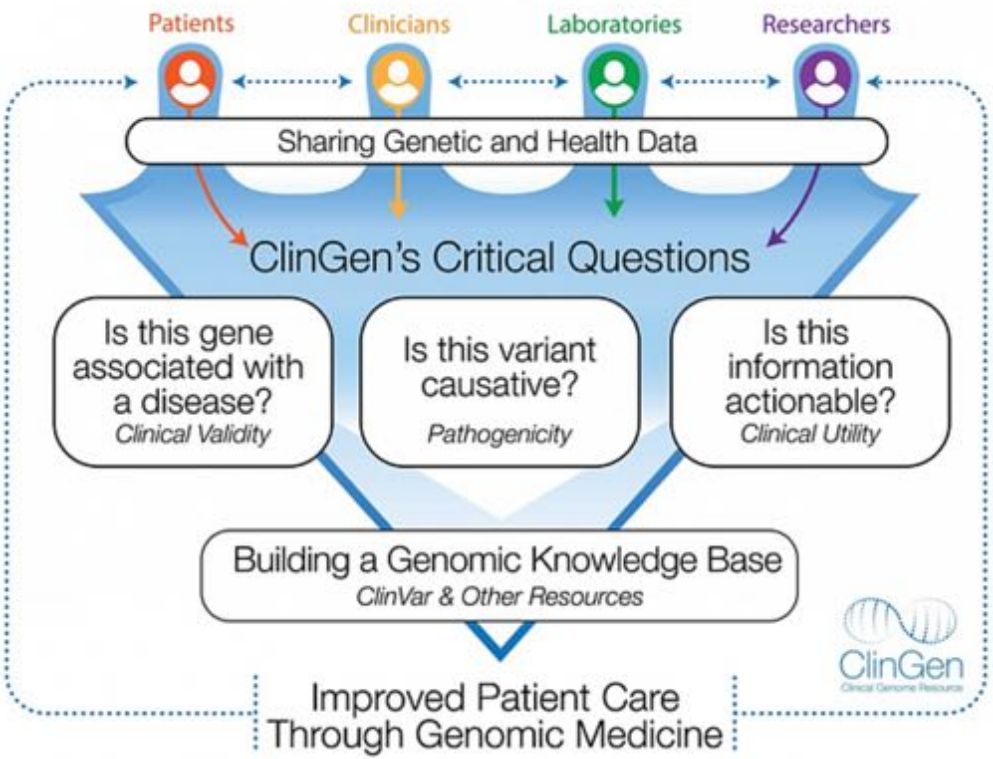
ClinGen Linked Data Hub

Aleks Milosavljevic
Baylor College of Medicine

Curating the Clinical Genome
Washington, DC May 30th, 2019

Clinical Genome Resource (ClinGen)



Clinical Genome Resource (ClinGen)

U.S. FOOD & DRUG ADMINISTRATION

Search | Menu

FDA NEWS RELEASE

FDA takes new action to advance the development of reliable and beneficial genetic tests that can improve patient care

Share | Tweet | LinkedIn | Email | Print

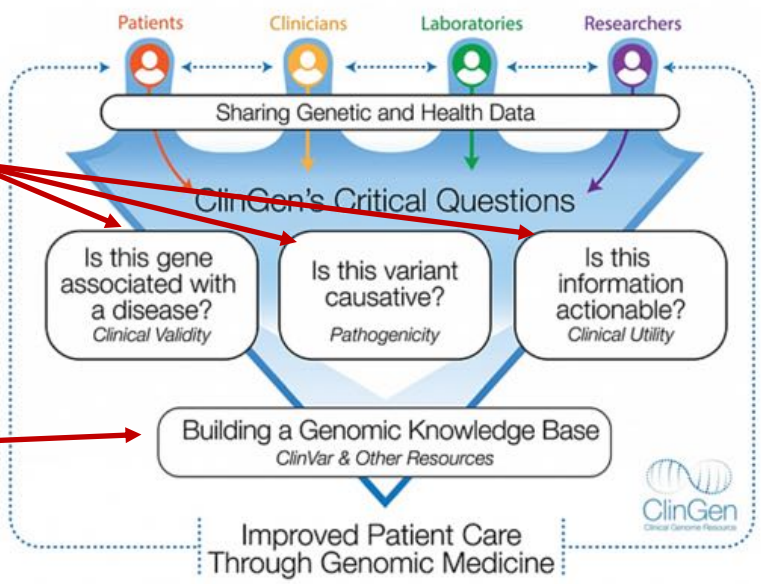
diagnose genetic diseases and guide medical treatments. For the first time, the agency has formally recognized a public database that contains information about genes, genetic variants and their relationship to disease. The FDA is recognizing the genetic variant information in the Clinical Genome Resource (ClinGen) consortium's ClinGen Expert Curated Human Genetic Data, which is funded by the National Institutes of Health (NIH) as a source of valid scientific evidence that can be used to support clinical validity in premarket submissions. The information contained in this open database has been

Clinical Genome Resource (ClinGen)

Data integration
to feed
knowledge curation

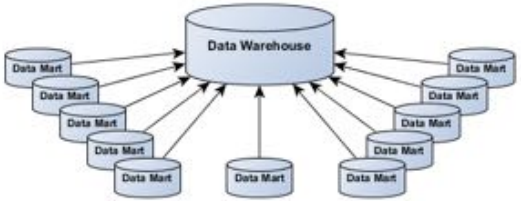
and

to provide
transparency
(evidential support for conclusions)



Two approaches to data integration: Data Warehousing vs. Data Linking

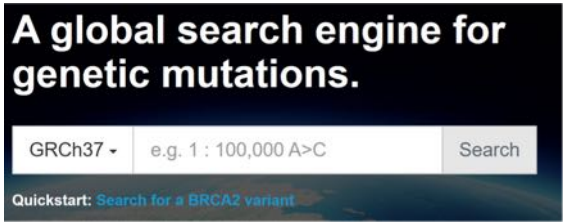
Data Warehousing



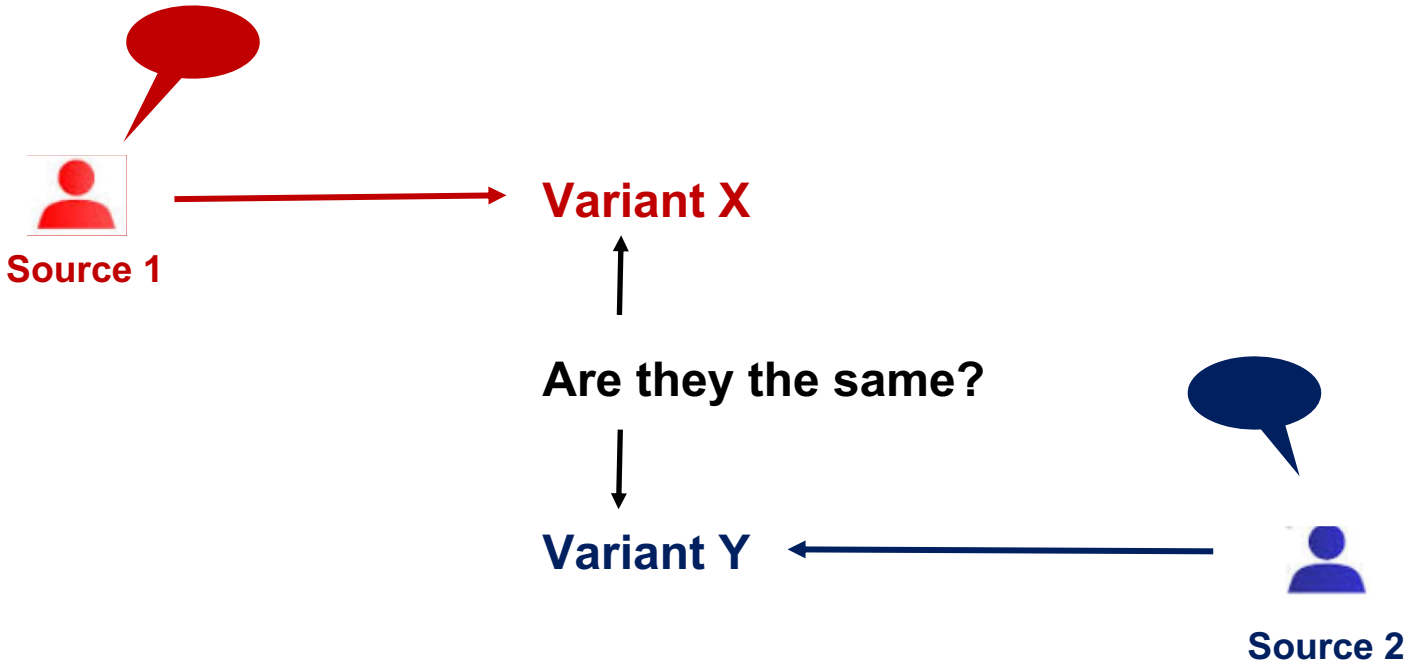
Example: MyVariant.info

Data Linking

Example: GA4GH Beacon



Data Linking: the “deduplication” problem



ClinGen Allele Registry

ClinGen Allele Registry provides deduplication as a web service:

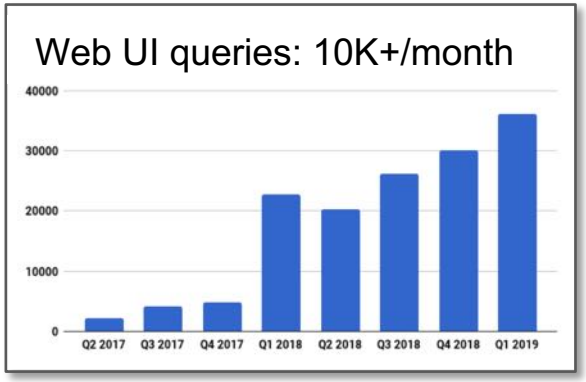
- Unique variant IDs
  (“CAid” + dereferenceable URI)

- Fast query

- High-bandwidth registration of new variants
  (many millions per request)

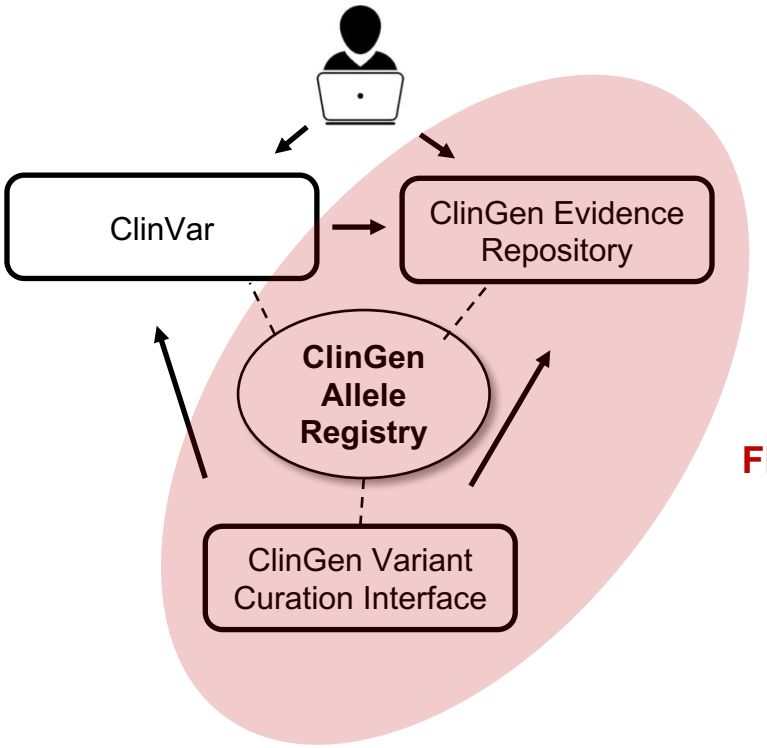
- Accessibility: Web UI + REST API

900M+ variants registered



Pawliczek et al. ***Human Mutation***. 2018;39:1690–1701.

ClinGen Allele Registry



FDA recognized "database"

Pawliczek et al. ***Human Mutation***. 2018;39:1690–1701.

ClinGen Evidence Repository (ERepo)

Link to SEPIO compliant JSON-LD document

Guideline: *ACMG variant classification (PAH)*

NM_000277.2(PAH):c.1A>G (p.Met1Val)

CA114360

586 (ClinVar)

Gene: PAH
Condition: phenylketonuria
Inheritance Mode: Autosomal recessive inheritance
Link to MONDO

HGVS expressions

NM_000277.2:c.1A>G
NM_000277.1:c.1A>G
NM_001354304.1:c.1A>G
NM_00
ENST00
ENST00

Met criteria codes

PP4_Moderate 0 ✓ Seen in PKU patients. BH4 disorders ruled out.

Analysis of RFLP haplotypes and mutations revealed a novel mutation, an A-to-G transition (met----val) in codon 1 (the translation-initiation codon). It occurred on 5 of the 18 mutant chromosomes. A proband homozygous for this mutation had the PKU phenotype. The appropriate biochemical tests were done to rule out disorders of tetrahydrobiopterin homeostasis. PubMed

Pathogenic

Met criteria codes 4
PP4_Moderate PVS1 PM2 PS3

Evidence submitted by expert panel

PAH VCEP

PAH-specific ACMG/AMP criteria applied: PVS1: Initiation codon variant; PM2: gnomAD MAF=0.00002; PP4_Moderate: Seen in PKU patients. BH4 disorders ruled out. (PMID:2574002); PS3: <3% (PMID:9450897). In summary this variant meets criteria to be classified as pathogenic for phenylketonuria in an autosomal recessive manner based on the ACMG/AMP criteria applied as specified by the PAH Expert Panel: (PVS1, PM2, PP4_Moderate, PS3).

Met criteria codes

PP4_Moderate 0 ✓ Seen in PKU patients. BH4 disorders ruled out.

Analysis of RFLP haplotypes and mutations revealed a novel mutation, an A-to-G transition (met----val) in codon 1 (the translation-initiation codon). It occurred on 5 of the 18 mutant chromosomes. A proband homozygous for this mutation had the PKU phenotype. The appropriate biochemical tests were done to rule out disorders of tetrahydrobiopterin homeostasis. PubMed

PVS1 0 ✓ Initiation codon variant

PM2 0 ✓ gnomAD MAF=0.00002

PS3 0 ✓ <3%

< 3 PAH enzyme activity as % of wild type. (p91023(B)/COS PubMed

Approved on: 2018-08-07
Published on: 2018-12-10

ClinGen Allele Registry

ClinGen
Clinical Genome Resource

ClinGen Allele Registry

Allele Registry | Pathogenicity Calculator | Login | Forgot Password?

Canonical Allele Identifier: CA000559

Gene: PTEN HGNC NCBI

Identifiers and link-outs to other resources

ClinVar Variation Id: 142269

ClinVar RCV Id: RCV000131292 RCV000212883 RCV000613031 RCV000660237

dbSNP Id: rs587782350

COSMIC: COSM5111

MyVariant Identifiers: chr10:g.89717712C>T (hg19) chr10:g.87957955C>T (hg38)

User contributed link-outs

PubMed: PMID:10076877 PMID:10400993 PMID:12938083 PMID:14566704 PMID:17942903 PMID:21628076 PMID:22261759 PMID:22595938 PMID:23161105 PMID:23934111 PMID:24375884 PMID:24744697 PMID:26681312 PMID:27477328

ERepo: CA000559/MONDO:0017623/003

134,000+ PMID links

Calculator JSON-LD

ClinGen Allele Registry

ClinGen
Clinical Genome Resource

ClinGen Allele Registry

Allele Registry | Pathogenicity Calculator | Login | Forgot Password?

Canonical Allele Identifier: CA000559

Gene: PTEN

Identifiers and link-outs to other resources ← **"Core" links**

ClinVar Variation Id: 142269

ClinVar RCV Id: RCV000131292 RCV000212883 RCV000613031 RCV000660237

dbSNP Id: rs587782350

COSMIC: COSM5111

MyVariant Identifiers: chr10:g.89717712C>T (hg19) chr10:g.87957955C>T (hg38)

User contributed link-outs ← **"Crowdsourced" links**

PubMed: PMID:10076877 PMID:10400993 PMID:12938083 PMID:14566704 PMID:17942903 PMID:21628076 PMID:22261759 PMID:22595938 PMID:23161105 PMID:23934111 PMID:24375884 PMID:24744697 PMID:26681312 PMID:27477328

← **134,000+ PMID links**

ERepo: CA000559/MONDO:0017623/003

API-based submission of links to external databases
CIViC: 970

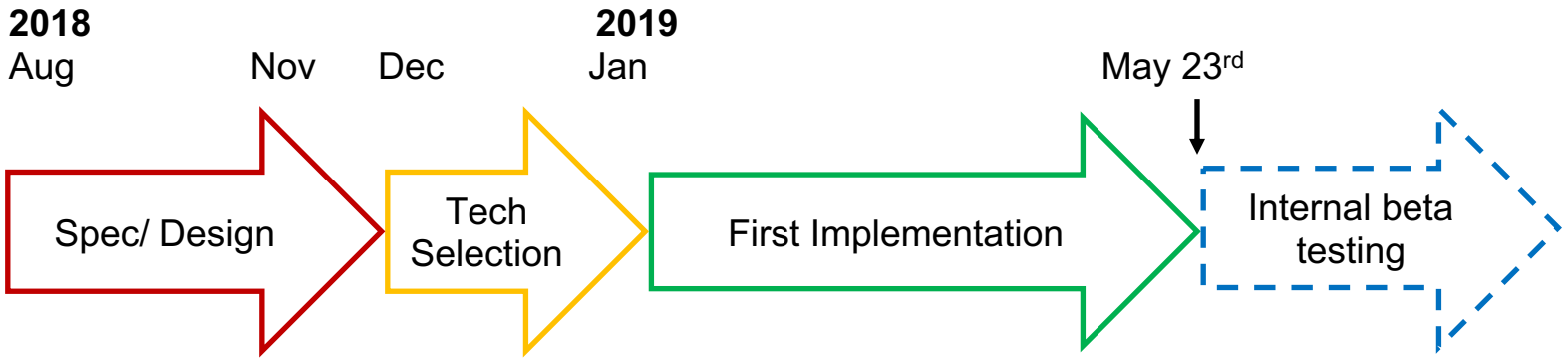
Calculator

JSON-LD

Linking: the next batch of problems

- **PMID** does not point to **specific relevant piece of text** that may be identified through “Community Curation”
- The **text is not structured** (machine-readable) and controlled vocabularies/ ontologies (e.g., disease ontology) are not used
- The link does not identify **type of evidence** making it **hard to “target” information** to the appropriate curation context (functional, allele frequency, etc.)
- In addition, **links to database entries** do not constitute **permanent reference** to support conclusions about pathogenicity of variants
- **Managing updates** to database entries across the web (e.g., allele frequency)

Linked Data Hub (LDH): development timeline



Linked Data Hub (LDH): technology selection

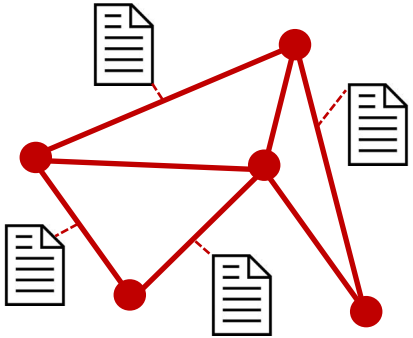
W3C Web Annotation Model

- Web-wide standard data model and format defined in February 2017
- Defines “envelope” for exchange of “annotation” about any “sequential” or “multi-dimensional” data type: web pages, PDFs, maps, … and also anything identified by a web link (genetic variant!)
- Useful for representing information about web links (LDH entries!)

Linked Data Hub (LDH): technology selection

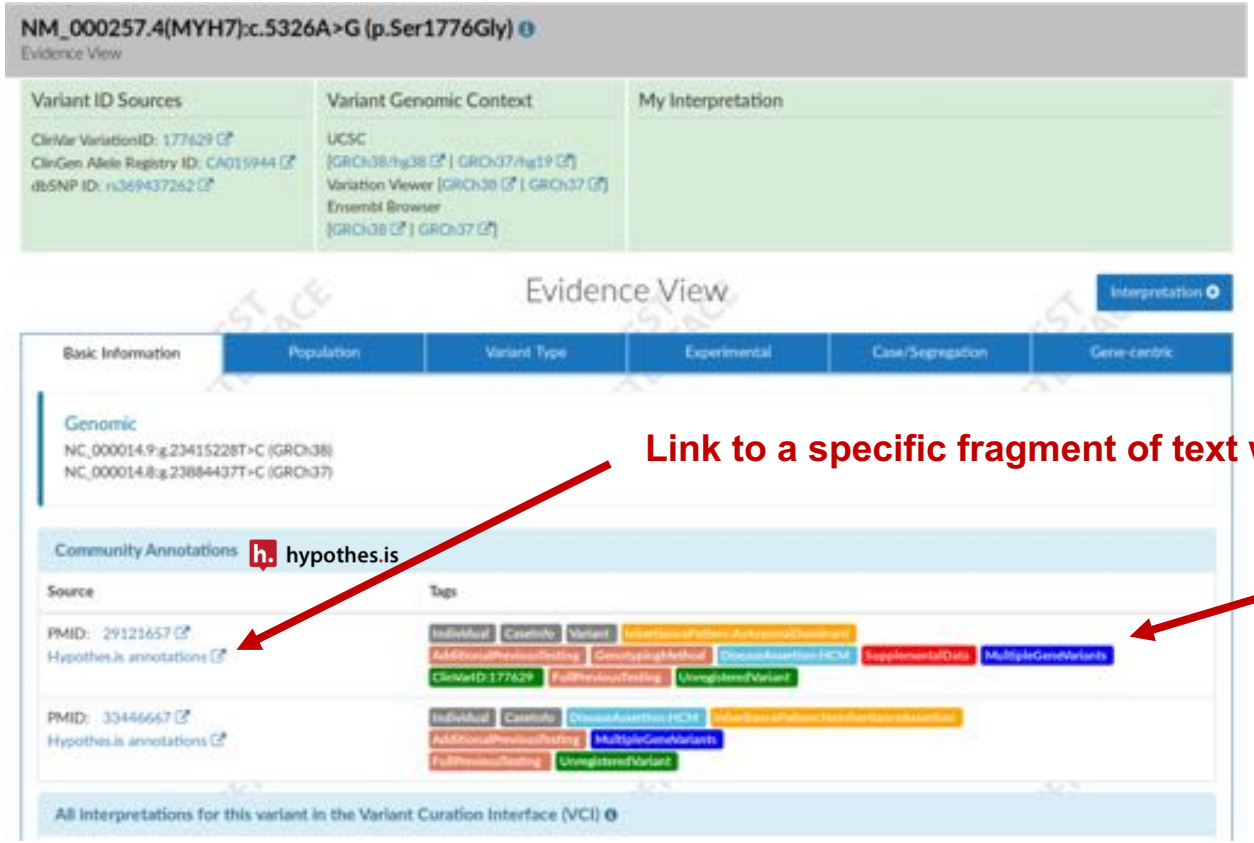
ArangoDB:

- Latest generation noSQL (**document** + **graph**) database
- Extremely scalable
- Cross-cloud, open source
- Full support for data linking



LDH use case 1: Community Curation

Linking output of Community Curation using Hypothes.is into ClinGen Variant Curation Interface



Link to a specific fragment of text within the article PDF

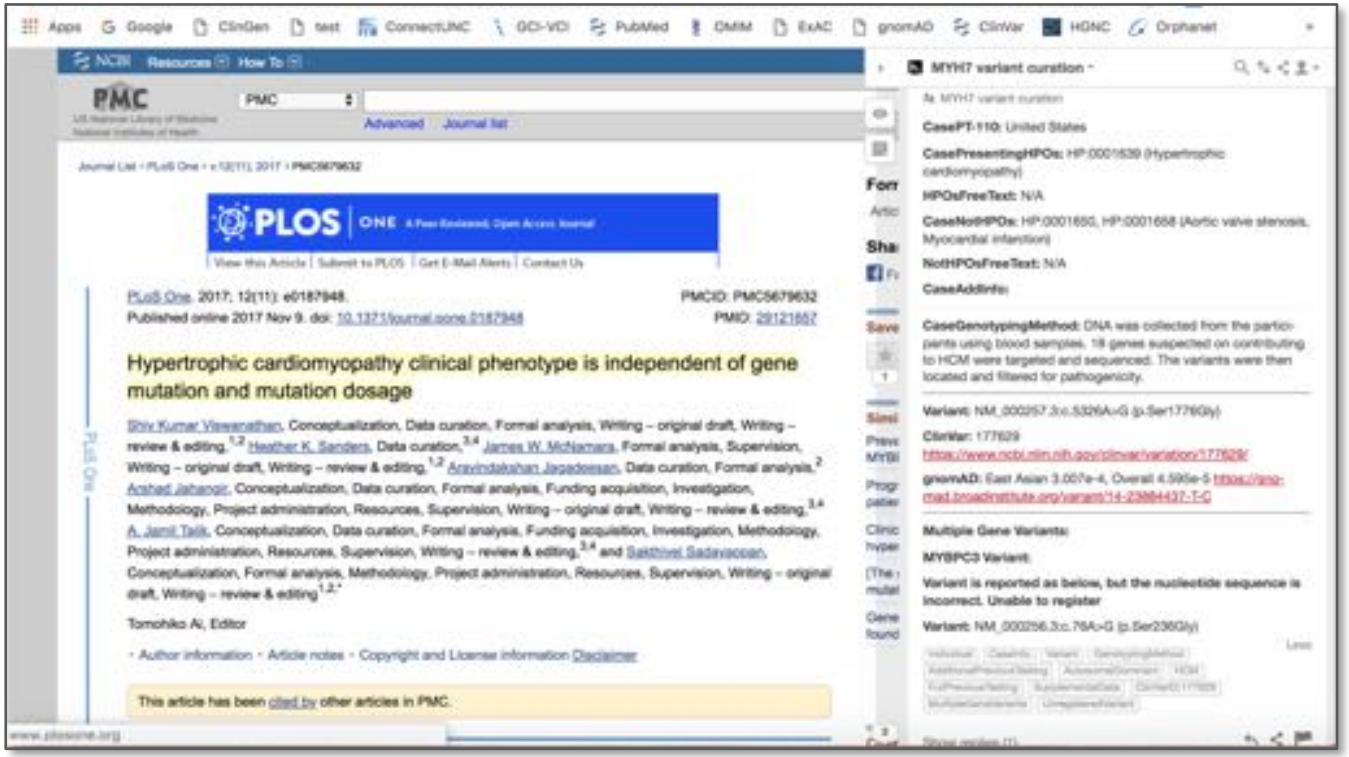
Tags annotating specific fragments of text

Also: variant identifier, disease ontology terms, other structured information

LDH use case 1: Community Curation

Linking output of Community Curation using Hypothes.is into ClinGen Variant Curation Interface

Click-through to Hypothes.is annotations:



LDH use case 2: ClinGen Variant Curation Interface

Delivery of structured “functional effect” data from literature curation.

NM_000527.4(LDLR):c.-152C>T

Evidence View

Evidence View

Interpretation

Basic Information | Population | Variant Type | Experimental | Case/Segregation | Gene-centric

Hotspot or functional domain

Curated Literature Evidence (Hotspot or functional domain)

No evidence added.

Experimental Studies

Structured Narrative of Functional Impact

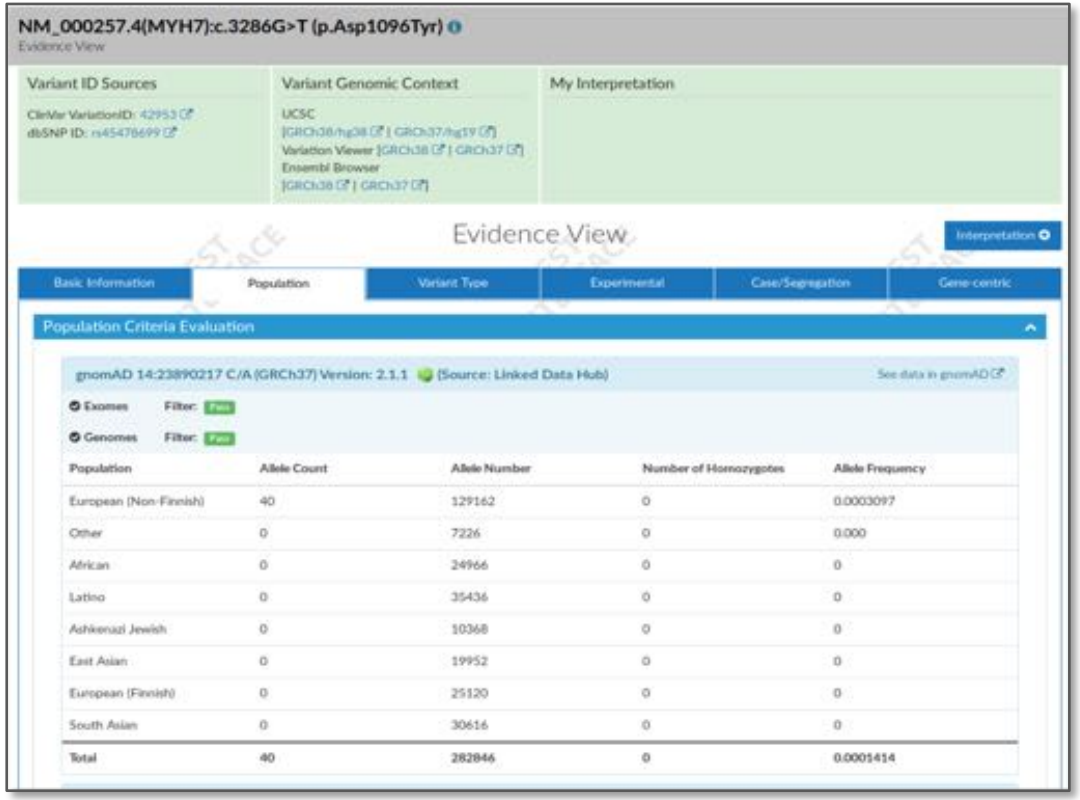
Source: PMID:10484771
Sholtz CL et al. (1999) "Mutation -59c–>t in repeat 2 of the LDL receptor promoter: reduction in transcriptional activity and possible allelic interaction in a South African family with familial hypercholesterolaemia." Hum Mol Genet Oct; 8(11):2025-30.

Experiment 1
Method: Luciferase reporter gene assay
Material(s): Hep G2 cell
Functional impact(s): 40% gene expression
Functional impact comments: none
Experimental repeats: 3
Experimental controls: Hep G2 cells with wild-type *LDLR* vector
Validation controls: Hep G2 cells with known pathogenic promoter variant *LDLR*:c.-337C>T
Statistical analysis: none
Additional notes: none

Contributor: Familial Hypercholesterolemia VCEP

LDH use case 3: Variant Curation Interface

Targeted delivery of data from allele frequency databases for variant curation.



Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Non-Finnish)	40	129162	0	0.0003097
Other	0	7226	0	0.000
African	0	24966	0	0
Latino	0	35436	0	0
Ashkenazi Jewish	0	10368	0	0
East Asian	0	19952	0	0
European (Finnish)	0	25120	0	0
South Asian	0	30616	0	0
Total	40	282846	0	0.0001414

LDH use case 4: Evidence Repository

Structured data as supporting evidence for pathogenicity assessments.

ClinGen
Clinical Genome Resource

EVIDENCE REPOSITORY

Link to SEPIO compliant JSON-LD document

Guideline: *MYH7-associated inherited cardiomyopathies - adapted from ACMG/AMP*

Molecular Consequence

ClinVar Submission

Community Annotation

InSilico Prediction Score

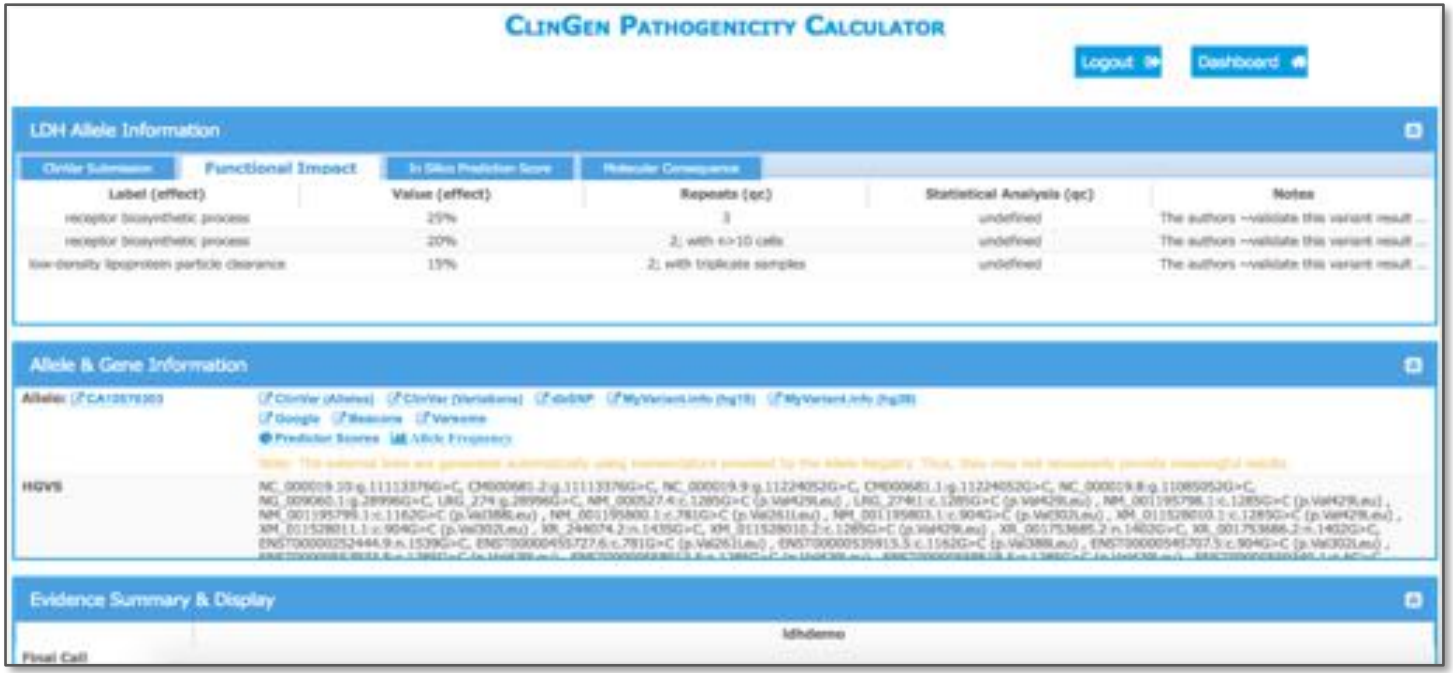
Population Frequency

Allele Count	Allele Frequency	Allele Number	Homozygous Allele Individual Count	Method	Population
3	0.000196027	15304	0	gnomad-exome	afr
2	5.95557e-05	33582	0	gnomad-exome	amr
0	0	9850	0	gnomad-exome	asj
2	0.000115955	17248	0	gnomad-exome	eas

LDH use case 5: ClinGen Pathogenicity Calculator

Supporting evidence in LDH is reusable by other variant curation tools.

LDH as a “PubMed” for variant data.



Acknowledgements

Software development team at Baylor College of Medicine

Andrew R. Jackson

Sai Lakshmi Subramanian

Neethu Shah

Sameer Paithankar

Ronak Patel

Sharon Plon
Baylor PI

Carlos Bustamante
Stanford PI

Software development team at Stanford

Helio Costa

Matt Wright

Bryan Wulf

Christine Preston

UNC

Courtney Thaxton
UNC

Broad Institute

Larry Babb