

Breakout Session 2: Track B

Application of Genomic Knowledge Standards to the Genome Aggregation Database

Dr. Alex Wagner

Principal Investigator, Nationwide Children's Hospital

Application of Genomic Knowledge Standards to the Genome Aggregation Database



*Supplement to: Development and validation of a computable
knowledge framework for genomic medicine (R35 HG011949)*

Alex Wagner, PhD



NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.™



THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

Genomic Medicine



Patient / Tumor DNA



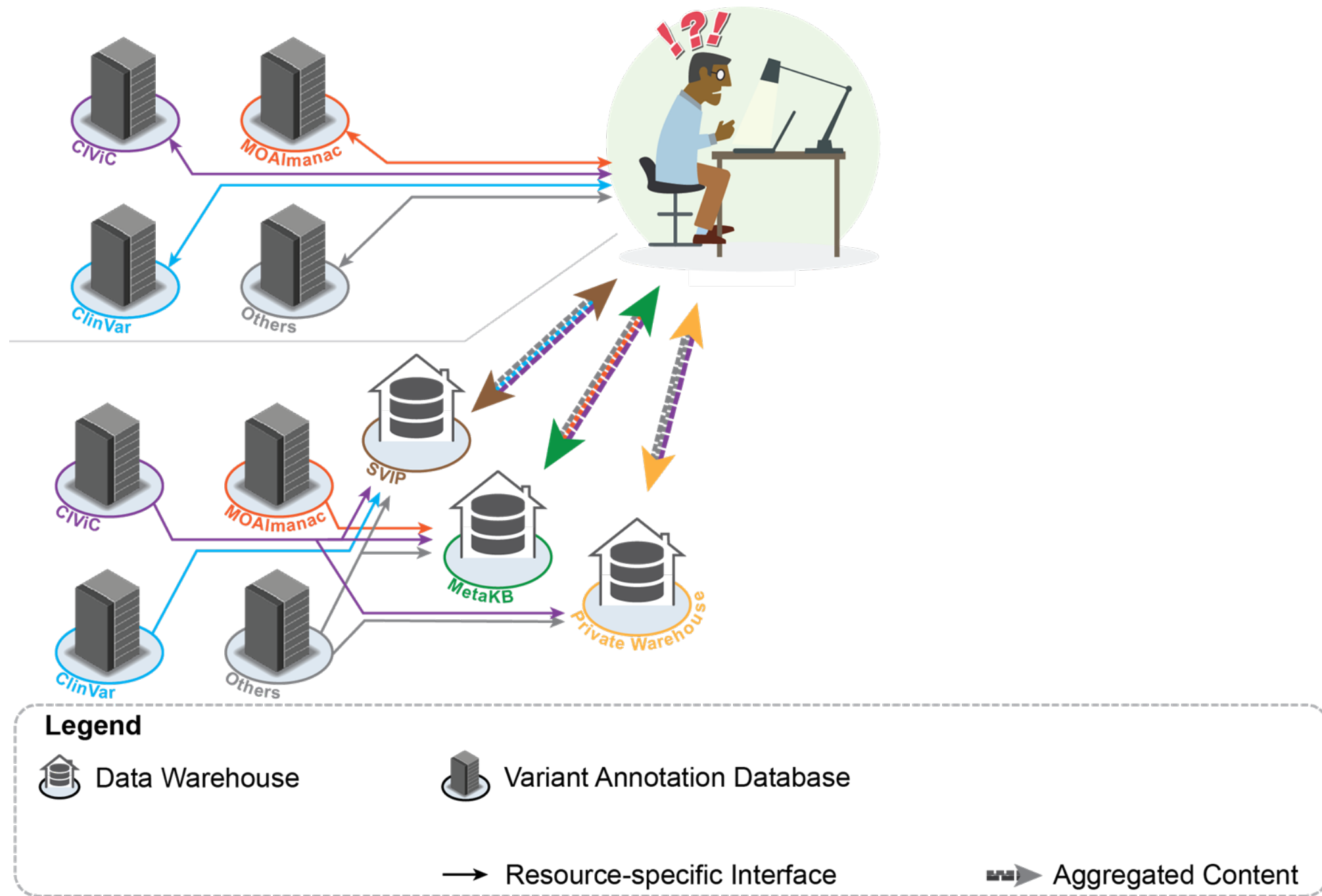
Clinical Genomic Variant Report



Patient Care

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in <i>trans</i> with a dominant variant BP2 Observed in <i>cis</i> with a pathogenic variant BP2		For recessive disorders, detected in <i>trans</i> with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Pathogenic	<ul style="list-style-type: none"> (i) 1 Very strong (PVS1) AND <ul style="list-style-type: none"> (a) ≥1 Strong (PS1–PS4) OR (b) ≥2 Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) AND <ul style="list-style-type: none"> (a) ≥3 Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) AND ≥2 Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5)
Likely pathogenic	<ul style="list-style-type: none"> (i) 1 Very strong (PVS1) AND 1 moderate (PM1–PM6) OR (ii) 1 Strong (PS1–PS4) AND 1–2 moderate (PM1–PM6) OR (iii) 1 Strong (PS1–PS4) AND ≥2 supporting (PP1–PP5) OR (iv) ≥3 Moderate (PM1–PM6) OR (v) 2 Moderate (PM1–PM6) AND ≥2 supporting (PP1–PP5) OR (vi) 1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5)
Benign	<ul style="list-style-type: none"> (i) 1 Stand-alone (BA1) OR (ii) ≥2 Strong (BS1–BS4)
Likely benign	<ul style="list-style-type: none"> (i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) OR (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	<ul style="list-style-type: none"> (i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory



Legend



Data Warehouse



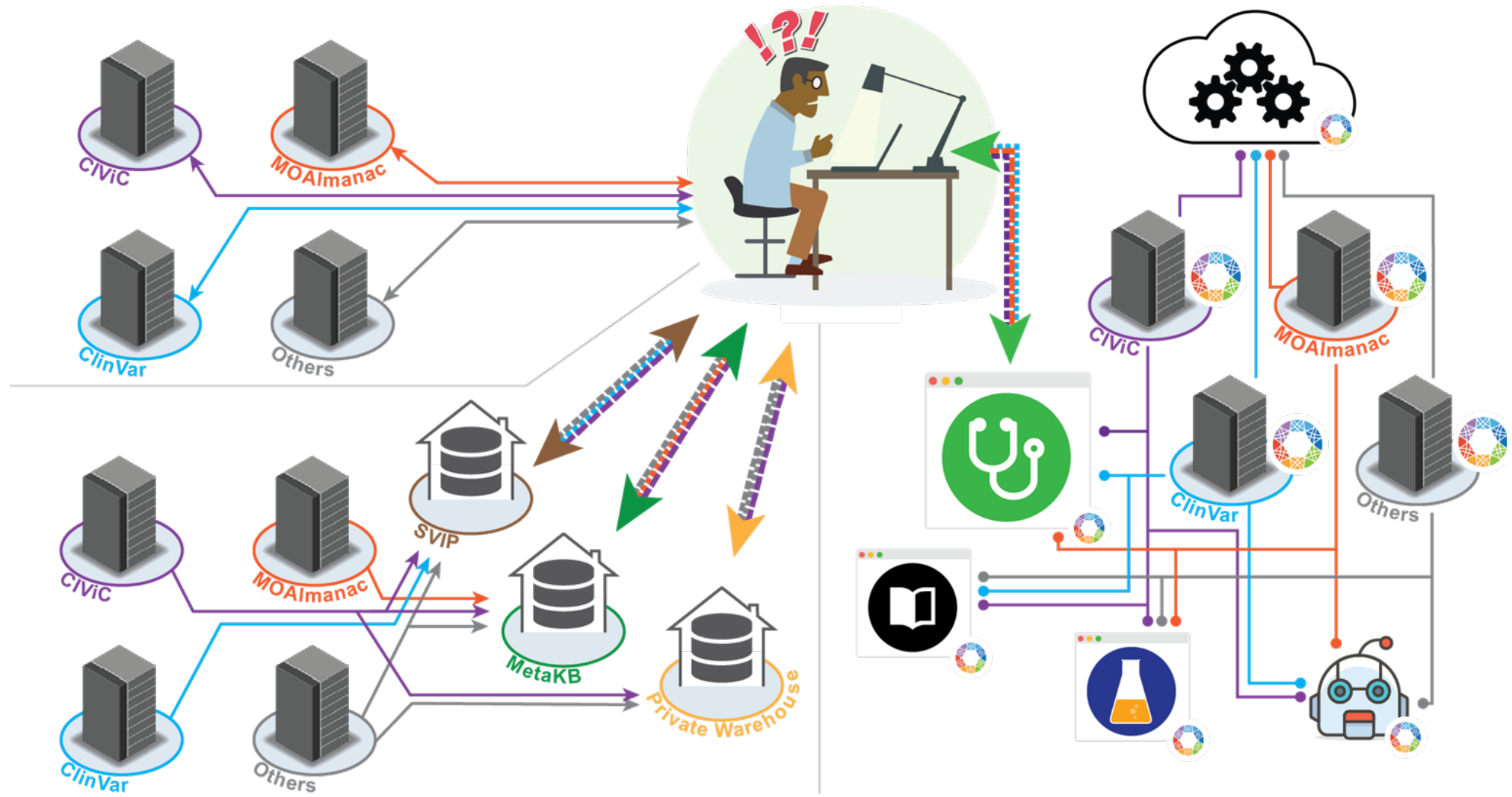
Variant Annotation Database












Resource-specific Interface



Aggregated Content



Legend

-  Data Warehouse
-  Variant Annotation Database
-  Support Tools
-  Web Application
-  Genomic Knowledge Framework
-  Machine Learning
-  Standardized Interface
-  Resource-specific Interface
-  Aggregated Content

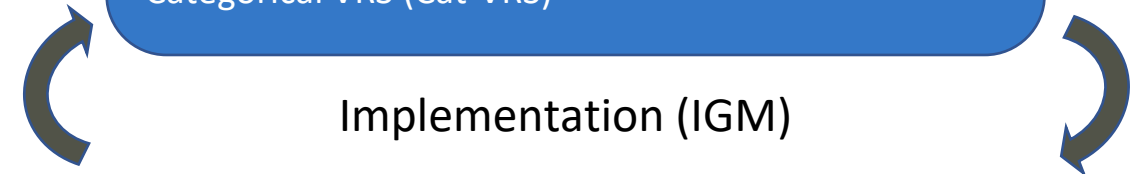
A GA4GH Genomic Knowledge Framework



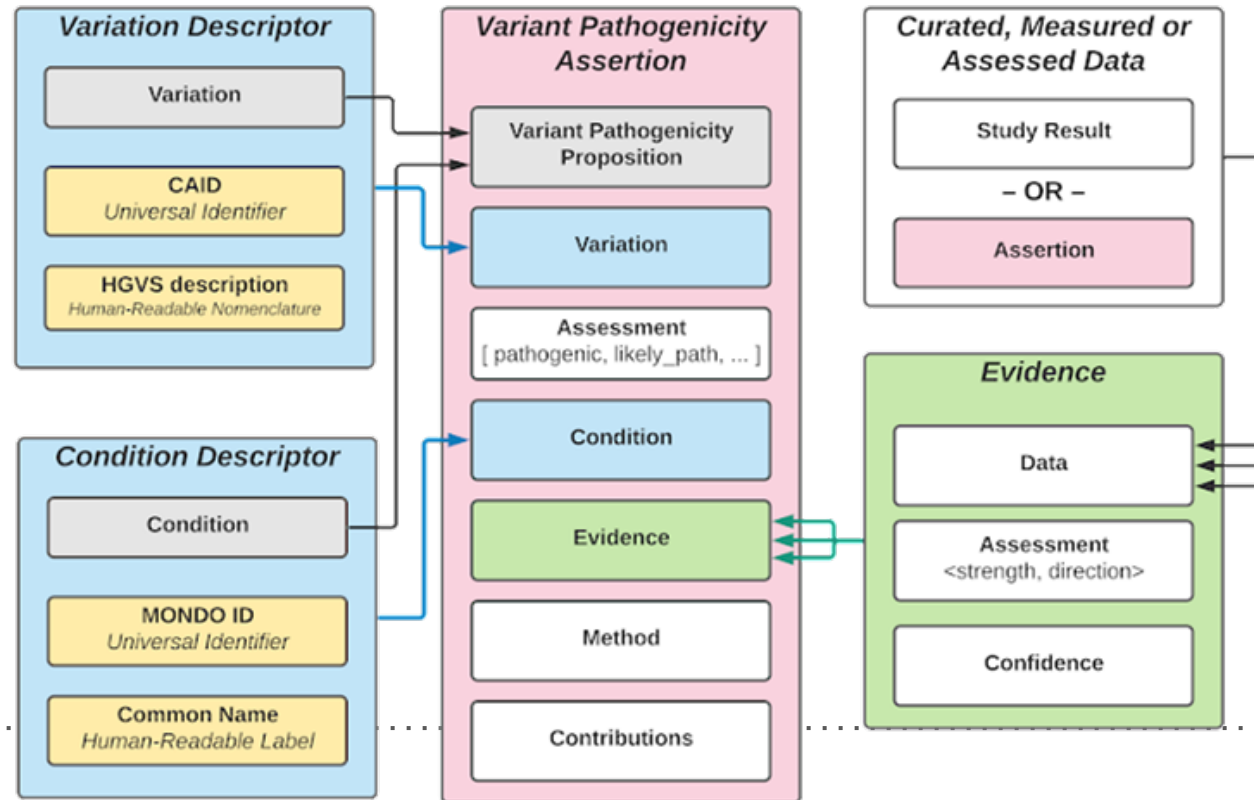
Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Genomic Knowledge
Standards Work Stream

Variation Representation Specification (VRS; "verse")
Variation Annotation Specification
Categorical VRS (Cat-VRS)



Implementation (IGM)



Collaboration with:



NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.™

Supported by [R35 HG011949](#)



Where do gnomAD samples come from?

308 data contributors

>100 studies

>25 countries*

Australia, Bangladesh, Belgium, Canada, China, England, Finland, France, Germany, Israel, Italy, Japan, Kenya, Korea, Lithuania, Mexico, Netherlands, Pakistan, Scotland, Singapore, Spain, Sweden, United Arab Emirates, USA, Wales

*Based on country of the study's institutional review board (IRB)

Contributing projects

1000 Genomes
1958 Birth Cohort
African American Coronary Artery Calcification project (AACAC)
ALSGEN
Alzheimer's Disease Sequencing Project (ADSP)
Atrial Fibrillation Genetics Consortium (AFGen)
Duke Catheterization Genetics (CATHGEN)
Bangladesh Risk of Acute Vascular Events (BRAVE) Study
BIOcd-plus
BioHeart
BioMe Biobank
BioVU
BipEx
Bulgarian Trios
CCDG IBD sequencing project
COPD-Genie
Crohn's & Colitis Foundation (CCFA) Genetics Initiative
ENGAGE-TIMI
Estonian Genome Center, University of Tartu (EGCUT)
Finland-United States Investigation of NIDDM Genetics (FUSION)
Finnish Migraine Study
Finnish Twin Cohort Study
FINN-ADGEN
FINRISK
Framingham Heart Study
Gene Discoveries in Subjects with Crohn's Disease of African Descent
Genetics of Cardiometabolic Health in the Amish
Genizon Biobank
Génome Québec - Genizon Biobank
Genomic Psychiatry Cohort
GoT2D
Genotype-Tissue Expression Project (GTEx)
Health2000
Human Genome Diversity Project
Inflammatory Bowel Disease:

1000IBD project
Helsinki University Hospital Finland
IBD Genomic Medicine Consortium (iGenoMed)
IBD: REMIND
IBD: Understanding the determinants of health outcomes
Inflammatory Bowel Disease Sequencing Study
NIDDK IBD Genetics Consortium
Quebec IBD Genetics Consortium
University of Miami IBD Collaborative
IMAGINE
International Genome Sample Resource (IGSR)
Jackson Heart Study
Jewish Genome Project - funded by Bonei Olam
Kuopio Alzheimer Study
LifeLines Cohort
Lung Tissue Research Consortium (LTRC)
Material and Information Resources for Inflammatory And Digestive Diseases Biobank
McLean Program for Neuropsychiatric Research, Psychotic Disorders Division
MESTA
METabolic Syndrome In Men (METSIM)
Mass General Brigham biobank
Molecular Genetics of Cognitive Disorders in Northern Finland
Multi-Ethnic Study of Atherosclerosis (MESA)
Myocardial Infarction Genetics Consortium (MIGen):
Leicester Exome Seq
North German MI Study
Ottawa Genomics Heart Study
Pakistan Risk of Myocardial Infarction Study (PROMIS)
Precocious Coronary Artery Disease Study (PROCARDIS)
Registre Gironi del COR (REGICOR)
South German MI Study
Variation in Recovery: Role of Gender on Outcomes of Young AMI Patients (VIRGO)
National Institute of Mental Health (NIMH) Controls
NHGRI CCDG
NHLBI-GO Exome Sequencing Project (ESP)
NHLBI TOPMed

NeuroDev
Nurses' Health Study
Osaka University Graduate School of Medicine
PEGASUS
Population Architecture Using Genomics and Epidemiology (PAGE) Consortium
PRISM
Pritzker Neuropsychiatric Disorders Research Consortium
Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA)
SCHEMA - Japan
SCHEMA - Spain
Schizophrenia Trios from Taiwan
Sequencing Initiative Suomi (SiSu)
SHARE
SIGMA-T2D
SubPopulations and Intermediate Outcome Measures In COPD Study (SPIROMICS)
SUPER Study – "A Finnish study of hereditary mechanisms of psychosis disorders"
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
BioMe
GoDARTS
Framingham Heart Study
T2D-SEARCH
~~The Cancer Genome Atlas (TCGA)~~ **TCGA removed**
The Fund for Resources for Psychiatric Research
The Genetics of Atrial Fibrillation
The Genetics of Cardiovascular Disease: Atrial Fibrillation and Atrioventricular Block
The Vanderbilt Atrial Fibrillation Ablation Registry (VAFAR)
TheWellcomeTrust Case Control Consortium
THL Biobank consent in accordance with the Finnish Biobank Act
UCSF atrial fibrillation cohort
UKIBDGC - Pharmacogenetic
UK BioBank
Whole Genome Sequencing in Psychiatric Disorders (WGSPD)
Women's Health Initiative (WHI)

126K from v2 exomes

76K from v3 genomes

417K exomes from UKBB

188K exomes from many new sources (sequenced at Broad)



Breakdown of gnomAD cohort phenotypes

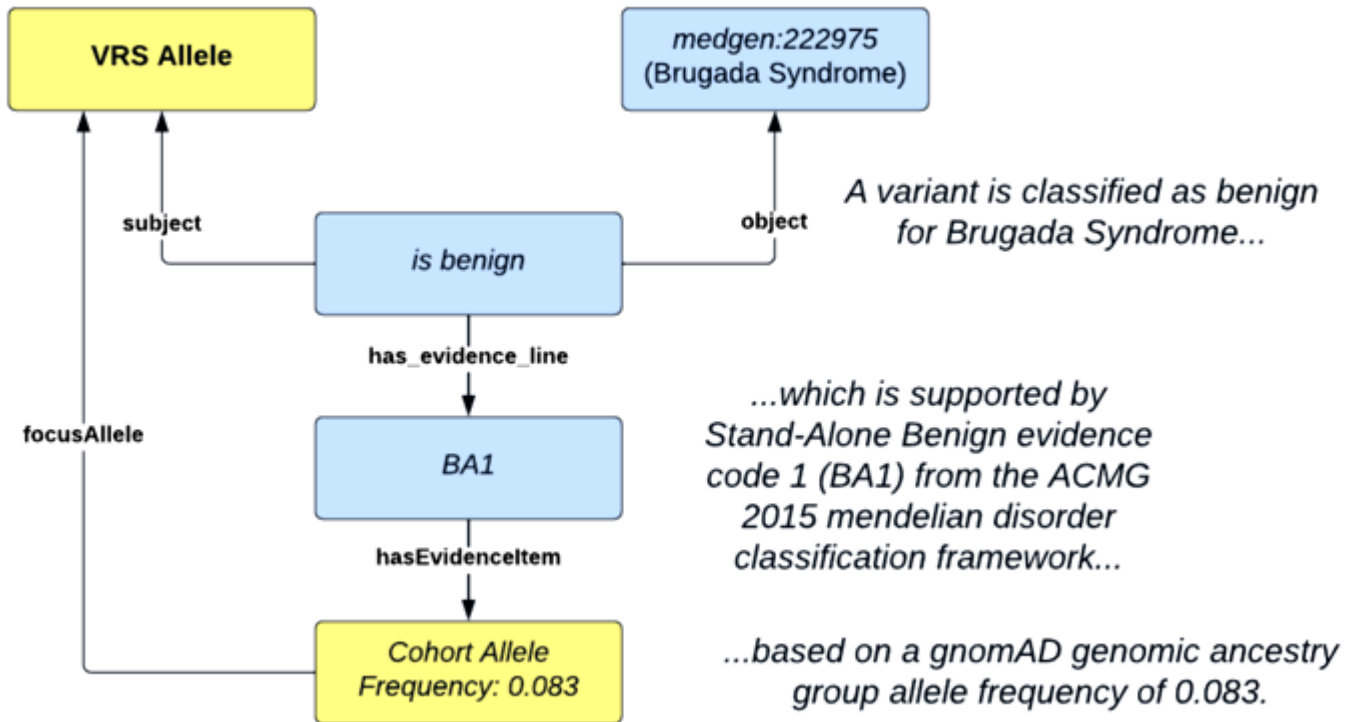
Phenotypes	Case	Control	Unknown	Total	% of cases out of all v4 exomes
Biobank or control dataset*	-	24,016	447,750	471,766	N/A
Neurodevelopmental**	-	132	-	143	N/A
Coronary heart disease	1,557	-	-	1,557	0.21%
Myocardial infarction	11,900	369	-	12,269	1.63%
Cardiac arrhythmia	458	-	-	458	0.06%
Atrial Fibrillation	4,398	3,546	38,289	46,233	0.60%
Non-specific cardiovascular disease	1,888	11,376	15,000	28,264	0.26%
Type 2 Diabetes	17,506	13,096	3,807	34,409	2.39%
Inflammatory bowel disease spectrum and related disorders^	35,008	11,928	280	47,217	4.79%
Bipolar disorder	19,284	16,383	80	35,747	2.64%
Schizophrenia spectrum and related disorders	30,278	17,689	39	47,994	4.14%
Alzheimer's disease	2,594	665	1,632	4,890	0.35%
Grand Total	124,871	99,200	506,877	730,947	17.08%

*This category includes: GTEx, 1KG, UKBB, and the Qatar Genome Project, as well as the FinnGen and MGB biobank samples when no phenotype was specified

^ includes diseases like Crohn's disease, irritable bowel syndrome, interstitial cystitis, ulcerative colitis

** Neurodevelopmental controls are unaffected parents of children with confirmed or suspected de novo cause of their neurodevelopmental disorder

Developing a Cohort Allele Frequency Model



Developed from the GA4GH Variant Annotation Specification (draft standard)

[CAF Schema](#)

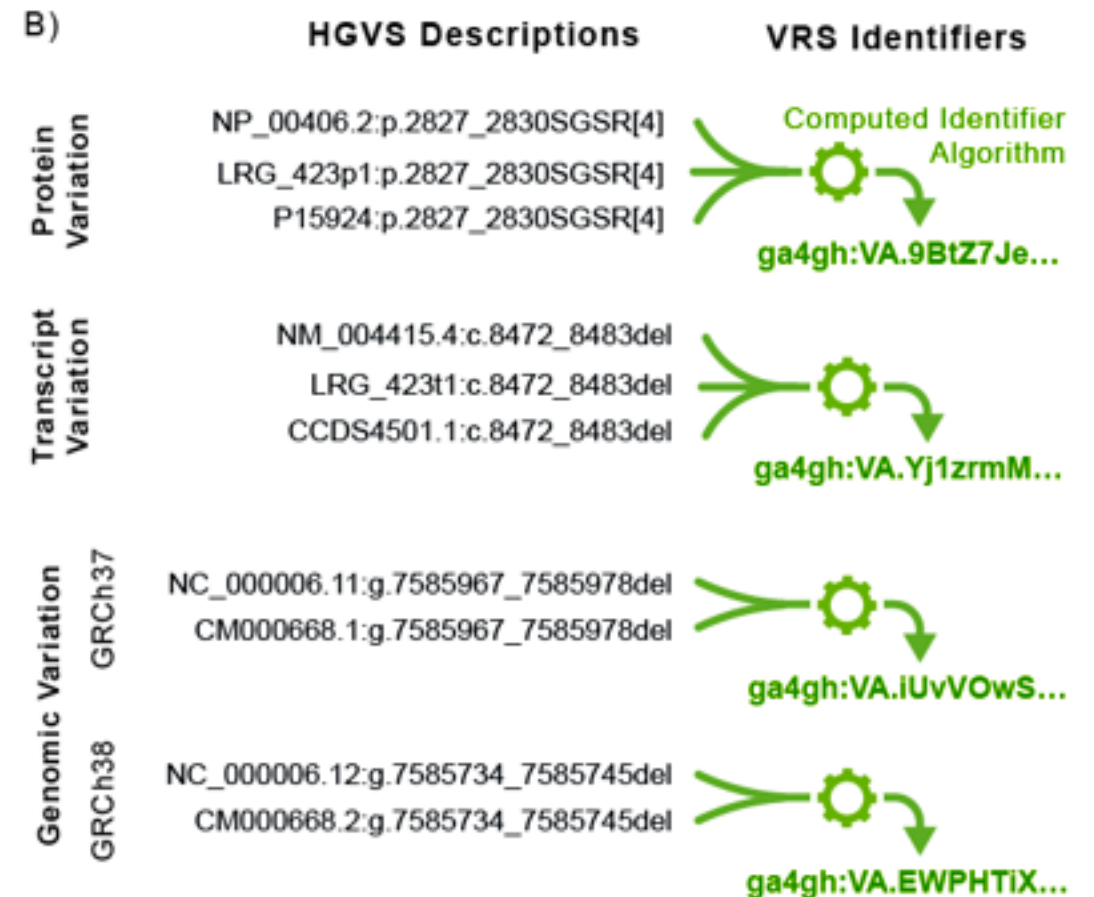
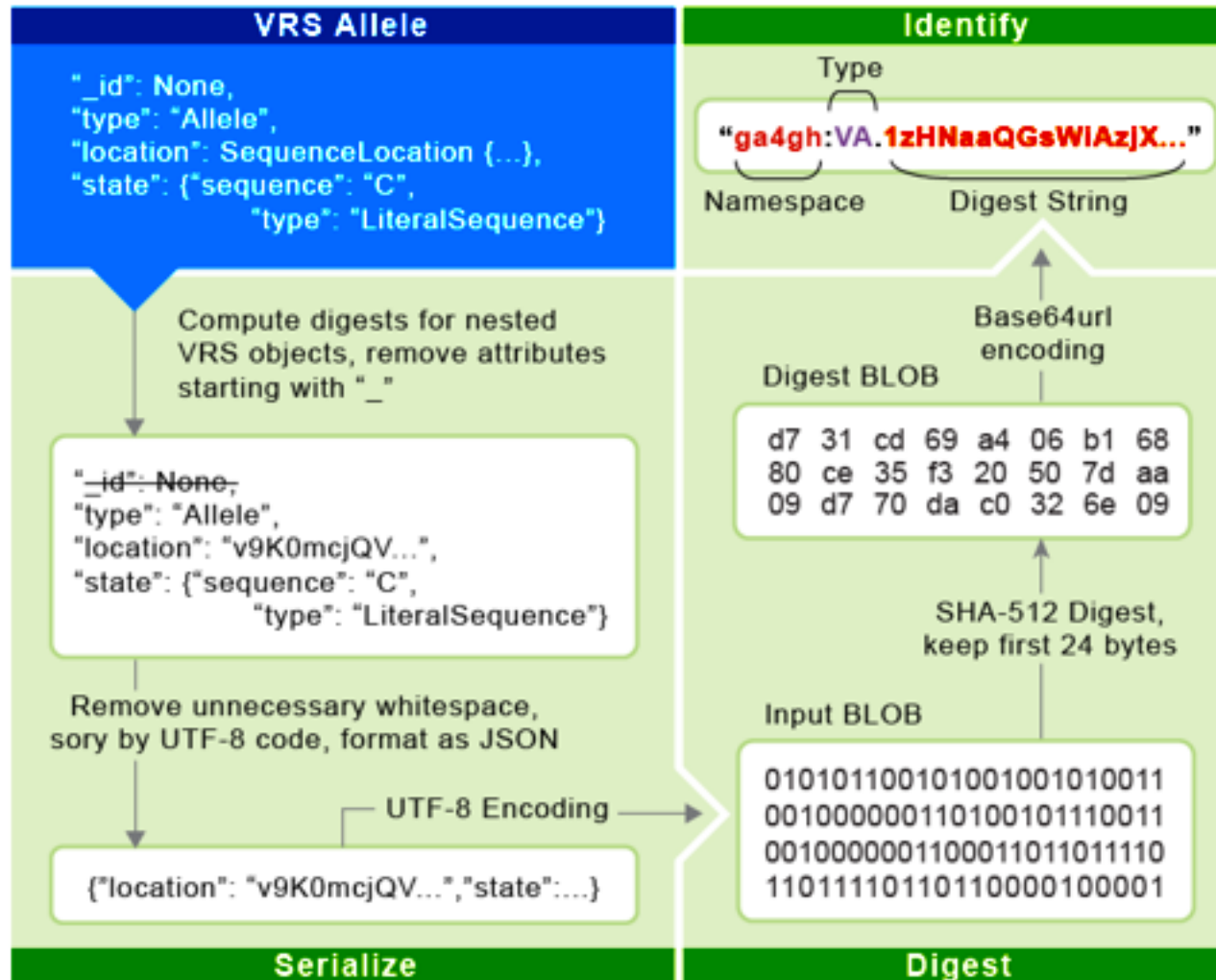
Uses the the GA4GH Variation Representation Specification (approved standard; v1.3)

[VRS Documentation](#)

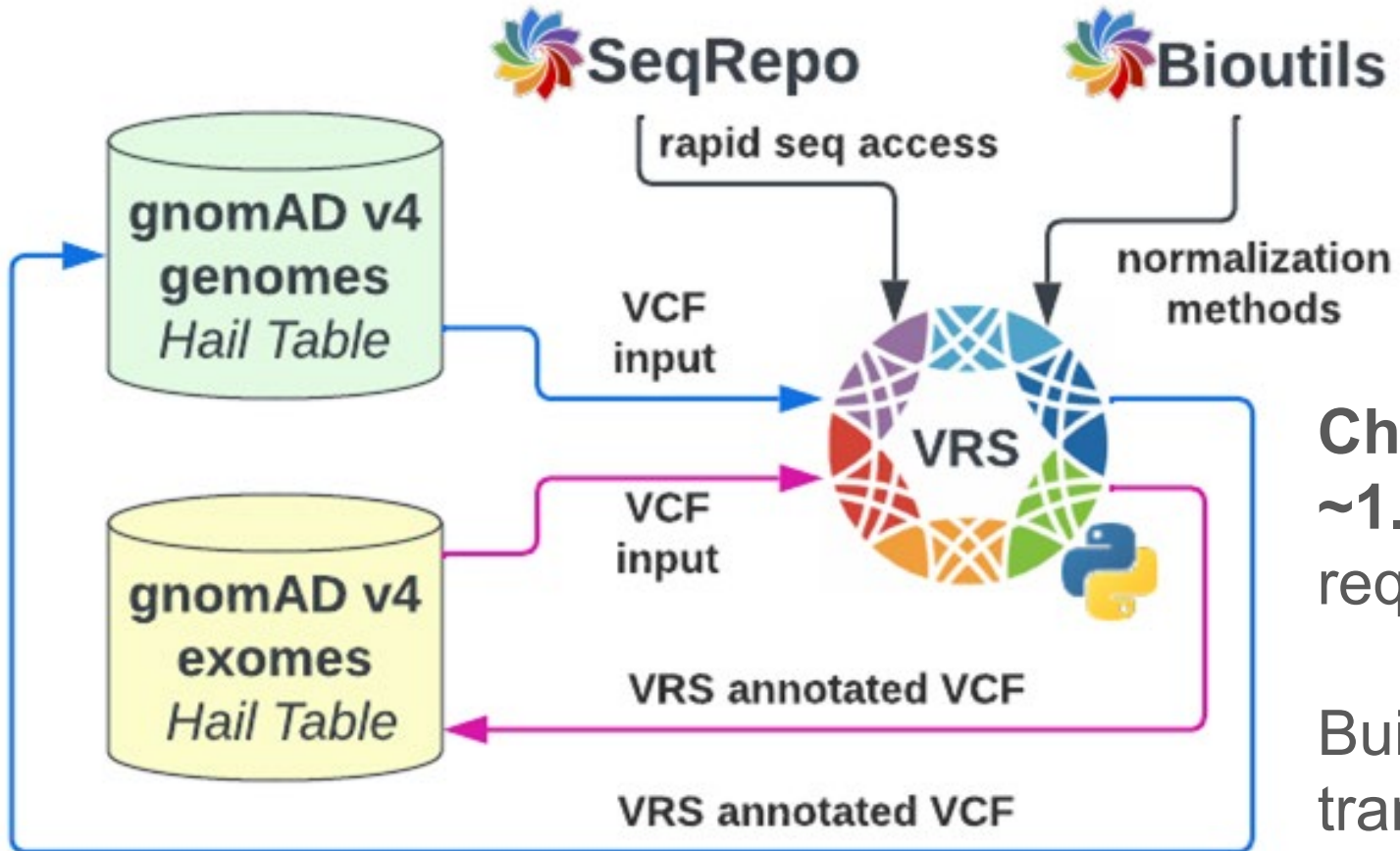
[GnomAD Blog Post](#)

Supported by [NOT-OD-22-067: Administrative Supplements to Support Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data](#)

VRS Variants and Global Identifiers



Developing High Throughput Variant Translation Tools

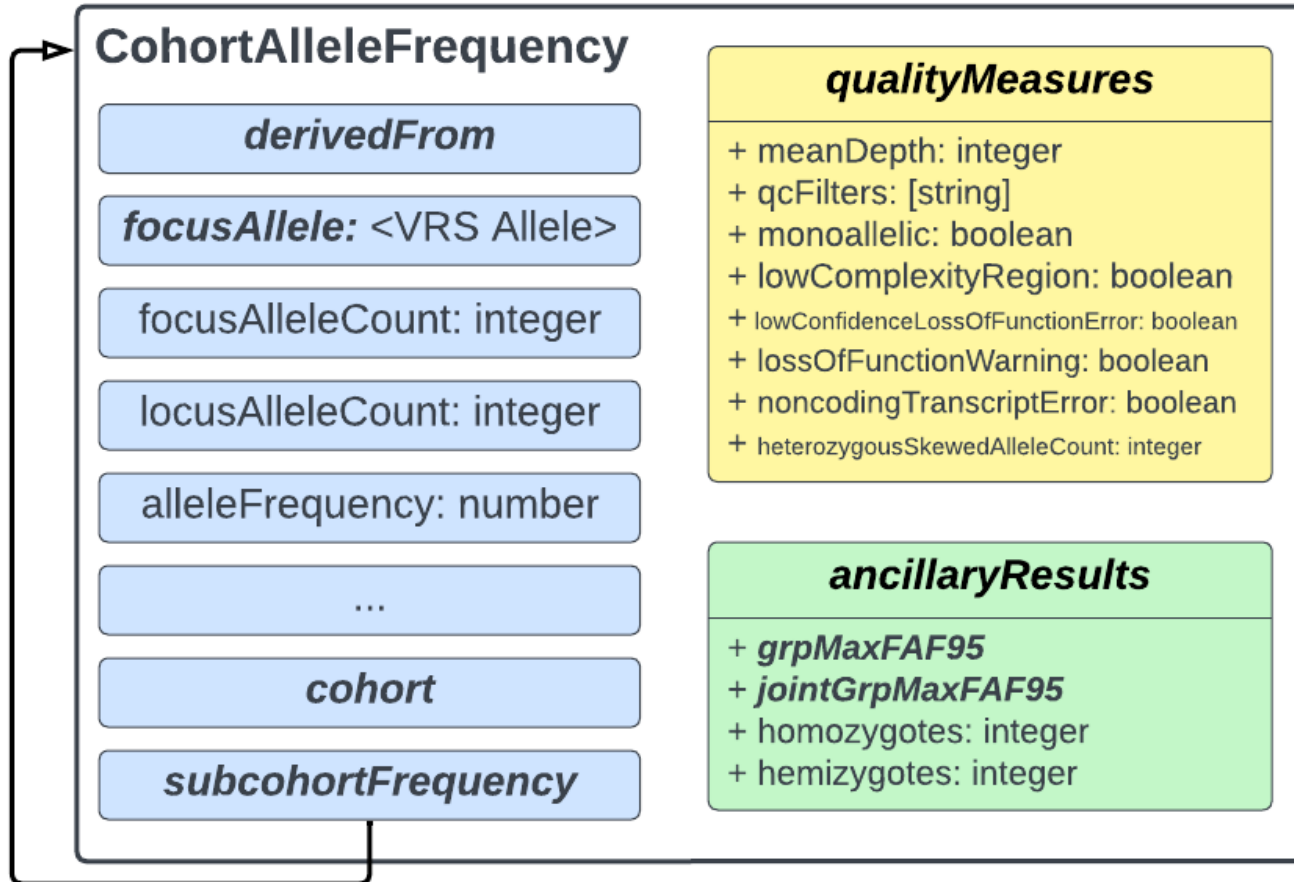


Challenge:
~1.89 billion Alleles in gnomAD
requires variant representation at scale

Built a high-throughput VRS / VCF
translation tool ([source code](#))

Leverages SeqRepo and Bioutils from
the Biocommons community

A Variant Annotation Model and Python Toolkit



Challenge:

CAF model includes global standard profile for core data, but also resource-specific *quality measures* and *ancillary results*

Created a CAF profile with global and resource-specific components

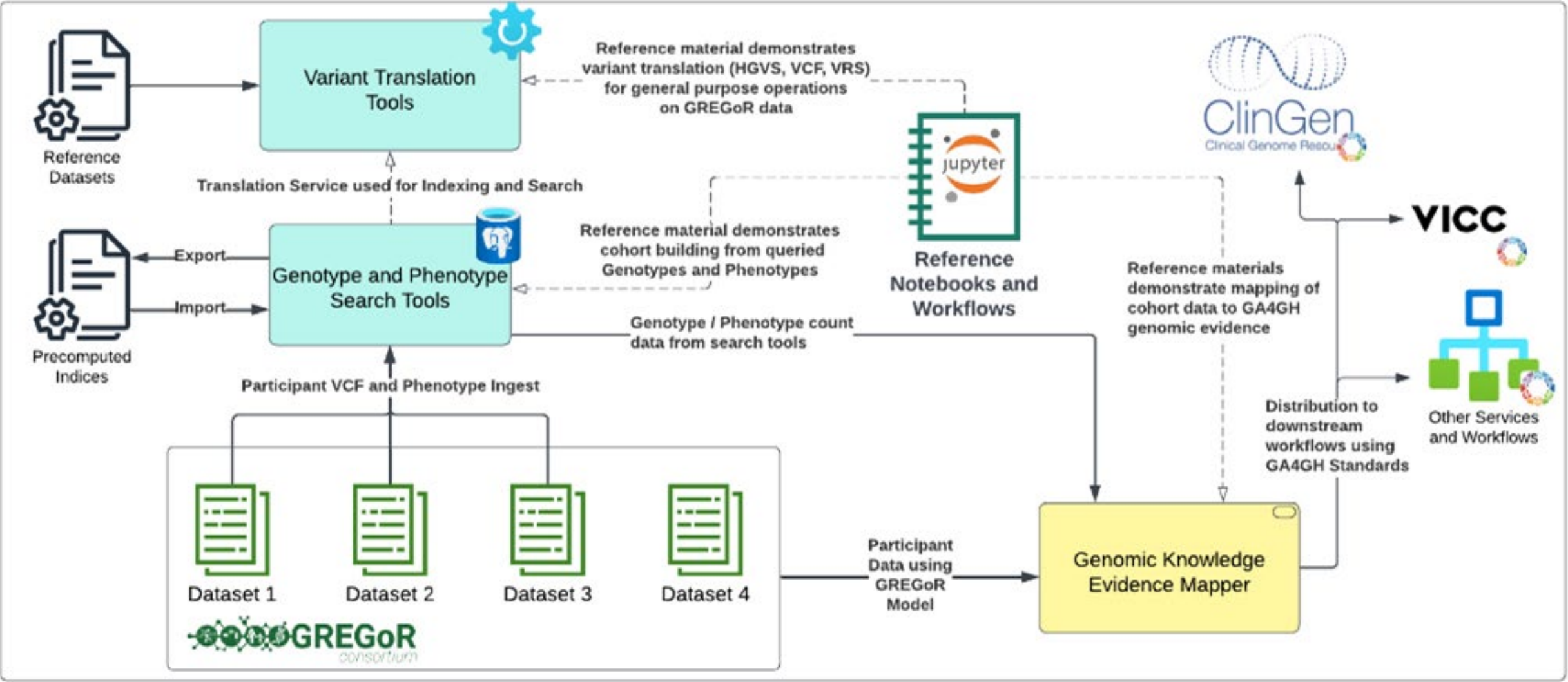
Added methods to the gnomAD Hail utils library to extract CAF-formatted data from annotated Hail tables ([source code](#))

Future work: CAF model and VRS at GA4GH Connect



1. Applications of this model to other Variant Annotation Specification profiles
2. Development of the VRS 2.0 specification and planned application to gnomAD

Coming in 2024-25: Cohort Allele Frequency from GREGoR



Acknowledgements

Wagner Lab

Wesley Goar
Kori Kuzma
James Stevenson

Broad Institute

Heidi Rehm
Larry Babb
Katherine Chao
Grace Tiao
Matthew Solomonson
Kyle Ferriter
Daniel Marten
Phil Darnowsky
Qin He

Funding



National Human Genome
Research Institute

[R35 HG011949](#)

and collaborative supplement