients enrolled in the NIAID Central Sequencing Program

Thanks

We would like to thank the NIAID GRIS Team for helping process and implement results and the NIAID Central Sequencing Program for all their help in testing and evaluation. Special thanks to NIH/ODSS and STRIDES for supporting this project.

Questions? Email: gris@mail.nih.gov

National Institute of Allergy and Infectious Diseases

NIAID