COMPUTABLE KNOWLEDGE PUBLICATIONS

Sync for Genes Phase 5: Computable artifacts for sharing dynamically annotated FHIR-formatted genomic variants

Robert Dolin[1] | Bret S. E. Heale[2] | Rohan Gupta[3] | Carla Alvarez[1] |
Justin Aronson[4] | Aziz Boxwala[1] | Shaileshbhai R. Gothi[5] | Ammar Husami[6] |
James Shalaby[1] | Lawrence Babb[7] | Alex Wagner[8] | Srikar Chamala[9]

[1]Elimu Informatics, El Cerrito, California, USA

[2]Humanized Health Consulting, Salt Lake City, Utah, USA

[3]Shri Mata Vaishno Devi University, Katra, India

[4]Northwestern University, Evanston, Illinois, USA

[5]University of Florida, Gainesville, Florida, USA

[6]Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[7]Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[8]Nationwide Children's Hospital/OSU, Columbus, Ohio, USA

[9]USC/Children's Hospital Los Angeles, Los Angeles, California, USA

Correspondence
Robert Dolin, Elimu Informatics, 1709 Julian Ct., El Cerrito, CA 94530-2009 USA.
Email: bdolin@elimu.io

Srikar Chamala, USC/Children's Hospital, Los Angeles, CA 90027-6016, USA.
Email: schamala@chla.usc.edu

Abstract

Introduction: Variant annotation is a critical component in next-generation sequencing, enabling a sequencing lab to comb through a sea of variants in order to hone in on those likely to be most significant, and providing clinicians with necessary context for decision-making. But with the rapid evolution of genomics knowledge, reported annotations can quickly become out-of-date. Under the ONC Sync for Genes program, our team sought to standardize the sharing of dynamically annotated variants (e.g., variants annotated on demand, based on current knowledge). The computable biomedical knowledge artifacts that were developed enable a clinical decision support (CDS) application to surface up-to-date annotations to clinicians.

Methods: The work reported in this article relies on the Health Level 7 Fast Healthcare Interoperability Resources (FHIR) Genomics and Global Alliance for Genomics and Health (GA4GH) Variant Annotation (VA) standards. We developed a CDS pipeline that dynamically annotates patient's variants through an intersection with current knowledge and serves up the FHIR-encoded variants and annotations (diagnostic and therapeutic implications, molecular consequences, population allele frequencies) via FHIR Genomics Operations. ClinVar, CIViC, and PharmGKB were used as knowledge sources, encoded as per the GA4GH VA specification.

Results: Primary public artifacts from this project include a GitHub repository with all source code, a Swagger interface that allows anyone to visualize and interact with the code using only a web browser, and a backend database where all (synthetic and anonymized) patient data and knowledge are housed.

Conclusions: We found that variant annotation varies in complexity based on the variant type, and that various bioinformatics strategies can greatly improve automated annotation fidelity. More importantly, we demonstrated the feasibility of an ecosystem where genomic knowledge bases have standardized knowledge (e.g., based on the GA4GH VA spec), and CDS applications can dynamically leverage that

knowledge to provide real-time decision support, based on current knowledge, to clinicians at the point of care.

KEYWORDS
clinical decision support, EHR, genomics

1 | INTRODUCTION

In 2017, the Office of the National Coordinator for Health Information Technology (ONC) launched the Sync for Genes program,[1] aiming to standardize the sharing of genomic information among laboratories, providers, patients, and researchers. In particular, ONC sought to achieve these objectives with the use of the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard.

Phase 1 through 4 of the Sync for Genes program sponsored a number of pilots focusing on various aspects of standardizing genomic data, integrating genomic data, engaging with laboratories, and sharing genomic variants for patient care. Fundamental to these pilots was the use and maturation of the FHIR Genomics Reporting Implementation Guide.[2] Sync for Genes Phase 5 sought to standardize the sharing of annotated variants, again using FHIR, but also using knowledge structures defined by the Global Alliance for Genomics and Health (GA4GH) standards organization.

Variants (also sometimes referred to as "mutations") may in and of themselves be meaningless to clinicians (e.g., patient is found to have a heterozygous NC_000006.11:26091178:C:G variant) without additional context. Annotating variants with various pieces of information (e.g., annotating NC_000006.11:26091178:C:G with a diagnostic implication of hemochromatosis; annotating CYP2C19 *1/*2 with a therapeutic implication of altered clopidogrel metabolism) is a critical component in next-generation sequencing (NGS) analysis and interpretation. Variant annotation is crucial to enable a sequencing lab to comb through a sea of variants in order to hone in on those likely to be most significant.[3,4] Sequencing labs leverage annotations as part of a variant prioritization and filtering process in an attempt to identify that subset of variants that warrant inclusion in a clinical report as clinically actionable. Clinicians receiving these reports rely on these annotations to provide proper context for decision-making.

Likewise, annotation is a key component of variant reanalysis. It is well recognized that with the rapid evolution of genomic knowledge, the subset of a patient's variants deemed clinically relevant changes over time,[5] rapidly enough that the snapshot represented in a report risks becoming dangerously outdated. Additionally, many reporting platforms are context limited by the purpose of the ordered report, only reporting, for instance, on variants present in a select set of genes. Consider, for instance, the genetic variants recommended for reporting as secondary/incidental findings by the American College of Genetics and Genomics (ACMG).[6] Studies have found as many as 7% of patients, across many different ancestries, harbor pathogenic or likely pathogenic variants in these 66 ACMG genes.[7-9] Analysis of the National Center for Biotechnology Information (NCBI) ClinVar data shows that from 2016 to 2018, the number of known pathogenic or likely pathogenic variants in these genes went from 10 137 to 18 718,[10,11] reflecting over 8000 new or recategorized clinically significant variants over the course of 2 years. What are the implications for patients who were sequenced in 2016? Evidence such as this has led to the widely held belief that since variant annotations change over time, there is a need for periodic variant reanalysis.[12-14]

Many barriers to variant reanalysis exist. A survey of 21 labs in the United States noted that few policies existed documenting laboratory procedures for reanalysis.[15] Significant time and financial commitments may be required[16] and there is evolving legal debate over who holds responsibility for initiating reanalysis.[17] Presently, there is no standards-based approach for refreshing the variants and variant annotations delivered to a clinician via a typical clinical report.

To address this need, under Sync for Genes Phase 5, our team sought to extend ONC's objectives by not only developing standardized methods for sharing statically annotated variants (e.g., those coming in a lab report), but also for developing standardized methods for sharing variants dynamically annotated (i.e., annotated on demand) with GA4GH-encoded knowledge so that clinical decision support (CDS) applications can surface up-to-date annotations to clinicians—a foundational consideration for a Learning Health System, where knowledge updates can be immediately surfaced in a clinical setting. By decoupling the updated knowledge from the reporting of sequencing results, our pipeline enables the most current annotations to be presented to clinicians. The intent of this manuscript is to summarize the computable biomedical knowledge (CBK) artifacts that were developed under the Sync for Genes Phase 5 that enable sharing of dynamically annotated variants.

2 | BACKGROUND

Here we summarize key preexisting standards and artifacts leveraged by the CBK tools developed and described in this report.

2.1 | Fast healthcare interoperability resources genomics

FHIR is a next-generation interoperability standard designed to enable health data, including clinical and administrative data, to be quickly and efficiently exchanged. Based on common World Wide Web technologies and core application programming interface (API) capabilities, coupled with base semantic resources that enable easy exchange of

conditions, medications, laboratory observations, and more, HL7 FHIR has gained rapid acceptance on a global scale as an innovative standard for enabling health data interoperability.

The FHIR Genomics Reporting Implementation Guide (FHIR Genomics)[2] defines FHIR representations for a range of genomic data structures (e.g., variants, haplotypes, and variant annotations), enabling a standards-based communication of simple and structural variants, germline and somatic variants, pharmacogenomic star alleles, Human leukocyte antigen typing, and other findings generated from sequencing, chip technology, and cytogenetic analysis. Annotation types supported by the standard include diagnostic implications (e.g., variant is associated with a particular disease), therapeutic implications (e.g., presence of variant predicts a treatment response to a particular drug, presence of a haplotype predicts altered pharmacokinetics of a particular drug), and molecular consequences (e.g., presence of a variant predicts loss of function of the containing gene). It should be noted that FHIR Genomics is designed to enable the structured communication of variant annotations, irrespective of whether these annotations are static (e.g., conveyed on a lab report) or dynamic (e.g., computed on the fly at the time the FHIR instance is created).

2.2 | Global alliance for genomics and health

GA4GH is an international standards development organization focused on genomic data sharing and interoperability. GA4GH is organized into a set of "work streams" that bring together users and developers to address real-world needs. Intersecting these work streams are "driver projects," real-world genomic data initiatives that help guide development efforts.

Not only is GA4GH, through the Large-Scale Genomics work stream, the shepherd for core bioinformatics data standards such as the Variant Call Format (VCF) and the Binary Alignment Map (BAM), GA4GH is also focused on data security and regulatory issues,

standardizing phenotype data, cloud-based data sharing, and more. Driver projects such as the Variant Interpretation for Cancer Consortium (VICC) and the Clinical Genome Resource (ClinGen) are advancing data sharing initiatives that promise to improve precision medicine and research.

The GA4GH Genomic Knowledge Standards work stream develops, adopts, and adapts standards-based components to enable the exchange of reference genomic information through common APIs, thereby enabling the downstream analysis of genomic data. In particular, the evolving Variant Annotation (VA) standard seeks to establish extensible data models to support representation of diverse kinds of statements made about genetic variation, and the evidence and provenance supporting these statements. The GA4GH VA standard can potentially be used to standardize the contents of genomics knowledge bases such as ClinVar,[18] Clinical Interpretation of Variants in Cancer (CIViC),[19] or Pharmacogenomics Knowledge Base (PharmGKB).[20] Membership in the Genomic Knowledge Standards work stream overlaps with membership in the HL7 Clinical Genomics committee that is charged with FHIR Genomics maintenance, leading to an ongoing collaboration, harmonization, and sharing of ideas between GA4GH and HL7. In late 2021, members of our Sync for Genes Phase 5 team helped harmonize portions of the GA4GH VA specification with FHIR Genomics, as shown in Figure 1. The outcome of this harmonization was the close alignment between variant annotation semantics in GA4GH VA and FHIR Genomics, with a straightforward mapping.

2.3 | FHIR genomics operations

Over the past several years, our team and other collaborators have extended FHIR Genomics with FHIR Genomics Operations.[21]

The FHIR standard describes a mechanism for extending basic FHIR query capabilities through the creation of "Operations." FHIR Operations are a standardized way to extend the RESTful FHIR API's
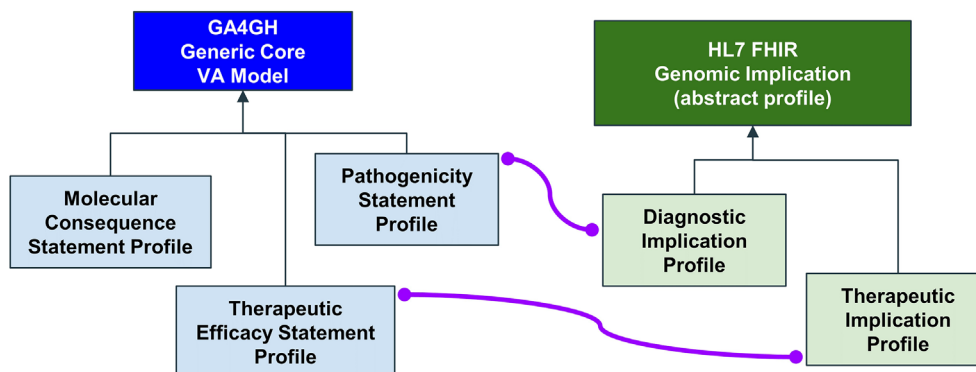

FIGURE 1　Global Alliance for Genomics and Health-Fast Healthcare Interoperability Resources (GA4GH-FHIR) variant annotation harmonization. A portion of the April 2022 GA4GH Variant Annotation model is shown on the left (in blue), whereas portions of the FHIR Genomics model are shown on the right (in green). The GA4GH Therapeutic Efficacy Statement Profile has been harmonized with the FHIR Genomics Therapeutic Implication Profile, and the GA4GH Pathogenicity Statement Profile has been harmonized with the FHIR Genomics Diagnostic Implication Profile.

Create/Read/Update/Delete actions and enable use cases where servers play an active role in formulating responses, where the intended purpose is to cause side effects such as the creation of new resources, and for data normalization to abstract away from variability in data representation. Many FHIR specifications supplement the standardization of data structures with the addition of FHIR Operations that define advanced API capabilities.

FHIR Genomics Operations extend the FHIR Genomics standard and basic FHIR search capabilities in order to simplify developer access to potentially complex and voluminous data structures. The Operations are based on the premise that genomic data, in FHIR format and/or some other format (e.g., VCF format), are stored in a repository, either in or alongside an Electronic Health Record (EHR), possibly along with phenotype annotations. The FHIR Genomics Operations essentially "wrap" the repository, presenting a uniform interface to applications, regardless of internal repository data structures.

We categorize FHIR Genomics Operations along two orthogonal axes—subject versus population, and genotype versus phenotype, as shown in Figure 2. For example, the "find-subject-variants" operation is categorized as a "subject" and a "genotype" operation, that retrieves genotype information for a single subject, whereas the "find-population-tx-implications" is categorized as a "population" and a "phenotype" operation, that retrieves a count or list of patients having specific phenotypes (such as being intermediate metabolizers of clopidogrel). The metadata operation retrieves metadata about the genomic studies that generated the data.

Phenotype operations are particularly relevant to the CBK artifacts developed under Sync for Genes Phase 5. These operations return diagnostic and therapeutic implications of a subject or a population's variants. Operations can return previously instantiated implications (e.g., those that came in via a static lab report) and/or dynamically computed implications (e.g., those computed on the fly using an associated knowledge base), thereby decoupling the reporting of variants from the maintenance of knowledge and the reporting of annotations.

In early 2022, our team created an open-source reference implementation of the FHIR Genomics Operations,[22] as shown in Figure 3. All patient data used in this implementation is either synthetic or publicly available and anonymized.

The reference implementation uses a standard Swagger interface to allow anyone to visualize and interact with the operations using only a web browser. All source code is available on the reference implementation's GitHub site. Patient data and knowledge data reside in a MongoDB repository, also publicly available. Detailed documentation is provided on the GitHub site's wiki.

3 | APPROACH

To address the challenge of sharing dynamically annotated variants in a standardized way, we developed a CDS pipeline that annotates patient's variants through an intersection with current knowledge and serves up the FHIR-encoded variants and annotations (diagnostic and therapeutic implications, molecular consequences, population allele frequencies) via FHIR Genomics Operations. Knowledge sources used were ClinVar, CIViC, and PharmGKB, encoded as per the draft GA4GH Variant Annotation specification. Data returned by the FHIR Genomics Operations were surfaced via two apps.

Two primary use cases helped to drive project priorities and design decisions: (1) Clinician wants to view up-to-date genetic implications, including germline cancer screening and germline pharmacogenomic screening; (2) Clinician wants to filter and prioritize potentially causative variants as part of a rare disease diagnostic workup. These use cases were chosen in part to help us examine our ability to selectively return only those annotations relevant to the specific context. In use case (1), we assumed that clinicians only want to see strongly evidence-based annotations that are deemed pathogenic or likely pathogenic, whereas in use case (2), we assumed that clinicians want to see a much broader scope of annotations, including variants of unknown significance and predicted molecular consequences. These use cases served as the basis for the two apps.

	Subject Operations	Population Operations
Genotype Operations	**Retrieve simple / structural variants, haplotypes** > find-subject-variants > find-subject-specific-variants > find-subject-structural-intersecting-variants > find-subject-structural-subsuming-variants > find-subject-haplotypes > find-subject-specific-haplotypes	**Identify cohorts with simple / structural variants, haplotypes** > find-population-specific-variants > find-population-structural-intersecting-variants > find-population-structural-subsuming-variants > find-population-specific-haplotypes
Phenotype Operations	Retrieve diagnostic and therapeutic **implications** > find-subject-tx-implications > find-subject-dx-implications	**Identify cohorts with diagnostic and therapeutic implications** > find-population-tx-implications > find-population-dx-implications
Metadata Operations	**Retrieve sequencing study metadata** > find-study-metadata	

FIGURE 2 Scope of Fast Healthcare Interoperability Resources (FHIR) Genomics Operations. We categorize FHIR Genomics Operations along two orthogonal axes—subject versus population, and genotype versus phenotype. Each cell shows an overall description of the operations in that group, along with a bulleted list of the actual operations.
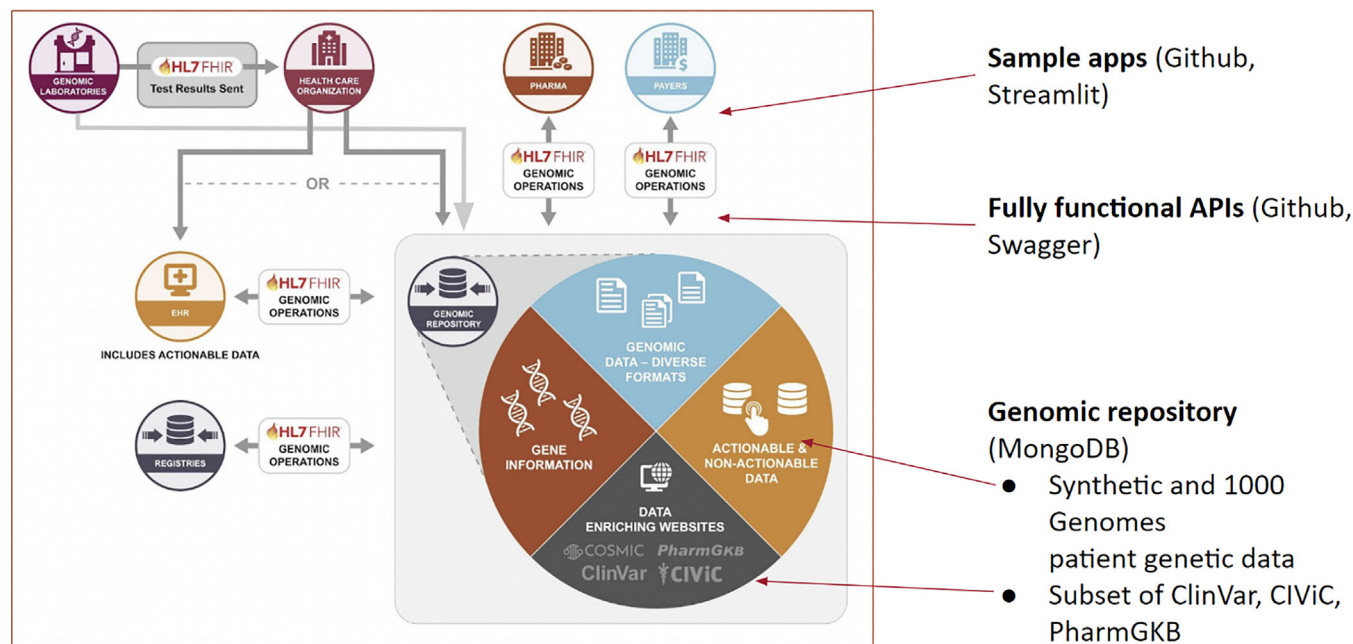
FIGURE 3   Fast Healthcare Interoperability Resources (FHIR) Genomics Operations public reference implementation. The reference implementation includes source code for the Operations, sample apps, functional application programming interfaces (via a Swagger interface), and a genomic repository containing patient genetic data and knowledge. (Image courtesy of HL7 FHIR Accelerator CodeX).

Our approach was to: (1) Enhance the FHIR Genomics Operations reference implementation to dynamically annotate genomic findings with GA4GH-encoded knowledge; (2) Enhance the reference implementation to include a molecular consequence and population allele frequency annotation pipeline; (3) Return annotated genomic information using FHIR Genomics Operations; and (4) Demonstrate the enhanced pipeline via two proof-of-concept apps.

3.1 | Enhance the FHIR genomics operations reference implementation to dynamically annotate genomic findings using GA4GH-encoded knowledge

We extracted representative subsets of ClinVar, CIViC, and PharmGKB from source knowledge bases, retaining native drug and condition codes used by each knowledge base. The Clinvar snapshot, drawn from both variant summary data and submission summary data, is limited to ACMG genes.[6] Conditions are coded with Medgen (https://www.ncbi.nlm.nih.gov/medgen/) codes. The PharmKGB snapshot is limited to Clinical Pharmacogenetics Implementation Consortium (CPIC) Level A star alleles in CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5, NUDT15, SLCO1B1, TPMT, UGT1A1. Medications are coded with RxNorm (https://www.nlm.nih.gov/research/umls/rxnorm/) ingredient codes. The CIViC snapshot is limited to simple variants. Conditions are coded with Disease Ontology (https://disease-ontology.org/) codes and medications are coded with RxNorm ingredient codes. Knowledge was manually formatted as per the latest GA4GH Variant Annotation draft specification, using specifications and data obtained from the GA4GH Genomic Knowledge Pilot

(https://gk-pilot.readthedocs.io/en/latest/index.html). In the future, once the GA4GH Variant Annotation specification is finalized, this process will be automated to execute at some frequency, thereby keeping knowledge in the reference implementation up-to-date.

We revised the reference implementation internal data structures, enabling a straightforward mapping from GA4GH VA-encoded data into our local schema. We imported the GA4GH VA-encoded-knowledge into the reference implementation's database, populating these revised data structures. As part of database loading, we normalized all simple variants (SNVs, MNVs, InDels), based on the NCBI canonical Sequence-Position-Deletion-Insertion (SPDI) format.[23]

The FHIR Genomics Operations "find-subject-tx-implications" retrieves genetic therapeutic implications for variants/haplotypes/genotypes, whereas "find-subject-dx-implications" retrieves genetic diagnostic implications for variants. In the reference implementation, these operations dynamically intersect patient variants (also normalized into canonical SPDI) against knowledge structures in order to compute implications. We modified the implementation of these operations to use the revised internal data structures, thereby enabling the dynamic intersection of patient variants against knowledge originating in GA4GH VA format.

3.2 | Enhance the reference implementation to include a molecular consequence and population allele frequency annotation pipeline

We had previously created a utility that automates loading VCF files into the reference implementation's database (based on vcf2fhir[24]). Here, we enhanced the loading process to invoke a molecular

```
1       86852621        rs2390096       A       G       .       .
NS=1;AN=2;AC=1;ANN=G|splice_donor_variant&intron_variant|HIGH|ODF2L|ODF2L|transcript|NM_0013955
25.1|protein_coding|2/16|c.-281+2T>C||||||,G|synonymous_variant|LOW|ODF2L|ODF2L|transcript|NM_0
01366781.1|protein_coding|2/17|c.90T>C|p.Cys30Cys|440/4198|90/1935|30/644||,G|5_prime_UTR_varia
nt|MODIFIER|ODF2L|ODF2L|transcript|NM_001395543.1|protein_coding|2/14|c.-171T>C||||3763|,G|int
ron_variant|MODIFIER|ODF2L|ODF2L|transcript|NM_001395544.1|protein_coding|1/11|c.-21-3742T>C|||
||||;LOF=(ODF2L|ODF2L|35|0.03);POPAF=3.27084e-01
GT:PS:FT:GQ:HQ:EHQ:CGA_CEHQ:GL:CGA_CEGL:DP:AD:CGA_RDP
1/0:.:PASS:100:100,100:99,99:39,40:-100,0,-100:-39,0,-40:13:5,8:8
```

FIGURE 4   Representative Variant Call Format (VCF) row, annotated by snpEff to show predicted molecular consequences and population allele frequency for a variant in the ODF2L gene. Each row of a VCF file represents a specific variant. The first field is the chromosome number, the second field is the position on that chromosome, the third field is an optional identifier for the variant, the fourth field is the reference allele, and the fifth field is the observed allele. The long INFO field is a series of subfields that further characterize the variant, and this is where snpEff inserts molecular consequence predictions for each known transcript (indicated by brown, red, yellow, and blue). Here, snpEff also predicts loss of gene function (purple "LOF"). After INFO is the FORMAT field, a list of subfields that characterize the genotype. Finally, there are SAMPLE fields for each sample tested, showing observed genotype information (e.g., heterozygous) for the variant. In the figure, the sample has a heterozygous variant on chromosome 1, at position 86 852 621, where the patient has a "G" instead of the reference "A".

consequence prediction tool, snpEff,[25] that provides transcript-level predictions regarding the molecular implications of a variant, and population allele frequencies drawn from gnomAD v2.1.1.[26] A representative VCF row is shown in Figure 4, showing predicted molecular consequences and population allele frequency for a variant in the ODF2L gene. We revised the reference implementation internal data structures to accommodate molecular consequence and population allele frequency data, and we further revised the VCF loader to import and parse this additional data.

Molecular consequence data was not derived from a GA4GH VA-encoded format.

3.3 | Return annotated genomic information using FHIR genomics operations

Having updated the reference implementation, the "find-subject-variants" operation now returns FHIR-encoded variants with molecular consequence annotations; the "find-subject-tx-implications" operation now returns FHIR-encoded variants with dynamically computed therapeutic implications (where those implications are drawn from CIViC and PharmGKB); and the "find-subject-dx-implications" operation now returns FHIR-encoded variants with dynamically computed diagnostic implications (where those implications are drawn from ClinVar).

3.4 | Demonstrate the enhanced pipeline via two proof-of-concept apps

To demonstrate the entire CDS pipeline, we developed two apps—a SMART-on-FHIR "face sheet" app that shows up-to-date genetic implications by superimposing highly curated ACMG and pharmacogenomics annotations onto a patient's face sheet; and a "variant summary" app that enables filtering/prioritization using all annotation data (diagnostic and therapeutic implications, molecular consequences, population allele frequencies).

4 | DEPLOYMENT

Primary knowledge artifacts from this project revolve around the FHIR Genomics Operations and the enhancements that were made to enable dynamic annotation of variants with GA4GH VA-encoded knowledge. Knowledge artifacts include: (1) GitHub repository containing source code (for Operations implementation, for genomic utilities, for apps developed using the Operations); (2) Swagger interface that allows anyone to visualize and interact with the Operations using only a web browser; (3) backend database where all (synthetic and anonymized) patient data and knowledge are housed. The reference implementation is written in Python and deployed on Heroku with data stored in MongoDB.

4.1 | FHIR genomics operations enhancements

Under the Sync for Genes Phase 5 project, a reference implementation of the FHIR Genomics Operations was enhanced to return diagnostic and therapeutic implications derived from GA4GH VA-encoded knowledge, and molecular consequence predictions and population allele frequencies that can be used to aid variant filtering and prioritization. The following examples illustrate these enhancements. URLs are "live" in that readers can simply copy and paste them into a browser. Alternatively, users can test these examples via Postman or via the Swagger interface. All Operations return JSON-formatted FHIR data. A detailed description of the response from each operation can be found on the HL7 site (http://build.fhir.org/ig/HL7/genomics-reporting/operations.html).

4.1.1 | Diagnostic and therapeutic implications derived from GA4GH VA-encoded knowledge

Patient CA12345 has metastatic non-small cell lung cancer. Biopsy shows two somatic variants felt to be oncogenic: NM_002524.5:

c.182A > C (NRAS:p.Gln61Pro), and NM_001354609.2:c.1799 T > A (BRAF:p.V600E). The clinician now wants to determine if there are any molecularly guided medication treatment options for this patient. Results, dynamically drawn from CIViC, show predicted resistance to dabrafenib, sensitivity to vemurafenib, and sensitivity to dabrafenib +trametinib (https://fhir-gen-ops.herokuapp.com/subject-operations/phenotype-operations/$find-subject-tx-implications?subject=CA12345&variants=NM_002524.5:c.182A>C,NM_001354609.2:c.1799T>A&conditions=https://disease-ontology.org|3908).

Patient HG02657 has liver disease, and the patient's clinician suspects hemochromatosis. The clinician wants to see if patient HG02657 has any variants associated with hereditary hemochromatosis. Results, dynamically drawn from ClinVar, show presence of variant NC_000006.11:26091178:C:G, pathogenic for hemochromatosis type 1 (https://fhir-gen-ops.herokuapp.com/subject-operations/phenotype-operations/$find-subject-dx-implications?subject=HG02657&conditions=https://www.ncbi.nlm.nih.gov/medgen|C3469186).

4.1.2 | Molecular consequence predictions and population allele frequencies can be used to aid variant filtering and prioritization

Patient HG00403 is suspected of having familial hypercholesterolemia but has no variants found in ClinVar. The clinician wants to see if patient HG00403 has any potentially causative novel variants in LDLR. Results show variant NC_000019.9:11210927:C:T predicted to cause loss of function (https://fhir-gen-ops.herokuapp.com/subject-operations/genotype-operations/$find-subject-variants?subject=HG00403&ranges=NC_000019.10:11089431-11133820&includeVariants=true).

4.2 | Sync for Genes Phase 5 apps

We developed two proof-of-concept apps under Sync for Genes Phase 5, shown in Figure 5. These apps illustrate potential scenarios for real-world use of the FHIR Genomics Operations and show how clinicians can be presented with optimized user interfaces while being shielded from the complexity of the underlying data.

The Face Sheet app (video demonstration available at https://vimeo.com/798557155) is a SMART-on-FHIR app used to show up-to-date genetic implications. Genetic interactions are displayed in four "widgets": (1) Genetic Screening widget: This widget shows the results of full ACMG and CPIC Level A screening. Clicking on a given interaction surfaces a dialog box with additional details; (2) Problem List widget: Where a variant detected through genetic screening is a potential etiology of an item on the patient's problem list, that problem list item is flagged. Clicking a DNA icon surfaces a dialog box with additional details; (3) Medication List widget: Where a variant detected through genetic screening potentially affects the behavior of a patient's medication, that medication is flagged. Clicking a DNA icon

surfaces a dialog box with additional details; and (4) Allergy/Intolerance List widget: Where a variant detected through genetic screening is a potential basis for an identified allergy, that allergy is flagged. Clicking a DNA icon surfaces a dialog box with additional details.

The Face Sheet source code and executable are not public at the time of this writing. The rationale is that (1) the app has proprietary code dependencies; (2) the app is deployed in an environment where open access presents security challenges. As noted above, a video demonstration and all FHIR Genomics Operations called by the app are available. In the future, we plan to deploy the app in a sandbox environment to enable public accessibility.

The Variant Summary app (source code available at https://github.com/FHIR/genomics-operations/blob/main/genomics-apps/getMolecularConsequences.py; run the app at https://getmolecularconsequences.streamlit.app/; video demonstration available at https://vimeo.com/798557133) is an open source app used to show a wide summary of implications and consequences for variants in specified regions of a patient's genome. The intent is to enable filtering/prioritization of variants for rare disease discovery, as described by Austin-Tse, et al.[3]

5 | DISCUSSION

We found that it is possible to share dynamically annotated genomic information, using GA4GH-encoded knowledge, delivered in FHIR Genomics format using FHIR Genomics Operations. Prior harmonization work described above enabled a simple mapping from GA4GH VA-encoded knowledge into FHIR Genomics diagnostic and therapeutic implication profiles. Minor issues encountered will be brought back to respective committees for resolution, and largely relate to the need for further harmonization around molecular consequences and knowledge provenance.

Our approach to "dynamic annotation" was to computationally intersect a patient's genomic variants with knowledge. We found, as others have previously reported, that dynamic annotation varies in complexity based on the variant type, as shown in Figure 6.

Dynamic annotation of single nucleotide variants (SNVs) and insertion-deletion variants (InDels) benefits from normalization of both patient variants and knowledge into a common canonical form. Both FHIR Genomics Operations and GA4GH leverage canonical SPDI format, making annotation of these variant types relatively straightforward.

A multi-nucleotide variant (MNV) exists where two or more SNVs combine into a larger variant (e.g., MNV AG > TC is comprised of SNV A > T and adjacent SNV G > C). Dynamic annotation of MNVs is complicated by the fact that (1) some variant callers will report MNVs whereas others only report component SNVs; (2) knowledge bases may contain MNVs and/or SNVs depending on what was submitted; and (3) bioinformatics tools may predict different molecular consequences for an MNV versus component SNVs. In fact, literature suggests that misannotation of MNVs is common and carries significant clinical implications.[27-29] Our approach has been to leverage variant
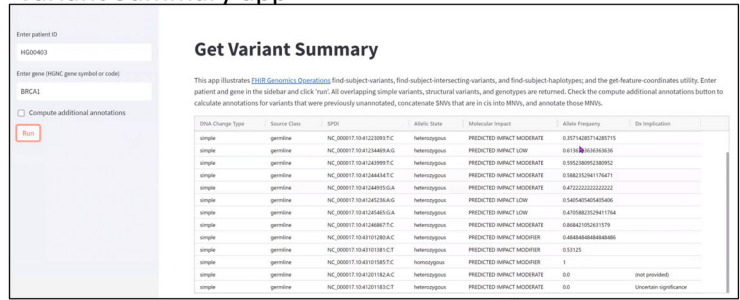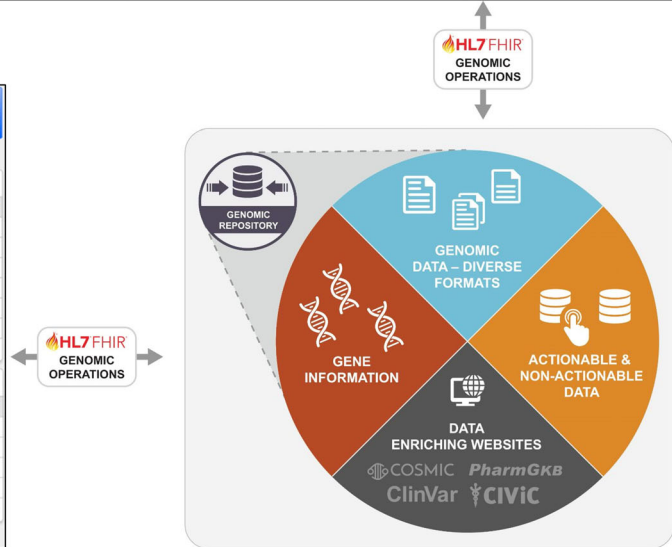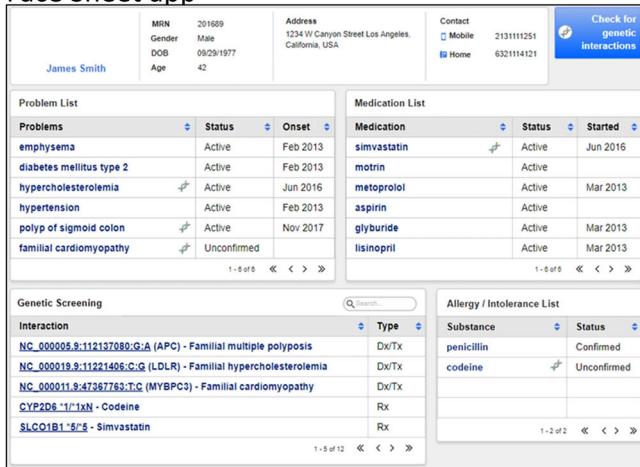
FIGURE 5 Proof-of-concept apps developed under Sync for Genes Phase 5. Apps are designed to address different use cases, and were chosen in part to help us examine the ability to selectively return only that subset of annotations relevant to the specific context of the app.
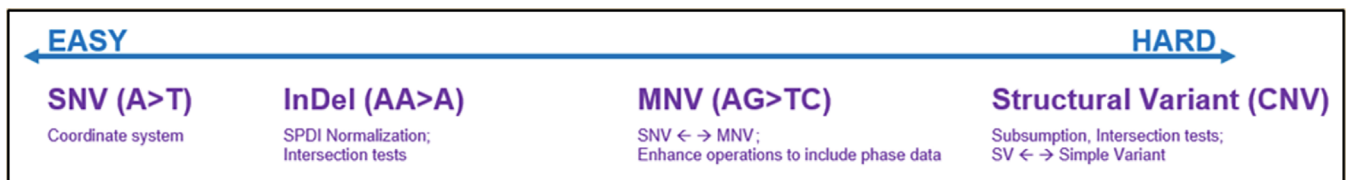

FIGURE 6 Dynamic annotation varies in complexity based on the variant type. Shown are four categories of variants, of increasing annotation complexity. Whereas it is relatively straightforward to automatically annotate single nucleotide variants, it is fairly complicated to annotate structural variants. ("A," "AA," "AG," "T," "TC": nucleotide base sequences; CNV: Copy Number Variant; InDel: insertion and/or deletion variant; MNV: multi-nucleotide variant; SNV: single nucleotide variant).

phase data to combine SNVs into MNVs where possible in order to compute additional potentially relevant annotations.

Dynamic annotation of structural variants such as duplications, large deletions, inversions, and copy number variants poses considerable challenges, in large part due to the imprecision of variant boundaries and the variability in boundaries across individual patients. Whereas with simple variants, the annotation process is essentially "look to see if this patient's variant is present in the knowledge base," for structural variants, it is unlikely to find an exact match of a patient's variant in a knowledge base. FHIR Genomics Operations include the "find-subject-structural-intersecting-variants" operation that determines if structural variants are present that overlap given ranges, and the "find-subject-structural-

subsuming-variants" operation that determines if structural variants are present that fully subsume a range. We also found that in some cases, structural variants with precise boundaries can be alternatively represented as simple variants, yielding additional annotations.

6 | CONCLUSIONS

Under Sync for Genes Phase 5, we have developed an open source pipeline for the sharing of dynamically annotated genomic information using GA4GH-encoded knowledge, delivered using FHIR Genomics Operations.

This project demonstrates the feasibility of an ecosystem where genomic knowledge bases have standardized knowledge (e.g., based on the GA4GH VA spec), and CDS applications can dynamically leverage that knowledge to provide real-time decision support, based on current knowledge, to clinicians at the point of care. As Dr. Friedman has described,[30] a Learning Health System improves "individual and population health by marrying discovery to implementation." Here, where updates to a knowledge base can be immediately translated into updated CDS, we have demonstrated a foundational tenet of a Learning Health System. A genomic data repository wrapped in a set of standardized APIs (e.g., FHIR Genomics Operations) enables us to manage a person's entire genome, manage evolution in our understanding of a person's genome, and provide up-to-date and contextually relevant genomics findings and recommendations at the point of care through a growing array of EHR integration options.

ACKNOWLEDGMENTS

FUNDING INFORMATION

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

- The most current Operations definitions are here (http://build.fhir.org/ig/HL7/genomics-reporting/operations.html).
- An open-source reference implementation of the HL7 FHIR Genomics Operations is here (https://github.com/FHIR/genomics-operations).
- A Swagger interface to the reference implementation is here (https://fhir-gen-ops.herokuapp.com/)

ORCID

Robert Dolin https://orcid.org/0000-0002-6235-6964
Rohan Gupta https://orcid.org/0000-0003-0770-8629
Lawrence Babb https://orcid.org/0000-0002-2455-2227

REFERENCES

1. Sync for Genes|HealthIT.gov. 2023. https://www.healthit.gov/topic/sync-genes
2. HL7.FHIR.UV. GENOMICS-REPORTING\Home Page – FHIR v4.0.1. 2022. http://build.fhir.org/ig/HL7/genomics-reporting/index.html
3. Austin-Tse CA, Jobanputra V, Perry DL, et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. *Npj Genomic Med*. 2022;7:1-13.
4. Chamala S, Majety S, Mishra SN, et al. Indispensability of clinical bioinformatics for effective implementation of genomic medicine in pathology laboratories. *ACI Open*. 2020;04:e167-e172.
5. Hiatt SM, Amaral MD, Bowling KM, et al. Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin Genet*. 2018;94:174-178.
6. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med off J Am Coll Med Genet*. 2017;19:249-255.
7. Thompson ML, Finnila CR, Bowling KM, et al. Genomic sequencing identifies secondary findings in a cohort of parent study participants. *Genet Med off J Am Coll Med Genet*. 2018;20:1635-1643.
8. Jang M-A, Lee S-H, Kim N, Ki CS. Frequency and spectrum of actionable pathogenic secondary findings in 196 Korean exomes. *Genet Med off J Am Coll Med Genet*. 2015;17:1007-1011.
9. Dorschner MO, Amendola LM, Turner EH, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet*. 2013;93:631-640.
10. NCBI. ClinVar gene-specific summary. 2016. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/2016/gene_specific_summary_2016-01.txt.gz
11. NCBI. ClinVar gene-specific summary. 2018. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/2018/gene_specific_summary_2018-01.txt.gz
12. Romero R, de la Fuente L, Del Pozo-Valero M, et al. An evaluation of pipelines for DNA variant detection can guide a reanalysis protocol to increase the diagnostic ratio of genetic diseases. *NPJ Genomic Med*. 2022;7:1-10.
13. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med*. 2017;19:209-214.
14. Li Q, Agrawal R, Schmitz-Abe K, et al. Reanalysis of clinical exome identifies the second variant in two individuals with recessive disorders. *Eur J Hum Genet*. 2023;31:1-4.
15. O'Daniel JM, McLaughlin HM, Amendola LM, et al. A survey of current practices for genomic sequencing test interpretation and reporting processes in US laboratories. *Genet Med off J Am Coll Med Genet*. 2017;19:575-582.
16. Bowdin S, Gilbert A, Bedoukian E, et al. Recommendations for the integration of genomics into clinical practice. *Genet Med off J Am Coll Med Genet*. 2016;18:1075-1084.
17. Clayton EW, Appelbaum PS, Chung WK, Marchant GE, Roberts JL, Evans BJ. Does the law require reinterpretation and return of revised genomic results? *Genet Med off J Am Coll Med Genet*. 2021;23:833-836.
18. ClinVar. 2023. https://www.ncbi.nlm.nih.gov/clinvar/
19. CIViC – Clinical Interpretation of Variants in Cancer. 2023. https://civicdb.org/
20. PharmGKB. 2023. https://www.pharmgkb.org/
21. Dolin RH, Heale BSE, Alterovitz G, et al. Introducing HL7 FHIR genomics operations: a developer-friendly approach to genomics-EHR integration. *J Am Med Inform Assoc*. 2022;30:246-493.
22. HL7 FHIR. Genomics Operations – Reference Implementation. 2023. https://github.com/FHIR/genomics-operations
23. Holmes JB, Moyer E, Phan L, Maglott D, Kattman B. SPDI: data model for variants and applications at NCBI. *Bioinformatics*. 2020;36:1902-1907.
24. Dolin RH, Gothi SR, Boxwala A, et al. vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration. *BMC Bioinformatics*. 2021;22:104.
25. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 2012;6:80-92.

26. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat*. 2022;43:1012-1030.

27. Srinivasan S, Kalinava N, Aldana R, et al. Misannotated multinucleotide variants in public cancer genomics datasets Lead to inaccurate mutation calls with significant implications. *Cancer Res*. 2021; 81:282-288.

28. Misannotation of multiple-nucleotide|Wellcome Open Research. 2023. https://wellcomeopenresearch.org/articles/4-145/v2

29. Wang Q, Pierce-Hoffman E, Cummings BB, et al. Landscape of multinucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat Commun*. 2020;11:2539.

30. Friedman CP. What is unique about learning health systems? *Learn Health Syst*. 2022;6:e10328.