ClinGen
Clinical Genome Resource

Baylor College of Medicine

The ClinGen Allele Registry, Linked Data Hub, and Evidence Repository

H3A-RDWG / ClinGen Workshop

Kevin Riehle – Lead Product Owner, ClinGen

Bioinformatics Research Laboratory, Baylor College of Medicine

2-5-21

Overview

- Introduction
- Role of ClinGen
- Tools – Overview, solution, and live demo
  - ClinGen Allele Registry (CAR)
    - Variant naming
    - Short variants
    - Structural variants
  - Linked Data Hub (LDH)
    - Existing and planned external databases
  - Evidence Repository (ERepo)
    - Requirements for FDA recognition
- Acknowledgements
- Summary

ClinGen
Clinical Genome Resource

Defining the Problem – Genetic Variation

Baylor College of Medicine

- Goal of whole genome sequencing is to interrogate all forms of variation
  - Single nucleotide variants (SNVs), small insertions / deletions (indels), and structural variants (SVs)
  - 1000 Genomes Project (2008 – 2015) observed 4.1 – 5.0M sites affecting 20M bases of sequence
    - Mostly SNVs and indels (>99%)
    - Also, ~2,500 SVs per genome
- Yields per average genome more refined with recent (2019-2020) studies
  - ~800,000 short variants (<50bp) "usable" from low millions of sites
  - ~27,000 structural variants (>50bp)

https://doi.org/10.1038/s41467-018-08148-z https://doi.org/10.1038/s41586-020-2371-0 https://www.internationalgenome.org/

ClinGen
Clinical Genome Resource

What Questions Are Being Asked?

Baylor College of Medicine

Can I see all the curation evidence that exists for variant interpretation? What about for the curation of genes?

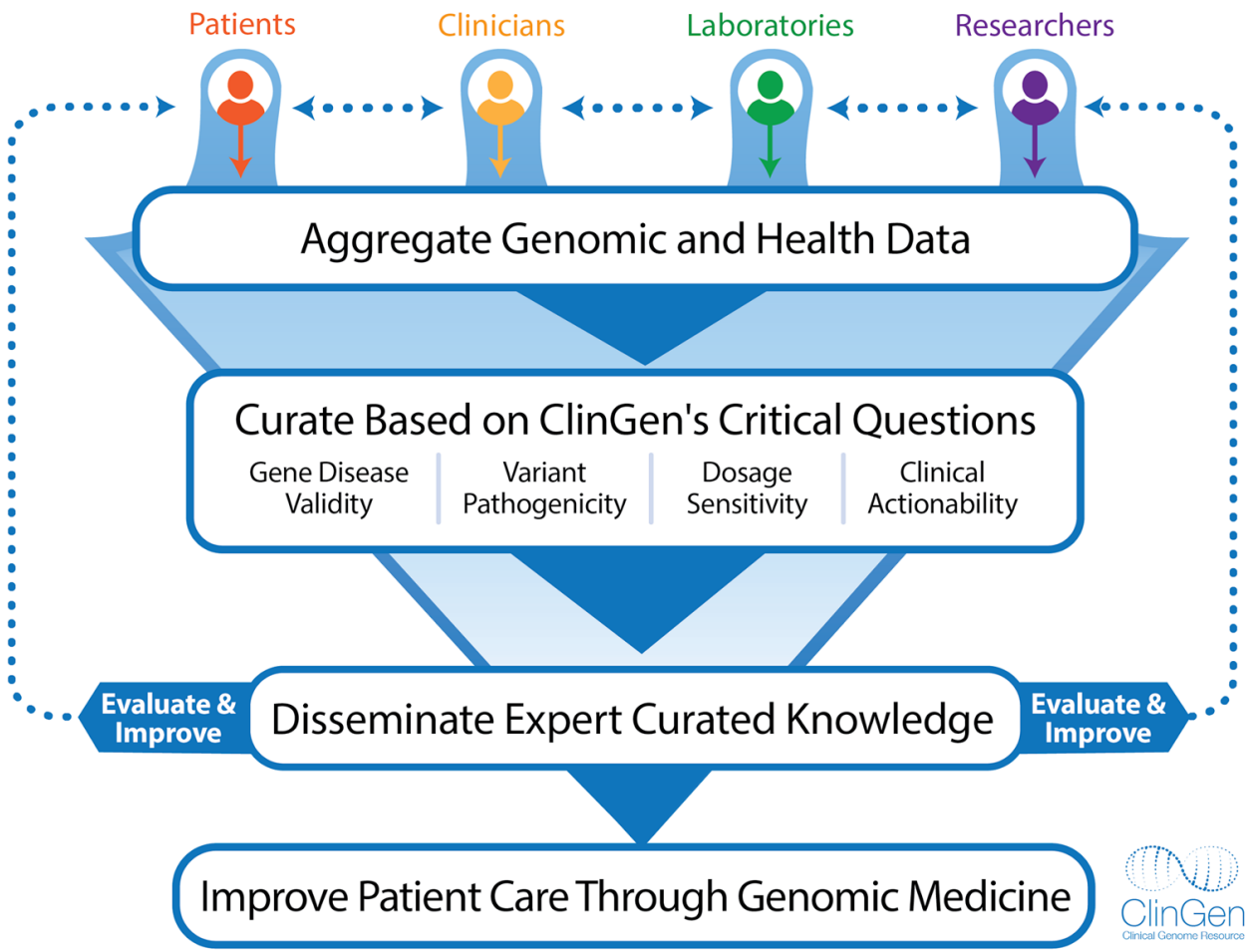
Is there a way to know when the curated evidence changes?

Is there a single service that aggregates data from 3rd party databases in a standardized format for integration on my side?

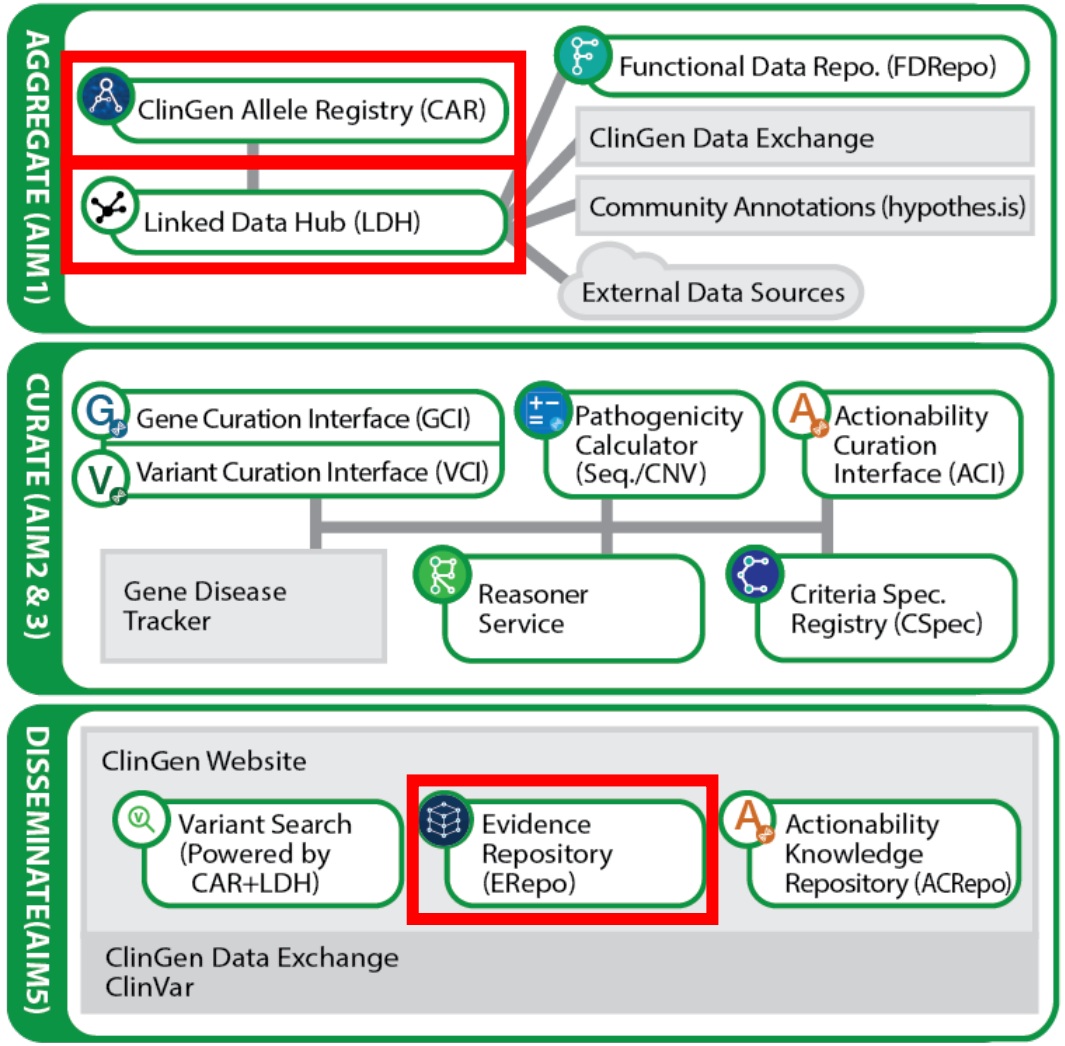
I see so many different identifiers and ways to represent the same variant, how do I know if I'm talking about the same thing?

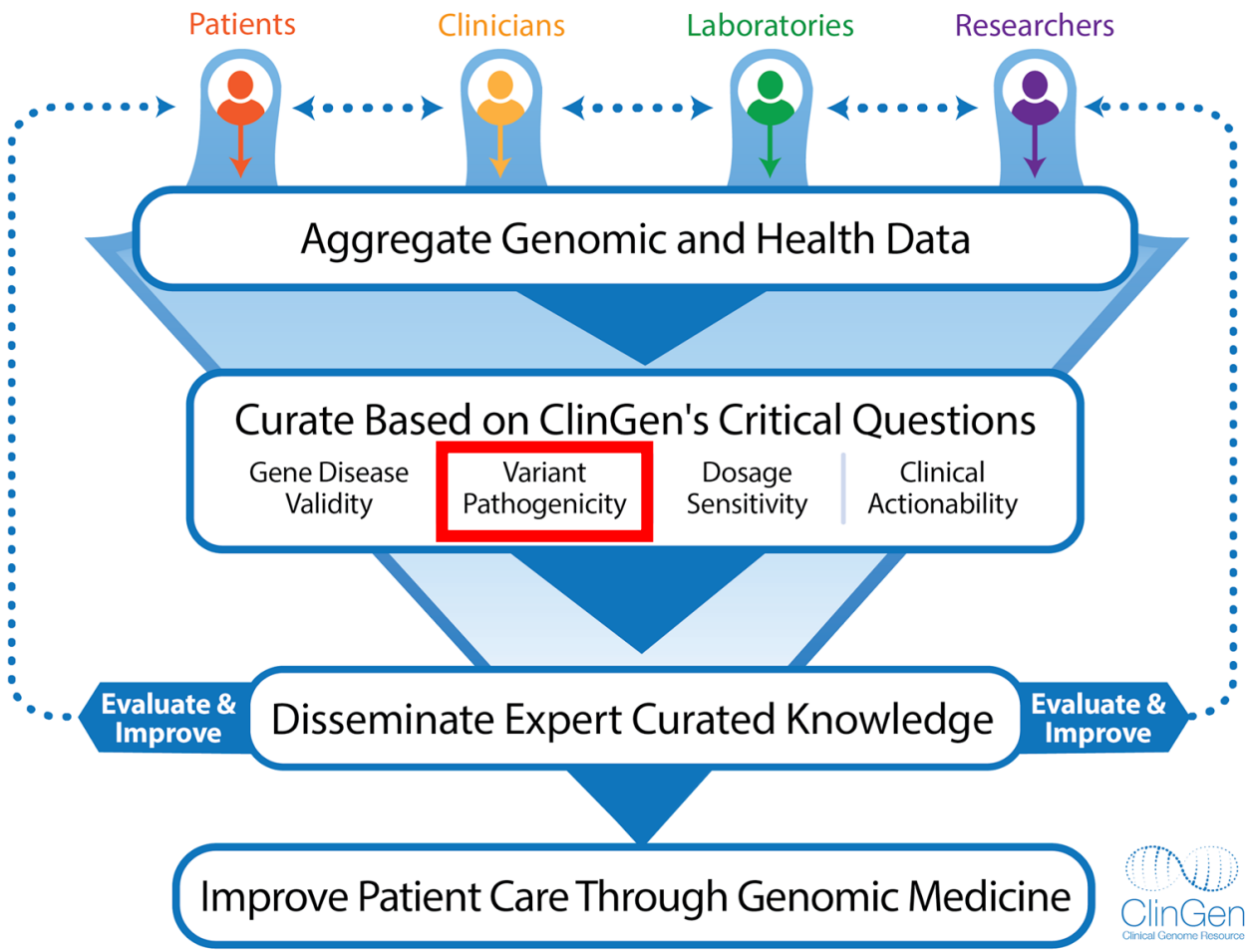
Defining the Problem - ClinGen

- NIH developed the Clinical Genome Resource Program (ClinGen) to address growing needs
  - Enabling clinical testing labs, researchers, patients, and providers to share data associated with systematic implementation of standards for clinically relevant genome variation
    - Genetic test reporting
    - Variant classification
    - Data exchange

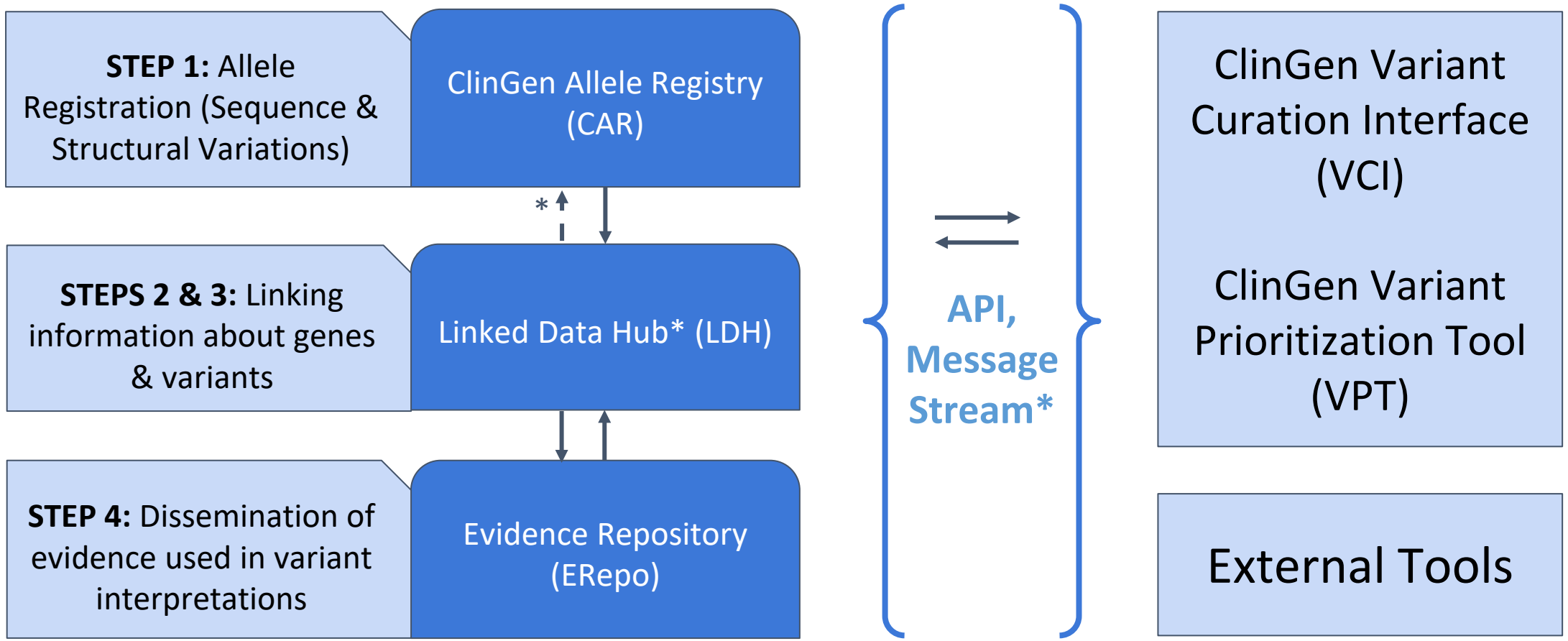


Aggregation, Curation, and Dissemination





Software to Facilitate Variant Interpretation



* Active development

FAIR Principles for Open Computable Data Sharing

ClinGen Allele Registry (CAR)

Linked Data Hub* (LDH)

Evidence Repository (ERepo)

Findable

Accessible

Interoperable

Reusable

Findable	Unique and persistent identifiers, rich metadata, searchable
Accessible	REST-APIs, Web UIs, Message Streams, SEPIO/GA4GH data standards
Interoperable	JSON/JSON-LD, standard ontologies, links to relevant data/metadata
Reusable	Track provenance, open source, microservice design

* Active development

ClinGen Allele Registry – Proposal

Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

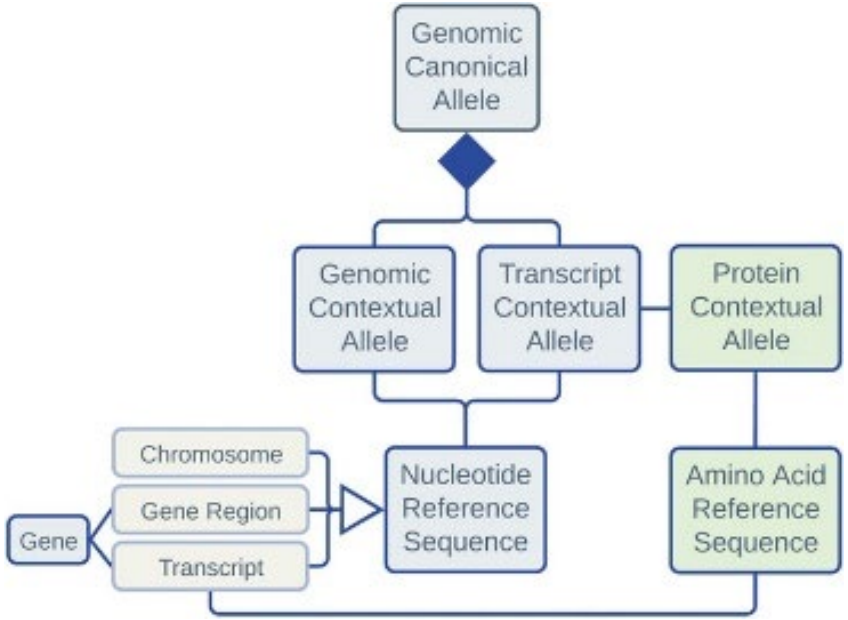
INFORMATICS | Open Access

ClinGen Allele Registry links information about genetic variants

Piotr Pawliczek, Ronak Y. Patel, Lillian R. Ashmore, Andrew R. Jackson, Chris Bizon, Tristan Nelson, Bradford Powell, Robert R. Freimuth, Natasha Strande, Neethu Shah, Sameer Paithankar, Matt W. Wright, Selina Dwight, Jimmy Zhen, Melissa Landrum, Peter McGarvey, Larry Babb, Sharon E. Plon, Aleksandar Milosavljevic, on behalf of the Clinical Genome (ClinGen) Resource ... See fewer authors

First published: 11 October 2018 | https://doi.org/10.1002/humu.23637 | Cited by: 1

For the ClinGen/ClinVar Special Issue



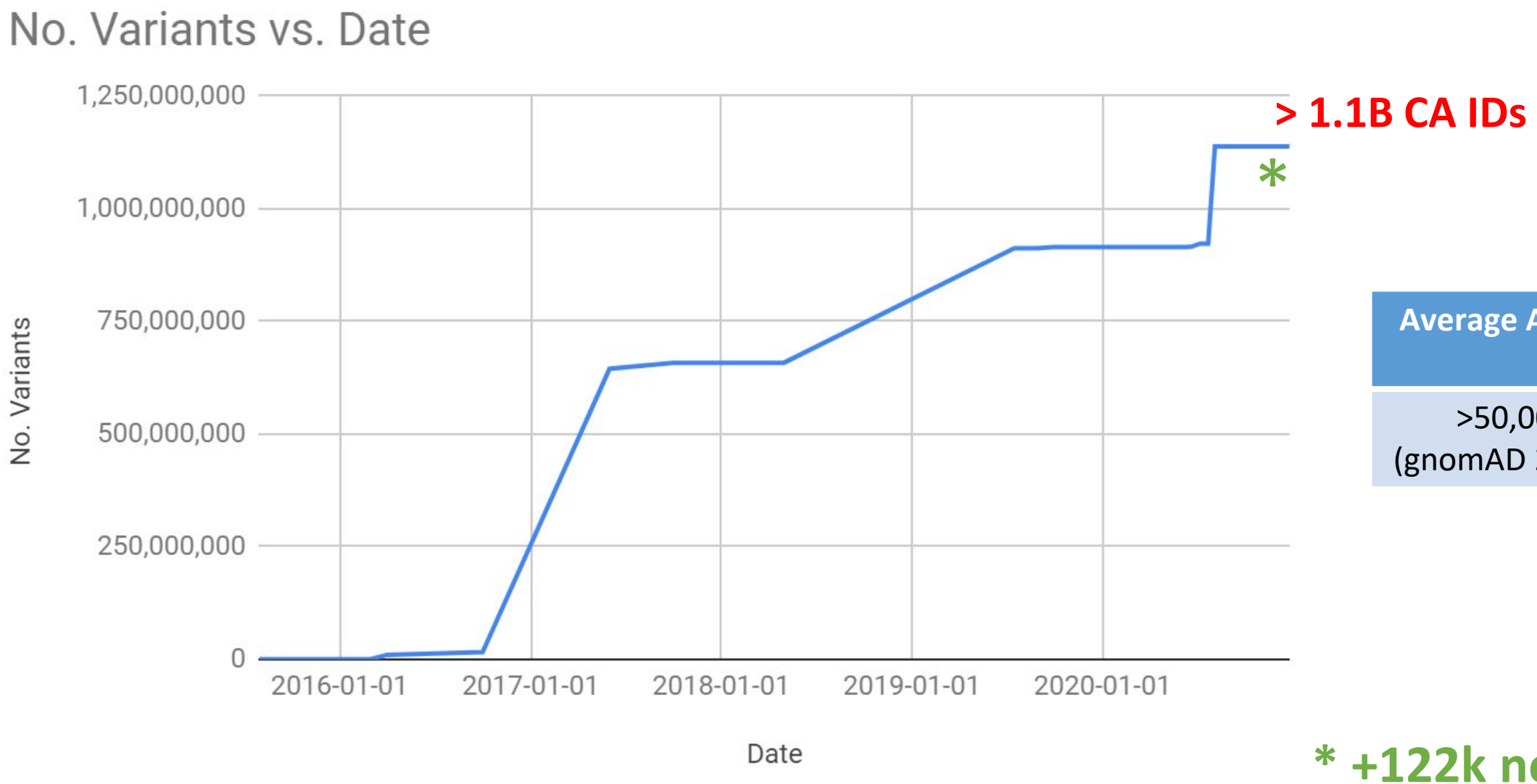
Genomic Alleles

HGVS	Genome Assembly
NC_000011.10:g.68032291C>T , CM000673.2:g.68032291C>T	GRCh38
NC_000011.9:g.67799758C>T , CM000673.1:g.67799758C>T	GRCh37
NC_000011.8:g.67556334C>T	NCBI36

ClinGen Allele Registry – Solution

- ClinGen Allele Registry aims to support
  - Stable and consistent variant identification and naming (curation / pathogenicity)
    - FDA requirement for classification process
    - CA IDs (e.g. CA321211)
  - Open sharing of genetic variant information
  - Aggregation of information from different sources
  - FAIR principles
    - Findable, Accessible, Interoperable, and Reusable
  - Identifiers on demand (high bandwidth, low latency)
    - ClinVar, dbSNP, etc. issue variant identifiers after completion of a robust process
  - API support for programmatic registration and access
    - Supporting large batch queries and registrations
    - JSON output for programmatic integration
  - Free, open user registration

Registered Variants Growth

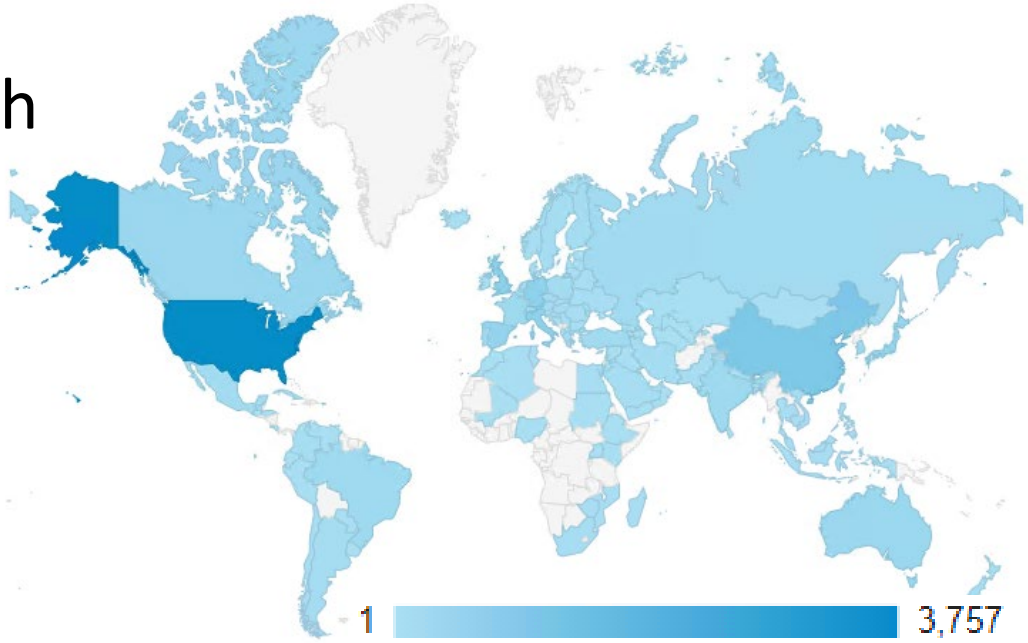


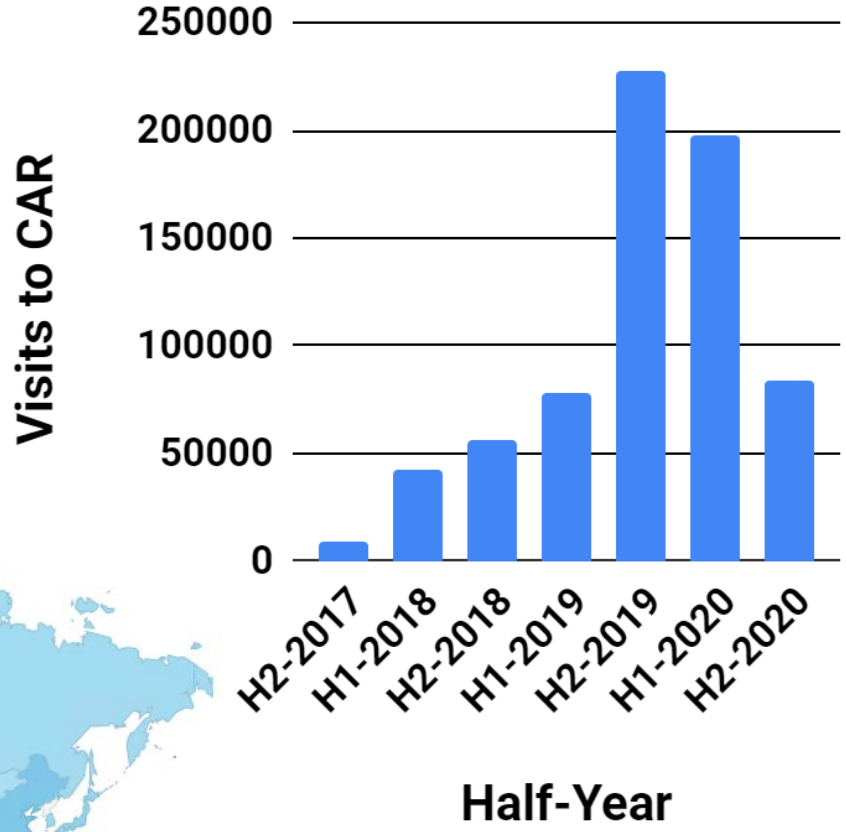
Average API Registration Rate
>50,000 / second (gnomAD 2.0 => ~7 hours)

*** +122k new variants**

Registry Usage – Users and Frequency

- >20 unique users register canonical alleles / month
- > 250k visits to landing page / month
- >14k UI uses / month
  - > 100 countries
- >1M API uses / month
  - > 25 countries





ClinGen
Clinical Genome Resource

Baylor College of Medicine

Registry Usage

- CA ID adoption
  - Databases
    - ClinVar, CIViC, myVariant.info, Ensembl, Database of Pathogenic Variants, etc.
  - Tools
    - Linked Data Hub, Variant Curation Interface, Evidence Repository, Functional Data Repository, etc.
  - Data aggregation
    - AllofUs clinical laboratories to identify variants classified by >= 2 laboratories
    - Variant Interpretation in Cancer Consortium (VICC) integration across six prominent somatic cancer variant knowledgebases to register variants and aggregate variant interpretations
    - Australian Genomics platform Shariant to facilitate the sharing of clinically curated variants and evidence among multiple diagnostic laboratories

Reference Databases

- References
  - NCBI36, GRCh37, GRCh38
- Transcripts
  - >700,000 (>450k RefSeq, >250k Ensembl)
- Update schedule
  - Monthly release schedule sourced from ClinVar
- MANE Select
  - Matched Annotations from NCBI and EMBL-EBI

ClinGen
Clinical Genome Resource

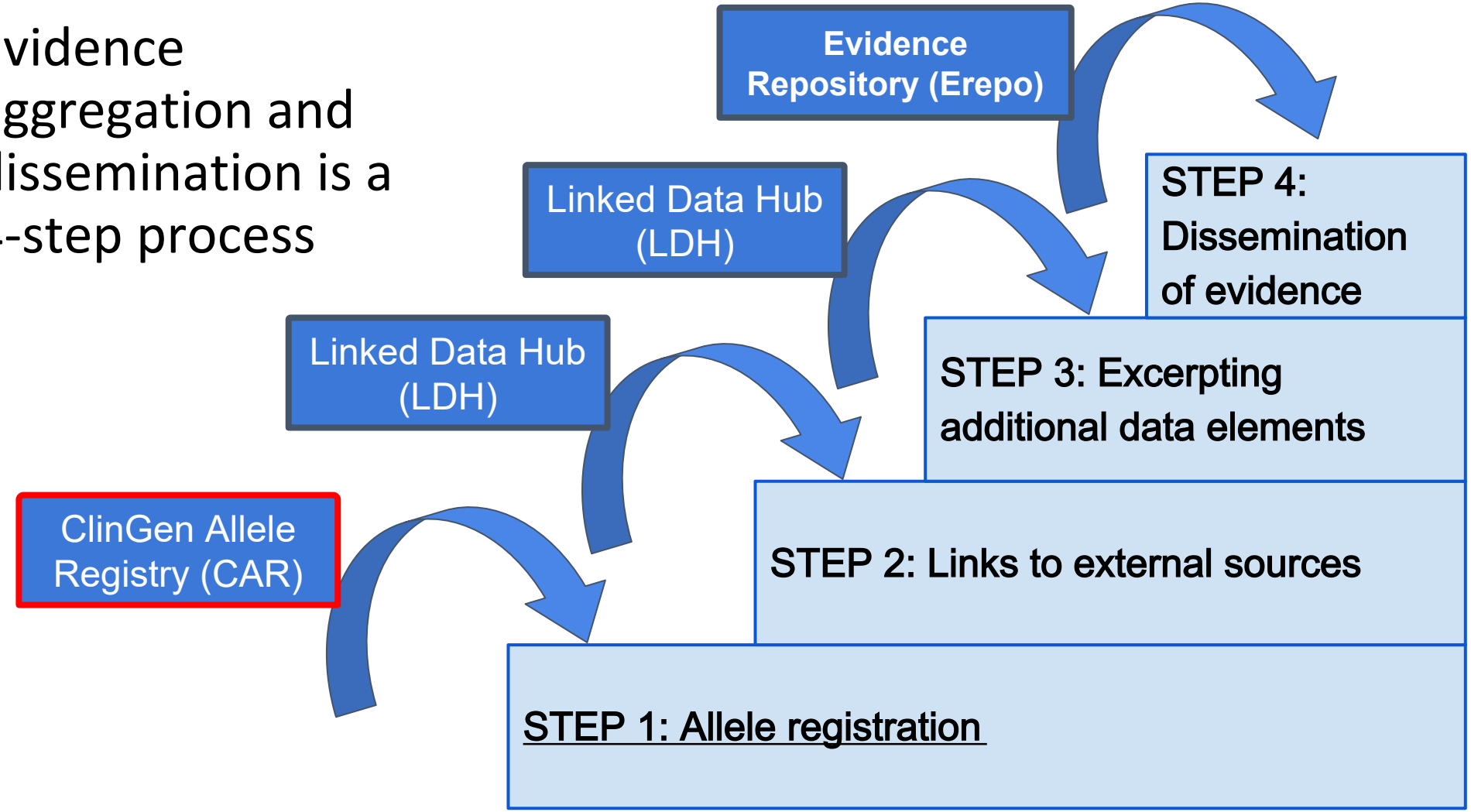
Variant Pathogenicity Interpretation

Baylor College of Medicine

- Variant curation challenges
  - *Variant identity* – same variant is defined differently in different sources
  - Diversity of evidence sources
  - Lack of standard ways to accumulate evidence
  - Tracking of provenance of evidence used in interpretation
  - Transparency and auditability – key FDA requirements
  - Alerting users / curators to knowledge / content changes

Variant Pathogenicity Interpretation – Registration

- Evidence aggregation and dissemination is a 4-step process



Registered CAR Variants with Canonical Allele Identifier (CA ID)

ClinGen
Clinical Genome Resource

Baylor College of Medicine

MyVariant.info

COSMIC
Catalogue Of Somatic Mutations In Cancer

ExAC

ClinVar
Clinically relevant variation

gnomAD

dbSNP
Organisms
dbVar

Registered Canonical Allele

Genomic HGVS

NC_000011.10:g.68032291C>T
NC_000011.9:g.67799758C>T
NC_000011.8:g.67556334C>T

ClinVar Variation Id: 214835

dbSNP Id: rs369602258

CA321211

Transcript HGVS

NM_002496.3:c.64C>T
XM_005274013.1:c.64C>T
XM_005274014.1:c.64C>T
XM_005274015.1:c.-244C>T
XM_011545053.1:c.64C>T
NM_002496.4:c.64C>T
ENST00000313468.9:c.64C>T
ENST00000432321.6:n.181C>T
ENST00000453471.6:c.64C>T
ENST00000525419.5:c.56-632C>T
ENST00000525628.1:c.64C>T
ENST00000526339.5:c.64C>T
ENST00000526446.5:c.64C>T
ENST00000528492.1:c.-67+1558C>T
ENST00000529645.1:n.55C>T
ENST00000531228.1:c.119C>T
ENST00000531796.1:n.221C>T
ENST00000532399.1:n.85C>T

MyVariant Identifiers: chr11:g.67799758C>T (hg19) chr11:g.68032291C>T (hg38)

ClinVar RCV Id: RCV000196794 RCV000276295 RCV000389629

gnomAD: 11:67799758 C / T

LDH info

Canonical Allele Identifier: CA321211

Gene: NDUFS8 HGNC NCBI

Identifiers and link-outs to other resources

ClinVar Variation Id: 214835

ClinVar RCV Id: RCV000196794 RCV000276295 RCV000389629

dbSNP Id: rs369602258

gnomAD: 11:67799758 C / T

ExAC: 11:67799758 C / T

MyVariant Identifiers: chr11:g.67799758C>T (hg19) chr11:g.68032291C>T (hg38)

Calculator JSON-LD

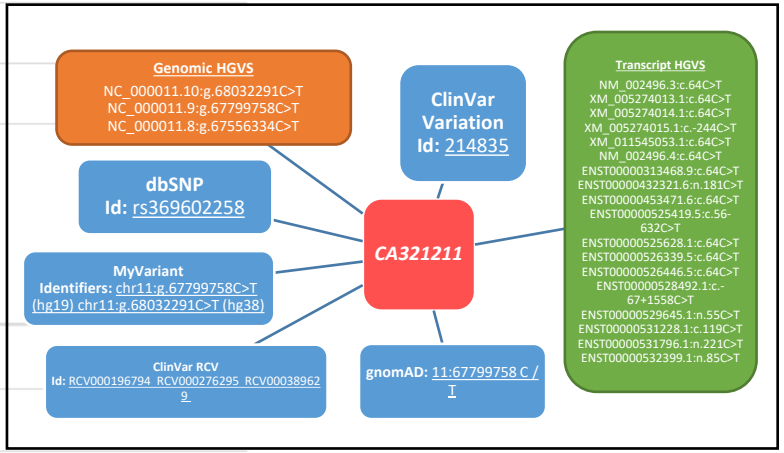
Genomic Alleles

HGVS	Genome Assembly
NC_000011.10:g.68032291C>T , CM000673.2:g.68032291C>T	GRCh38
NC_000011.9:g.67799758C>T , CM000673.1:g.67799758C>T	GRCh37
NC_000011.8:g.67556334C>T	NCBI36
NG_017040.1:g.6675C>T	

Transcript Alleles

HGVS	Amino-acid change
NM_002496.3:c.64C>T VV	NP_002487.1:p.Pro22Ser
XM_005274013.1:c.64C>T	XP_005274070.1:p.Pro22Ser
XM_005274014.1:c.64C>T	XP_005274071.1:p.Pro22Ser
XM_005274015.1:c.-244C>T	XP_005274072.1:p.=
XM_011545053.1:c.64C>T	XP_011543355.1:p.Pro22Ser
NM_002496.4:c.64C>T VV MANE Preferred	NP_002487.1:p.Pro22Ser
ENST00000313468.9:c.64C>T	ENSP00000315774.5:p.Pro22Ser

NOTE: Canonical allele page available in HTML and JSON formats



ClinGen Allele Registry – Biocurator Use Case

Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

METHODS

Single-molecule detection of cancer mutations using a novel PCR-LDR-qPCR assay

Cristian Ruiz, Jianmin Huang, Sarah F. Giardina, Philip B. Feinberg, Aashiq H. Mirza, Manny D. Bacolod, Steven A. Soper, Francis Barany

First published: 16 January 2020 | https://doi.org/10.1002/humu.23987 | Citations: 1

Present address:: Department of Pharmacology, Weill Cornell Medicine, New York, New York.

spatial sample dilution, and detection of multiple mutations in the same position. Our method was tested to interrogate the following common cancer somatic mutations: *BRAF*:c.1799T>A (p.Val600Glu), *TP53*:c.743G>A (p.Arg248Gln), *KRAS*:c.35G>C (p.Gly12Ala), *KRAS*:c.35G>T (p.Gly12Val), *KRAS*:c.35G>A (p.Gly12Asp), *KRAS*:c.34G>T (p.Gly12Cys), and *KRAS*:c.34G>A (p.Gly12Ser). The single-well version of the assay detected 2–5 copies of

- The HGVS expressions were not published
  - We only have gene name, location, and effect
- Do we have existing CA IDs for these variants?
  - http://reg.clinicalgenome.org
    - TP53:c.743G>A (p.Arg248Gln)
    - *BRAF*:c.1799T>A (p.Val600Glu)
    - *KRAS:c.35_39del*

https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23987

ClinGen Allele Registry Tour – Live Demo

- http://reg.clinicalgenome.org
  - Available queries (12)
  - Gene and variation-based query (substitution, indel) ✓
  - Create login ✓
  - Interactively generate variation
  - Bulk query

Small Variants

New Allele Registration

ClinGen Allele Registry

Allele Registry | Pathogenicity Calculator | Profile | Logout

This allele is not present in the registry. To get a canonical identifier, please click on the "Get identifier" below.

Canonical Allele Identifier: Get Identifier ☆

Gene: SBF2 HGNC NCBI

Genomic Alleles

HGVS	Genome Assembly
NC_000011.10:g.9990183T>A , CM000673.2:g.9990183T>A	GRCh38
NC_000011.8:g.9968306T>A	NCBI36
NC_000011.9:g.10011730T>A , CM000673.1:g.10011730T>A	GRCh37
NG_008074.1:g.309025A>T , LRG_267:g.309025A>T	

Transcript Alleles

HGVS	Amino-acid change
NM_030962.3:c.1297-588A>T , LRG_267t1:c.1297-588A>T	NP_112224.1:p.=
XM_005253154.3:c.1297-588A>T	XP_005253211.1:p.=

Built on APIs

- curl -X GET "http://reg.genome.network/allele/CA1234323"

```
[riehle@brlnode-048 ~]$ curl -X GET "http://reg.genome.network/allele/CA1234323"
{
  "@context": "http://reg.genome.network/schema/allele.jsonld",
  "@id": "http://reg.genome.network/allele/CA1234323",
  "externalRecords": {
    "ExAC": [
      {
        "@id": "http://exac.broadinstitute.org/variant/1-169519859-AG-A",
        "id": "1-169519859-AG-A",
        "variant": "1:169519859 AG / A"
      }
    ],
    "MyVariantInfo_hg19": [
      {
        "@id": "http://myvariant.info/v1/variant/chr1:g.169519860del?assembly=hg19",
        "id": "chr1:g.169519860del"
      }
    ],
    "MyVariantInfo_hg38": [
      {
        "@id": "http://myvariant.info/v1/variant/chr1:g.169550622del?assembly=hg38",
        "id": "chr1:g.169550622del"
      }
    ],
    "dbSNP": [
```

ClinGen
Clinical Genome Resource

Baylor College of Medicine

Recent API Additions

- Query alleles by specifying the sequence name directly
  - http://reg.genome.network/redmine/projects/registry/genboree_registry/alleles?refseq=NM_000546.5&begin=1&end=100&skip=0&limit=50
- Query alleles by specifying the chromosome directly (default GRCh38)
  - http://reg.genome.network/redmine/projects/registry/genboree_registry/alleles?refseq=chr17&begin=1&end=1000000&skip=0&limit=50
- Beta – specifying left, right, or explicit alignment

ClinGen
Clinical Genome Resource

Baylor College of Medicine

Structural Variants

- Tools and resources for structural variants have lagged behind the capabilities (sequencing technology, processing pipelines, representative models, naming conventions, etc.) that exist for smaller variants
- On average, structural variants:
  - 2.9 rare occurrences that alter coding regions
    - Affect dosage or structure of 4.2 genes and account for 4-11.2% of rare high-impact coding alleles
  - Account for 17.2% of rare alleles genome-wide
    - Predicted deleterious effects equivalent to loss-of-function coding alleles
  - 90% are noncoding deletions (mean of 19.1 per genome)
  - **3X more likely to associate with a GWAS study signal than SNVs**
  - Large SVs 50X more likely to affect the expression of a gene compared to an SNV
    - Enrichment at enhancers and other regulatory elements
    - Rare high-impact SVs associated with aberrant expression of nearby genes

ClinGen
Clinical Genome Resource

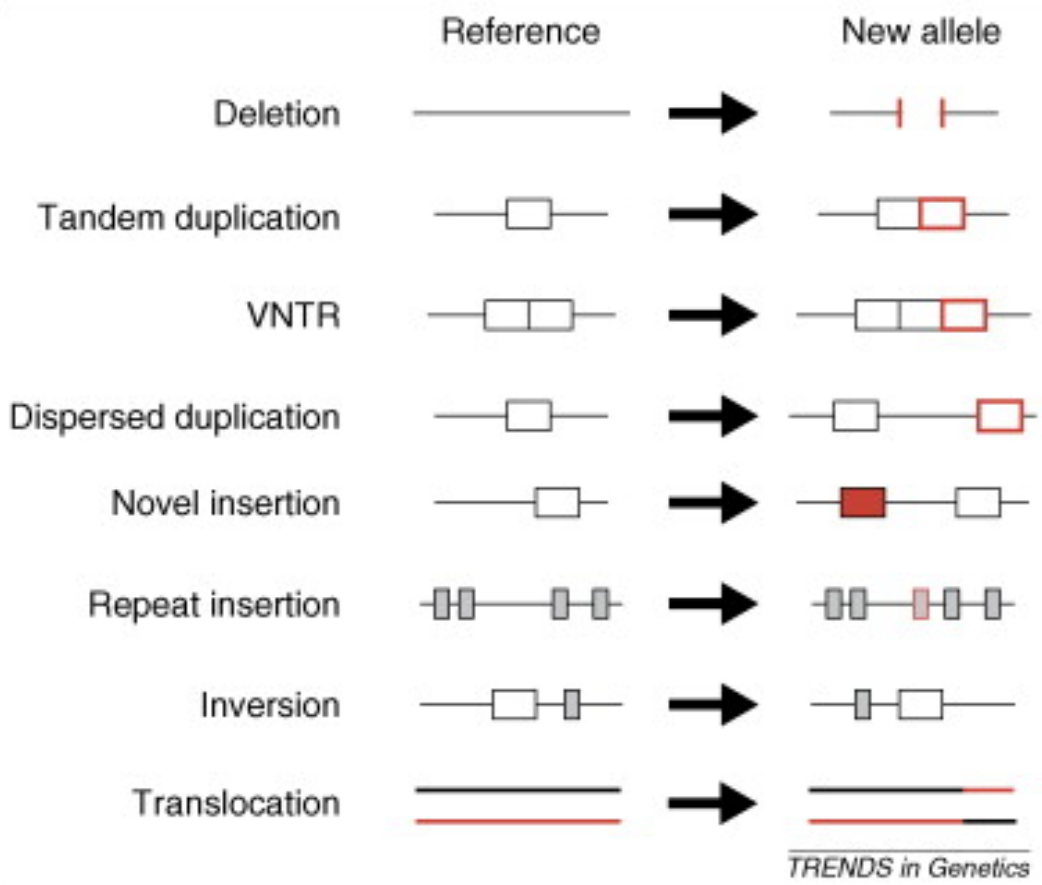
Baylor College of Medicine

Structural Variants

- Structural variants remain a challenge to researchers attempting to ascertain possible sources of disease-causing variants in unsolved Mendelian disorders
  - Profound consequences in evolution and human disease
- Detection has improved, which has resulted in multiple useful databases:
  - ClinVar, dbVar, gnomAD, ExAC, etc.
- However, there does not exist a dereferenceable naming convention in which to identify single SVs across genome builds and transcript versions
  - Stage 1: Copy number variant (CNV) structural variant support in Allele Registry – 2019
  - Stage 2: Non-CNV structural variant support in Allele Registry – ongoing
- Goal
  - Extension of the ClinGen Allele Registry (CAR) to support naming, registration, and querying of structural variants in addition to existing short variant support
    - A centralized resource for supporting clinical WGS, etc. interpretation
    - Serve the needs of ClinGen tools / users first and then extend into the larger community
    - Can we devise a generic model that would allow for additional types down the road?

https://doi.org/10.1038/s41467-018-08148-z

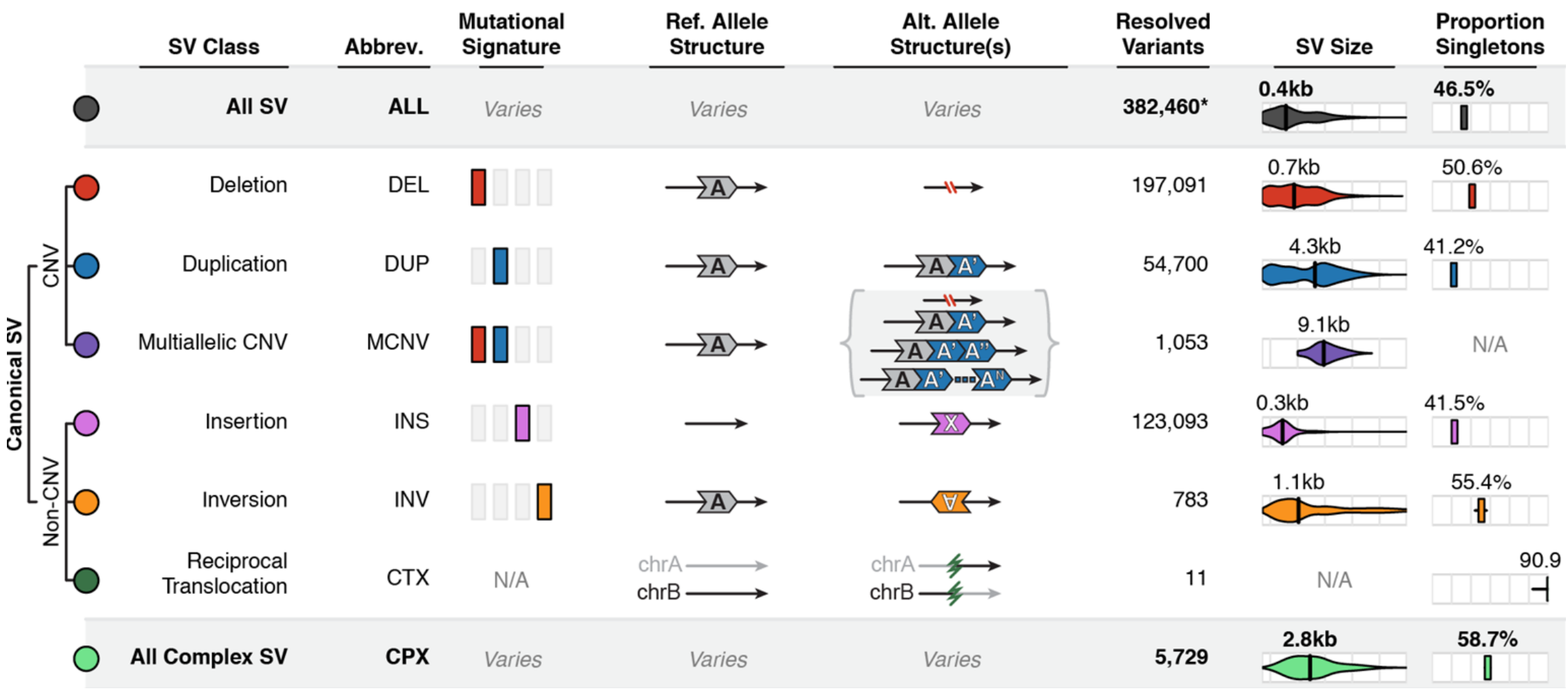
Structural Variants



- Structural variants
  - Variation in the structure of an organism's chromosome, generally 50bp < SV < Mbp
- Copy number variants (CNVs)
  - Structural variants of types deletions and duplications in the genome
- CNVs vs SVs
  - CNVs are unbalanced SVs with gain or loss of genomic material
    - Often detected with read-depth analysis
- Common formats
  - Variant Call Format (VCF)
  - Browser Extensible Data (BED)

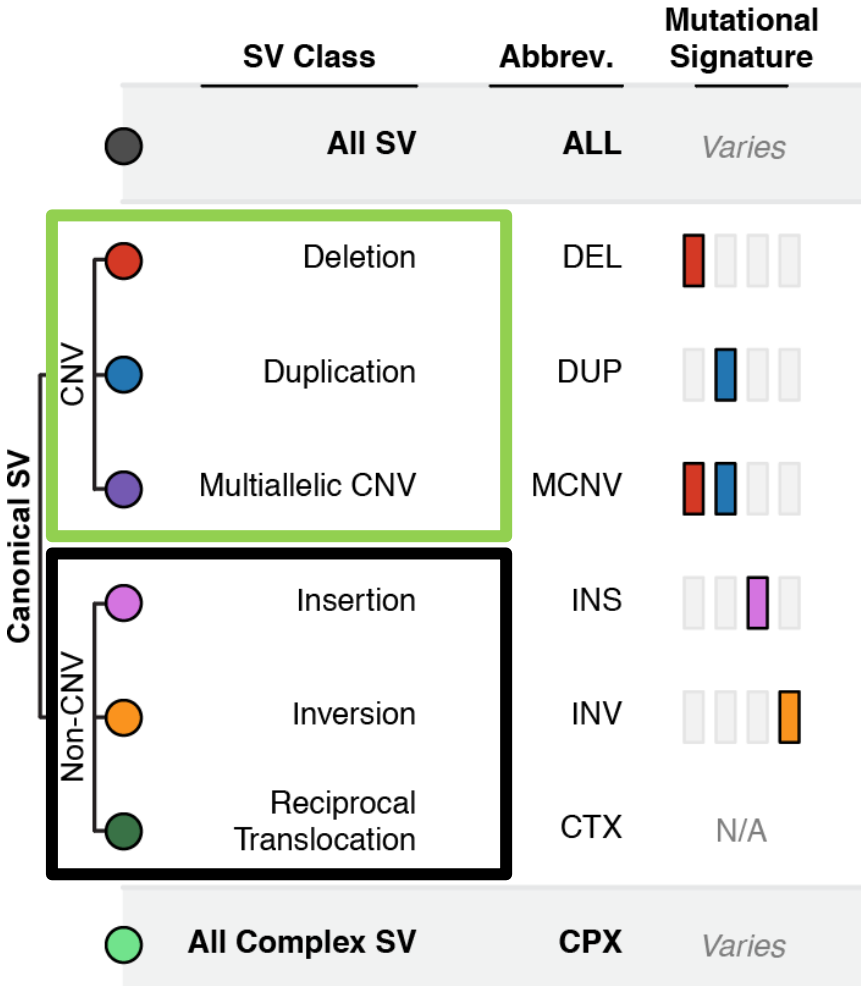
https://www.cell.com/trends/genetics/fulltext/S0168-9525(08)00078-4?large_figure=true

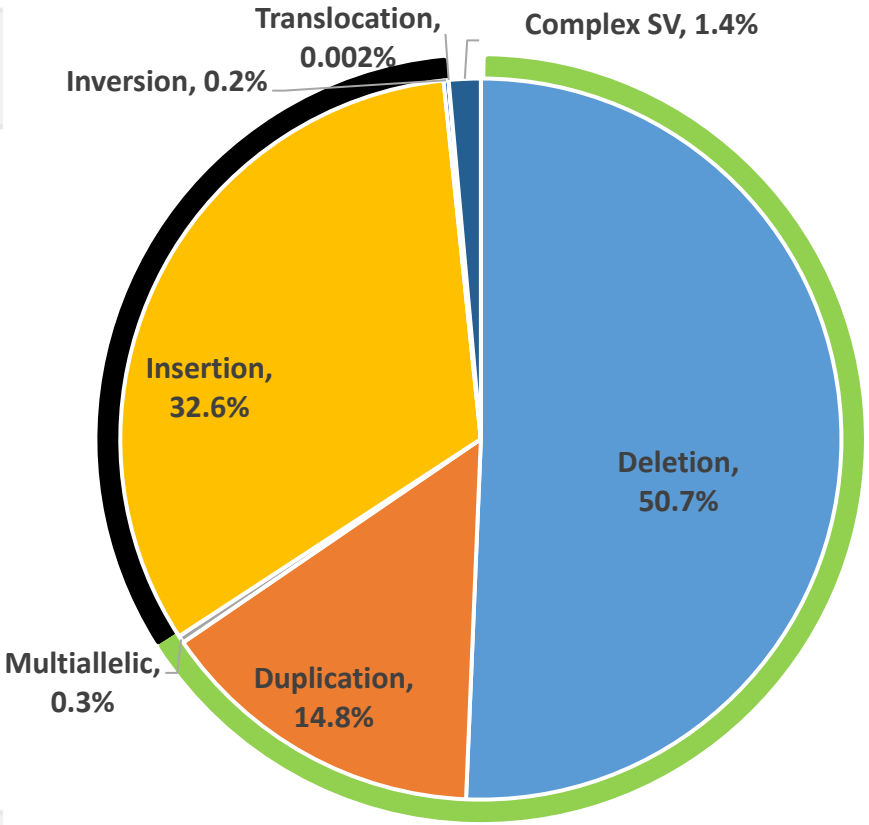
Structural Variants - gnomAD



https://macarthurlab.org/2019/03/20/structural-variants-in-gnomad/

Structural Variants - gnomAD





	Ins/ Inv/ Transloc	32.8%
	Complex	1.4%
	CNVs	65.8%

https://macarthurlab.org/2019/03/20/structural-variants-in-gnomad/

CNV descriptor
(derived from ISCN/ClinVar Preferred Name)

GRCh37 (chr15:?_22770421-32421780_?)x2,3

- GRCh37 — Assembly
- chr15 — Chromosome
- ?_22770421-32421780_? — Interval (precise or range at either end)
- x2,3 — Dosage (precise or range)

CAR – CNV Registration Summary

Source	Descriptors	CNVIDs Returned	Unique	Errors
dbVar	528588	523350	109150	5238
gnomAD	162334	161797	161793	537
ClinVar	55749	53094	40064	2655
ClinVar new	51745	51717	50713	28

Registry Event	Total CNVIDs
ClinVar	55721
ExAC	109111
dbVar	
gnomAD	340204
ClinVar (post bugs)	344679

- CNV metrics API endpoint
  - https://genboree.org/allele/cnv/srvc

ClinGen Allele Registry Tour – Live Demo

- http://reg.clinicalgenome.org
  - Available queries (4)

Copy Number Variants

Structural Variants – non-CNVs

- We have spec'd out initial descriptor models to support non-CNV SVs in addition to CNVs
  - Aim to roll out initial support in 2021

Type	Descriptor
DEL	GRCh37 (chr1:62399-62488del)
DUP	GRCh37 (chr1:137000-155499)x2
INS	GRCh37 (chr1:54665-54665)ins[52]
INS (with source)	GRCh37 (chr1:5583403-5583403)ins(chr1:4185537-4185712)
INV	GRCh37 (chr1:6611596-6612428inv)
CPX.INS_iDEL	GRCh37 (chr1:774568-828434del)ins(chr1:729293-729385)
CPX.INVdel	GRCh37 (chr1:31586010-31586409inv),(chr1:31586410-31588721del)
CPX.INVdup	GRCh37 (chr1:2811131-2812146inv),(chr1:2811953-2812146)x2
CPX.delINV	GRCh37 (chr1:2035872-2036262del),(chr1:2036263-2036671inv)
CPX.delINVdel	GRCh37 (chr1:6460677-6462570del),(chr1:6462571-6463866inv),(chr1:6463867-6467417del)
CPX.dupINV	GRCh37 (chr1:3045506-3045845)x2,(chr1:3045506-3046276inv)
CPX.dupINVdel	GRCh37 (chr1:12047053-12047110)x2,(chr1:12047053-12050131inv),(chr1:12050132-12052413del)
CPX.dupINVdup	GRCh37 (chr1:2988393-2996553)x2,(chr1:2988393-3145675inv),(chr1:2998372-3145675)x2

ClinGen
Clinical Genome Resource

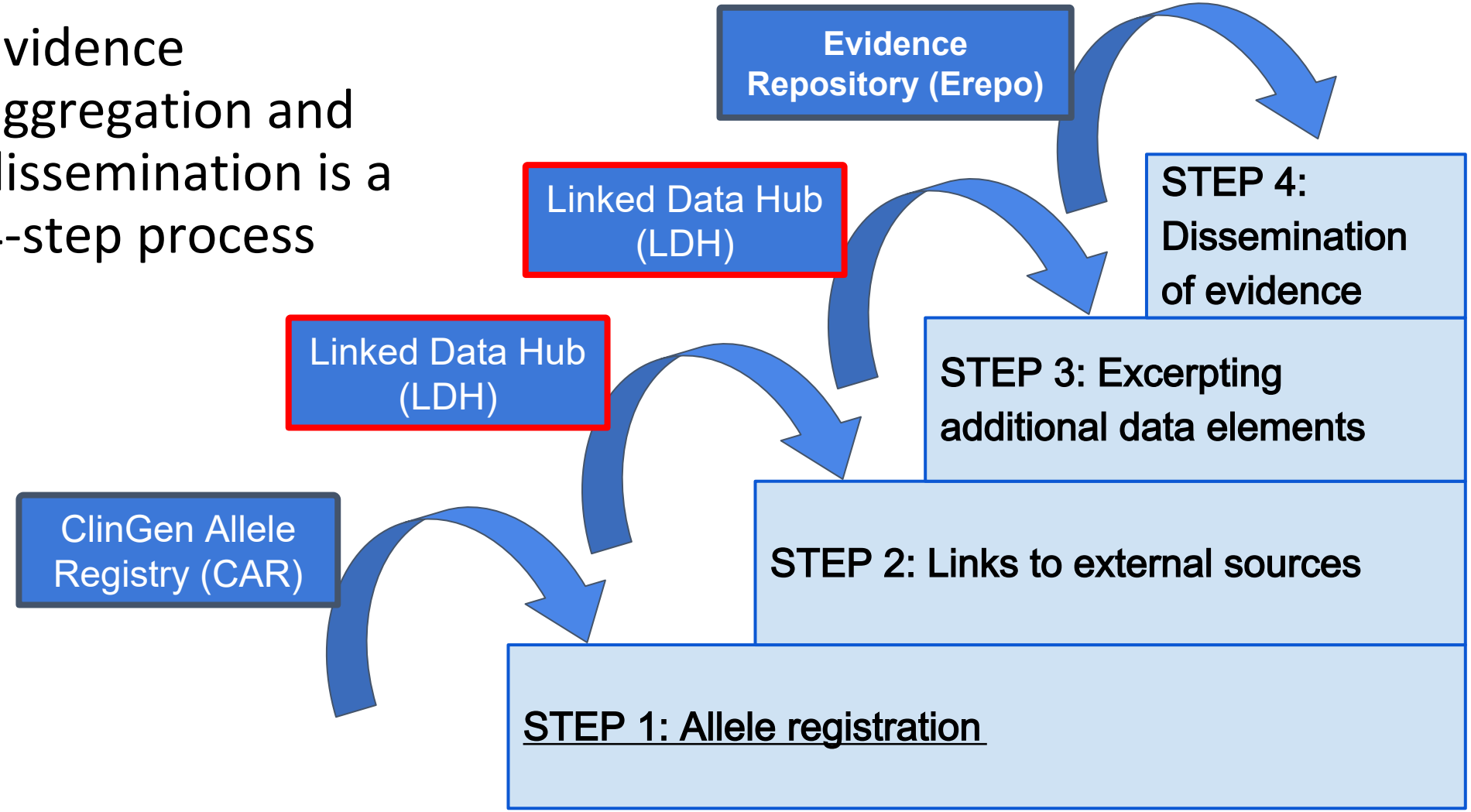
Baylor College of Medicine

Allele Registry – Next Steps

- Support of GA4GH Variant Representation Specification (VRS) spec
- Automation of updating process
- Multi-reference canonicalization support
- Non-coding variant support
- Non-CNV and complex structural variant support
- Variant Search Tool
- Preferred variant name
- Integration with Linked Data Hub (migration of link management)

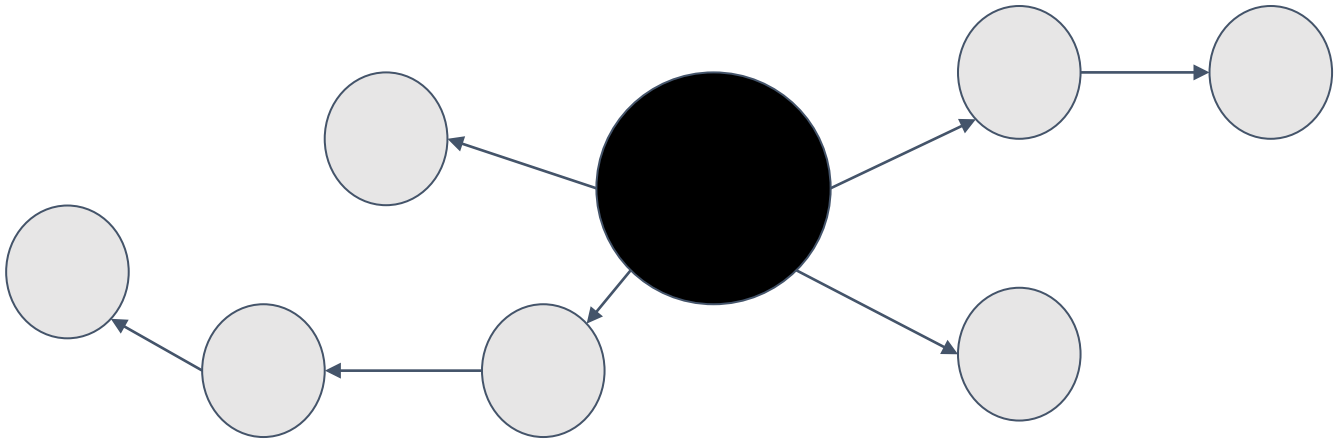
Variant Pathogenicity Interpretation – Linked Data

- Evidence aggregation and dissemination is a 4-step process



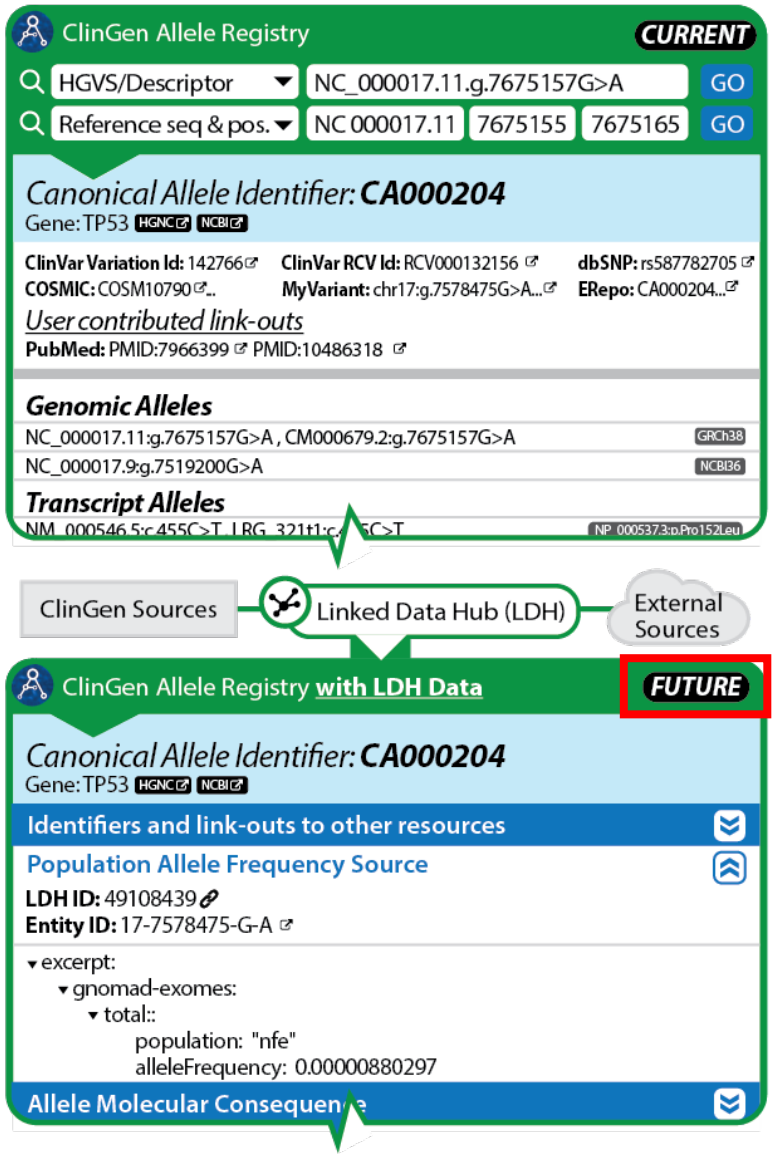
Linked Data – Problem

- What is the need for a system that 'links data' instead of directly integrating with 3rd party external databases?
  - No universal support for permanent references to be used as supporting evidence for variant interpretation
  - Versioning of evidence and metadata that may change
  - Receiving notifications updated data used in past assertions or curations
  - Tracking of provenance of when data was updated and by whom
  - Standardization of data and data access across sources



Allele Registry – Linked Data Vision

- Linked Data Hub
  - Aims to be software infrastructure like "PubMed" for gene and variant knowledge
  - Primary subject in database is CA ID variant
    - Facilitates linking of data and evidence
  - Producing low-level variant data to the VCI
    - Permanent reference for inclusion as "supporting evidence" in variant interpretation
    - Tracking of provenance
    - Standardization of data
    - Developing support for versioning of evidence
  - User-provided links in the CAR
    - Future plans include embedding user-provided links linked in the LDH within the CAR link-out section (e.g. PMIDs)



Linked Data Hub – Solution

- LDH can leverage CA IDs and other identifiers to link data about subjects
- Link to evidence and observations contained in external databases
- Provenance information
- ClinGen SEPIO / GA4GH data standards
- Available via REST-APIs (JSON format)
- Permanent record links

Variant

LDH ID: 40193030
Entity ID: CA023687
Modified: Thu, 03 Oct 2019 16:14:33 -0000


```
▼ Object
  entIri: "http://reg.genome.network/allele/CA023687"
  entId: "CA023687"
  ldhId: "40193030"
  ▼ ld: Object
    ▶ InSilicoPredictionScoreStatement: Array[1]
    ▶ ClinVarSubmission: Array[1]
    ▶ PopulationAlleleFrequencySource: Array[1]
    ▶ AlleleMolecularConsequenceStatement: Array[1]
    ▶ AlleleFunctionalImpactStatement: Array[3]
    ▶ PopulationAlleleFrequencyStatement: Array[19]
  ▼ entAliases: Array[1]
    ▶ 0: Object
  rev: "_Z2qyw7S--_"
  modifier: "sailakss"
  modified: "2019-10-03T16:14:33.640Z"
  entType: "Variant"
  ▼ ldFor: Object
    ▶ Gene: Array[1]
  ldhIri: "https://ldh.genome.network/ldh/Variant/id/40193030"
```


https://ldh.genome.network/ldh/ui/Variant/id/CA023687

Linked Data Hub – Demo

- LDH UI
- Search features
- JSON (via URL)
- Entity counts

Linked Data Hub – Content Updates

- LDH Service Endpoint - https://ldh.genome.network/ldh/srvc

LDH Property	Data Source	2020 Count	2021 Count
Allele Functional Impact	ClinGen FDRepo, etc.	428	428
Allele Molecular Consequence	Ensembl VEP	1,212,476	1,212,476
CIViC Evidence	CIViC	374	688
ClinVar Submission	ClinVar	30,598	30,598
Community Annotation	hypothes.is	1,376	1,376
Gene	ClinGen Allele Registry, ClinVar, HGNC	40	114
In-Silico Prediction	CADD, REVEL, M-CAP (MyVariant.info), S-CAP	252,510	252,510
Population Allele Frequency Source	gnomAD, etc.	324,338	324,338
Population Allele Frequency Statement	gnomAD, etc.	3,073,335	3,073,335+
Variant	ClinGen Allele Registry, ClinVar	1,212,720	4,843,147

Recently Updated | Recently Updated - Awaiting Stats

Linked Data Hub – Ingestion Use Case - CIViC

Civic Evidence

LDH ID: 49770376
Entity ID: 1294
Modified: Thu, 03 Oct 2019 20:49:29 -0000


```
▼Object
  ▼ldFor: Object
    ▼Variant: Array[1]
      ▼0: Object
        entId: "CA351747069"
        ...
  entId: "1294"
  rev: "_Z2qzoDi--F"
  entIri: "https://civicdb.org/api/variants/1294"
  ▼entContent: Object
    ▼evidence_items: Array[1]
      ▼0: Object
        ▶source: Object
        disease: "Von Hippel-Lindau Disease"
        rating: 2
        description: "A single case (named \"3GT\" in this study) wt
        evidence_level: "E"
        evidence_type: "Predisposing"
        clinical_significance: "Uncertain Signifi
        variant_origin: "Somatic Mutation"
        evidence_direction: "Supports"
    ▶variant_types: Array[1]
    ▶hgvs: Array[4]
  modifier: "sailakss"
  modified: "2019-10-03T20:49:29.774Z"
  entType: "CivicEvidence"
  ldhIri: "https://ldh.genome.network/ldh/CivicEvidence/id/49770376"
  ldhId: "49770376"
```



```
entIri: "http://reg.genome.network/allele/CA351747069"
ldhIri: "https://ldh.genome.network/ldh/Variant/id/49742989"
modified: "2019-10-03T20:48:19.173Z"
entId: "CA351747069"
```



```
▼evidence_items: Array[1]
  ▼0: Object
    ▼source: Object
      citation_id: "11536052"
      url: "http://www.ncbi.nlm.nih.gov/pubmed/11536052"
    disease: "Von Hippel-Lindau Disease"
    rating: 2
    description: "A single case (named \"3GT\" in this study
    evidence_level: "E"
    evidence_type: "Predisposing"
    clinical_significance: "Uncertain Significance"
    variant_origin: "Somatic Mutation"
    evidence_direction: "Supports"
```

Linked Data Hub – VPT / VCI Support

- JSON Schema scoping for Variant Prioritization Tool (VPT) messages (schema includes links from LDH to CAR, ClinVar, Ensembl)
  - Message schema for VPT
    - Alpha testing

- Content

  - event
  - id
  - ldh
  - hgnc
  - clinVarVariantId
  - caId
  - dbSNPId
  - clinVarVariantTitle
  - clinVarClinicalSignificanceSummary
  - clinVarSignificanceConflicts
  - clingenPreferredTitle
  - canonicalRefSeqTranscriptHGVS
  - maneTranscripts
  - molecularConsequences
  - hgvsNames
  - transcriptHGVS
  - proteinEffects
  - populationAlleleFrequencyStatements
  - revel
  - populationAlleleFrequencyMetadata

ClinGen
Clinical Genome Resource

Linked Data Hub – External Databases – Active Development

Baylor College of Medicine

Database	Introduction	Monthly Call Participant	Move Forward w/ LDH Integration?	Structured Data - First Pass	Initial LDH Ingestion
BRCA Exchange	✓	✓	✓		
CIViC	✓	✓	✓	✓	✓
LOVD	✓	✓	✓	✓	
MaveDB	✓	✓	✓	✓	
MSeqDR	✓	✓	✓	✓	

Linked Data Hub – External Databases Survey

- **BRCA Exchange**
  - Clinical significance and classification
  - Links to citations and supporting evidence
- **CIViC**
  - Actionability scores and assertions
  - Evidence
- **MaveDB**
  - Score sets (variant effect scores, counts, and associated metadata)
  - Experiment and experiment sets
- **MSeqDR**
  - Genomic and enhanced annotations
  - Consequence
- **LOVD**
  - Link out to LOVD entity
  - (Potentially listing gaited flags indicating presence or absence of pathogenicity assertions)

ClinGen
Clinical Genome Resource

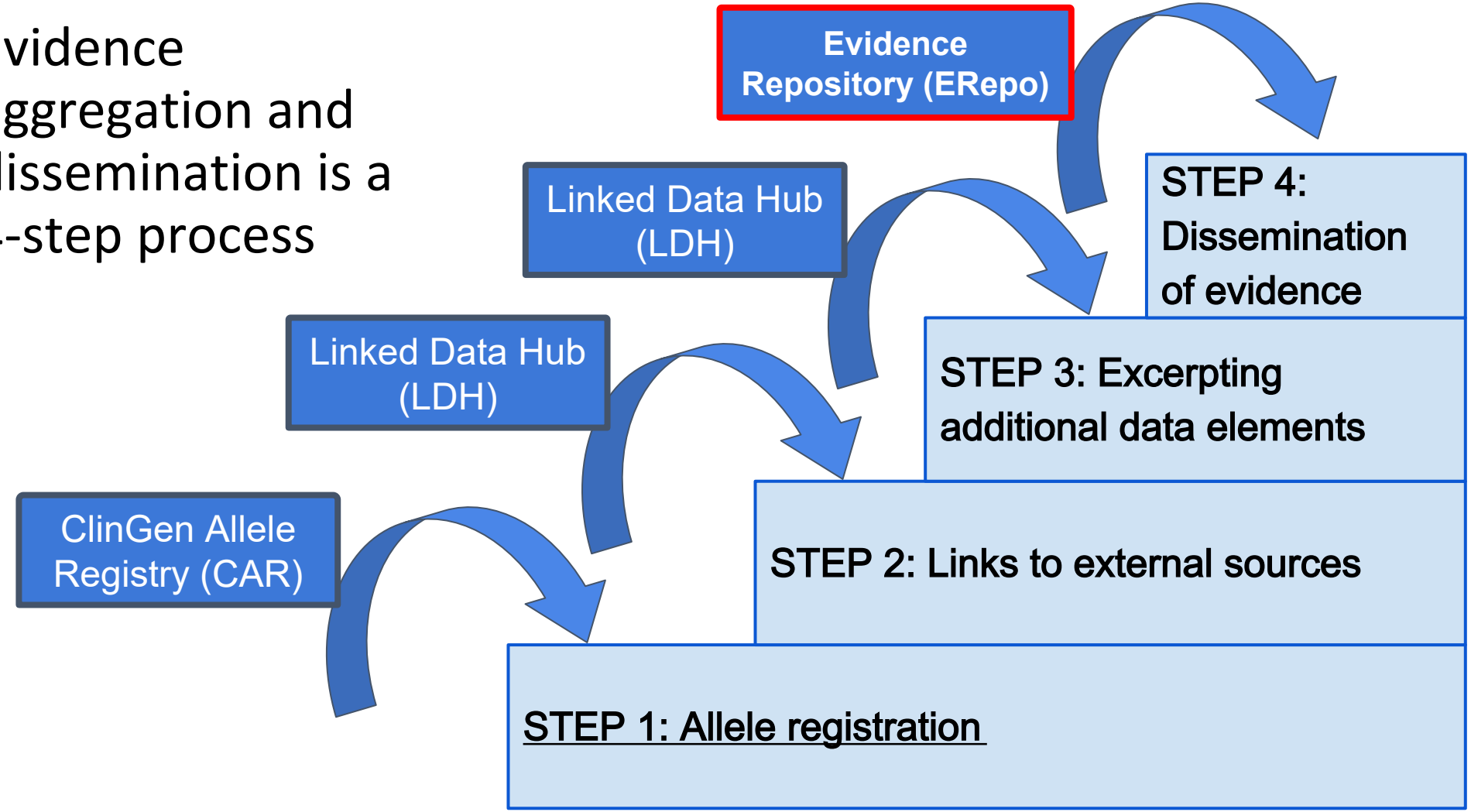
Linked Data Hub – Roadmap

Baylor College of Medicine

- 6-month roadmap
  - VPT connection operational
  - Versioning support (beta)
  - VCI schema design
  - Ingestion of external data
- 12-month roadmap
  - VCI connection operational
  - Versioning support (final)
  - ERepo links to versioned supporting data (FDA)
  - Self-service ingestion

Variant Pathogenicity Interpretation – Evidence

- Evidence aggregation and dissemination is a 4-step process



ClinGen
Clinical Genome Resource

Dissemination of Evidence – Problem

Baylor College of Medicine

- Can aggregated evidence help us solve critical questions?
  - Is variation in a gene disease-causing?
  - Which gene variants are responsible for the disease phenotype?
  - Can we provide auditability of the curation process?
- In order to be recognized by the FDA as the First Public Human Genetic Variant Database address these questions, we were required to solve:
  - Development of the expert panel process along with active participation
    - ?
  - Precisely defining variants under review
    - ?
  - Systemically applied ACMG/AMP evidence codes and allowing VCEPs to curate as teams
    - ?
  - Making all curated evidence rapidly accessible to the public
    - ?

ClinGen
Clinical Genome Resource

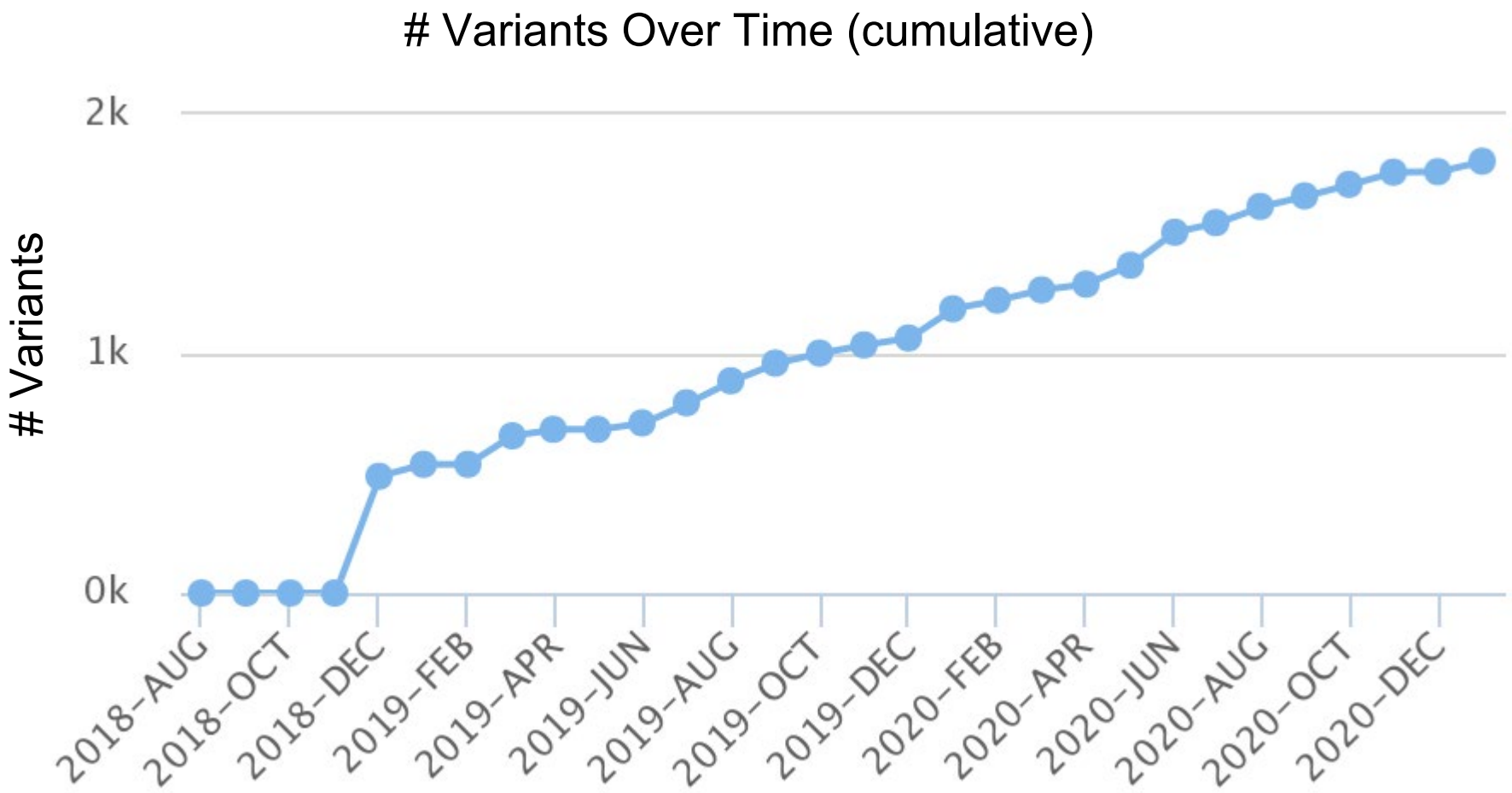
Dissemination of Evidence – Solution

Baylor College of Medicine

- Development of the expert panel process along with active participation
  - **Variant Curation Expert Panel (VCEP)**
- Precisely defining variants under review
  - **ClinGen Allele Registry (CAR)**
- Systemically applied ACMG/AMP evidence codes and allowing VCEPs to curate as teams
  - **ClinGen Variant Curation Interface (VCI) => publishing**
- Making all curated evidence rapidly accessible to the public
  - **ClinGen Evidence Repository (ERepo)**
    - **Designated source of FDA-recognized variant classifications**
- Recognition enables use of ClinGen variant classification in place of traditional FDA clinical validation studies normally required to establish variant pathogenicity
- Variant interpretations from the VCI are published in the ERepo
  - Via transmission of messages published in the ClinGen Data Exchange

Evidence Repository – Growth and Demo

- Currently has evidence for 1,805 variants
- Demo



ClinGen
Clinical Genome Resource

Acknowledgements

Baylor College of Medicine

- Baylor College of Medicine
  - Piotr Pawliczek
  - Ronak Patel
  - Sai Lakshmi Subramanian
  - Andrew Jackson
  - Neethu Shah
  - Anton Kodochygov
  - Aleksandar Milosavljevic (BRL Lab Director)
  - Sharon Plon (BCM Site PI)
  - Alejandro Zuniga
  - Matt Roth
  - Sameer Paithankar
  - Deb Ritter
- Stanford
  - Matt Wright
  - Rao Madhavrao
  - Christine Preston
  - Helio Costa
  - Mark Mandell
  - Gloria Cheung
  - Carlos Bustamante
  - Thomas Montine (Stanford Site PI)
- Broad / Geisinger
  - Larry Babb
  - Erin Riggs
  - Tristan Nelson
  - Christa Martin
- UNC
  - Julianne O'Daniel
  - Chris Bizon
  - Courtney Thaxton
  - Jonathan Berg (UNC Site PI)
- Variant Curation Expert Panels, additional ClinGen Consortium Members, etc.
- GA4GH
  - Reece Hart

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

NIH National Human Genome Research Institute

National Human Genome Research Institute (NHGRI) grants [U41HG006834], [U41HG009649], [U41HG009650]

ClinGen
Clinical Genome Resource

Baylor College of Medicine

Summary

- Role of ClinGen
- Tools – Overview, solution, and live demo
  - ClinGen Allele Registry (CAR)
  - Linked Data Hub (LDH)
  - Evidence Repository (ERepo)

- Thank you on behalf of many different ClinGen teams and groups for attending this meeting and today's workshop
  - Questions?
    - Email me (Kevin Riehle)

Riehle@bcm.edu

Supplementary Slides

Linked Data Hub – Leveraging Messaging Queues

- (Utilizing different message queues)
- Ingestion queue - deposition of messages via REST endpoint to be processed by consumer and ingested into the LDH
- Dissemination queue - LDH messages hosted for consumption by the community via subscribers (infinite retention)
- Aggregation queue - messages constructed for specific downstream tools (e.g. VPT) that contain a mixture of ClinGen + other content that can be deposited onto external queues (e.g. ClinGen Data Exchange)
- Versioning queue - internal only queue that provides versioning support of locally stored documents via permalinks
- Phase I - Capture versions (via slower queue queries)
- Phase II - Indexing enhancement (via quick key-value queries)

Linked Data Hub – Example

https://ldh.genome.network/ldh/ui/Variant/id/CA023687

Variant

LDH ID: 40193030

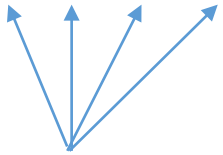
Entity ID: CA023687

Modified: Thu, 03 Oct 2019 16:14:33 -0000


```
▼ Object
  entIri: "http://reg.genome.network/allele/CA023687"
  entId: "CA023687"
  ldhId: "40193030"
  ▼ ld: Object
    ▶ InSilicoPredictionScoreStatement: Array[1]
    ▶ ClinVarSubmission: Array[1]
    ▶ PopulationAlleleFrequencySource: Array[1]
    ▶ AlleleMolecularConsequenceStatement: Array[1]
    ▶ AlleleFunctionalImpactStatement: Array[3]
    ▶ PopulationAlleleFrequencyStatement: Array[19]
  ▼ entAliases: Array[1]
    ▶ 0: Object
  rev: "_Z2qyw7S--_"
  modifier: "sailakss"
  modified: "2019-10-03T16:14:33.640Z"
  entType: "Variant"
  ▼ ldFor: Object
    ▶ Gene: Array[1]
  ldhIri: "https://ldh.genome.network/ldh/Variant/id/40193030"
```

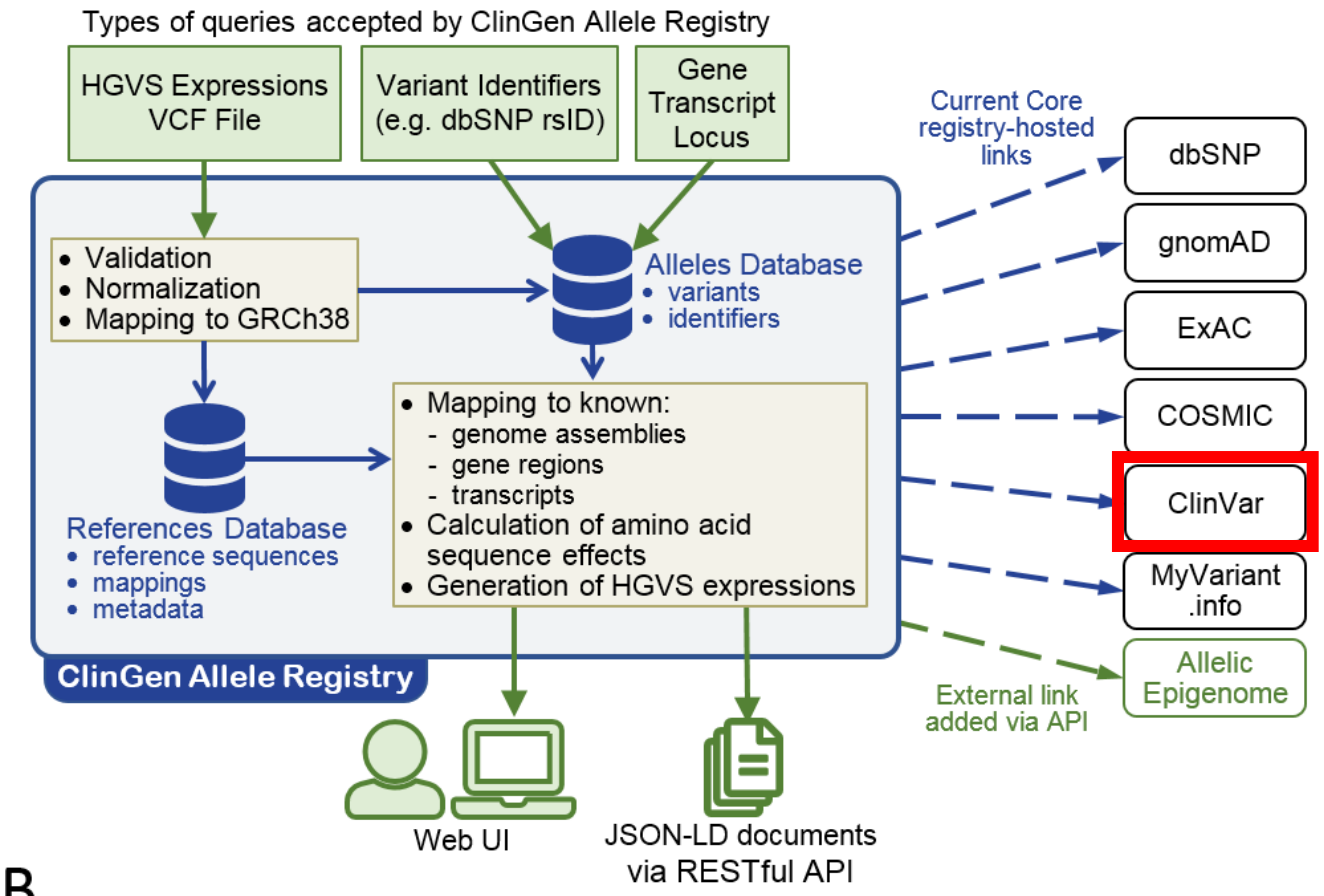
Naming – One More Point

- The CAR produces globally unique variant identifiers that enables aggregation of information from different data sources
- GRCh38 chr7 117,590,426 - 117,590,430
  - … G A A A T…

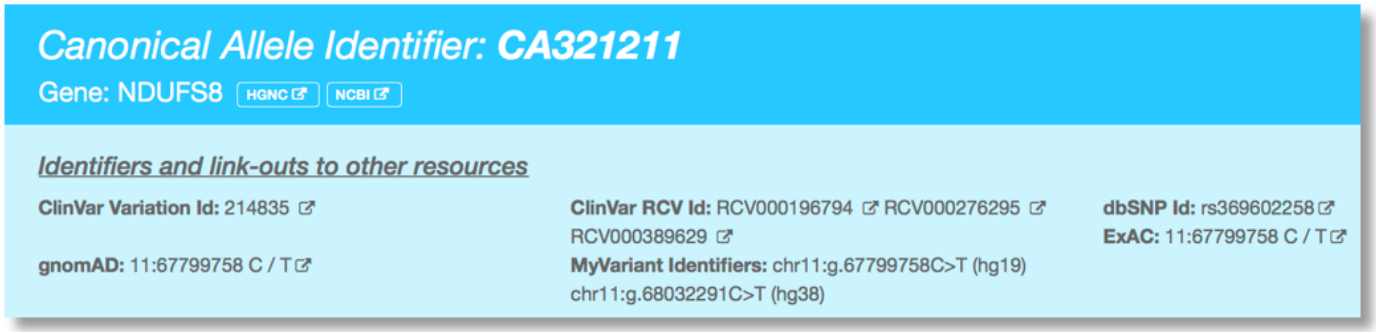


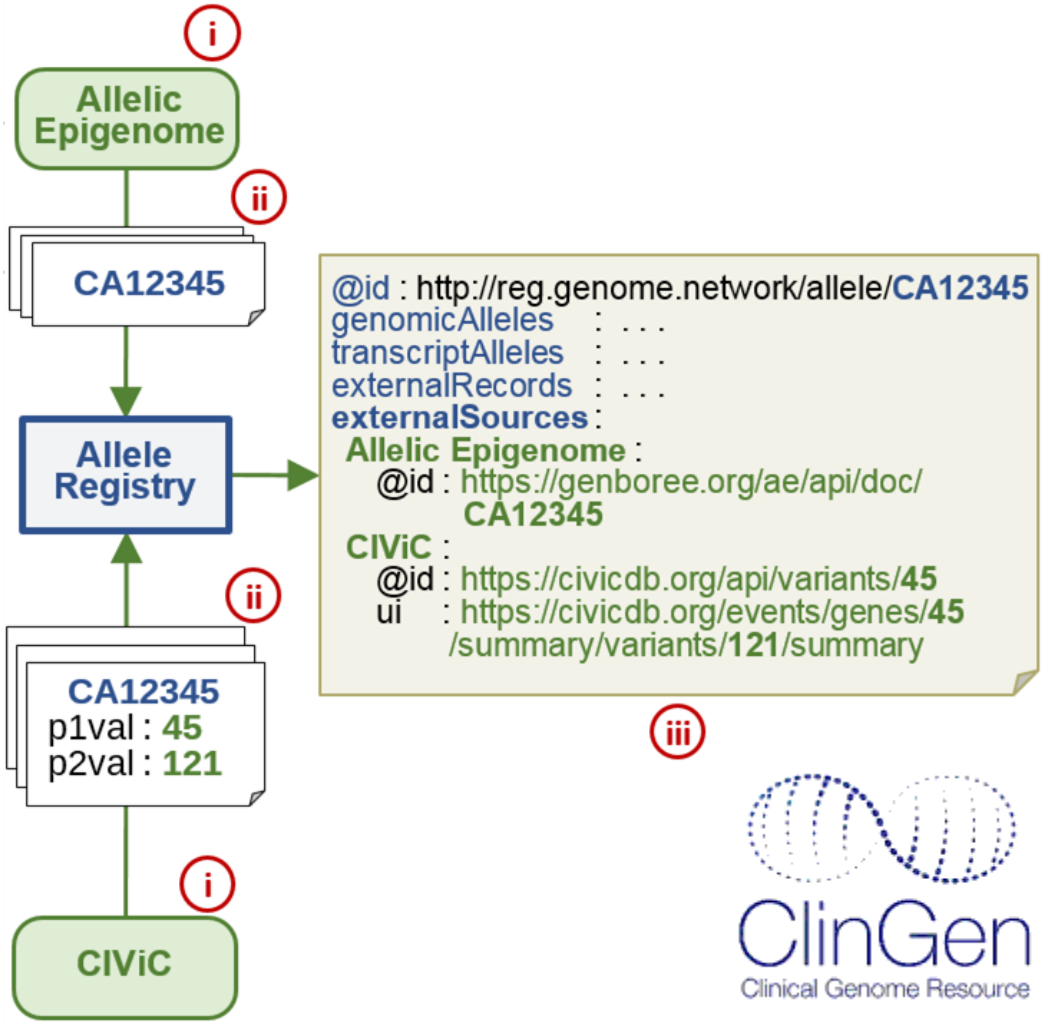
- Insert an A at each of the above positions results in the same net change
  - NC_000007.14:g.117590426_117590427insA => GAAAAT => CA1139532343
  - NC_000007.14:g.117590427_117590428insA => GAAAAT => CA1139532343
  - NC_000007.14:g.117590428_117590429insA => GAAAAT => CA1139532343
  - NC_000007.14:g.117590429_117590430insA => GAAAAT => CA1139532343
- All HGVS expressions served up as NC_000007.14:g.117590429dup

A.



B.

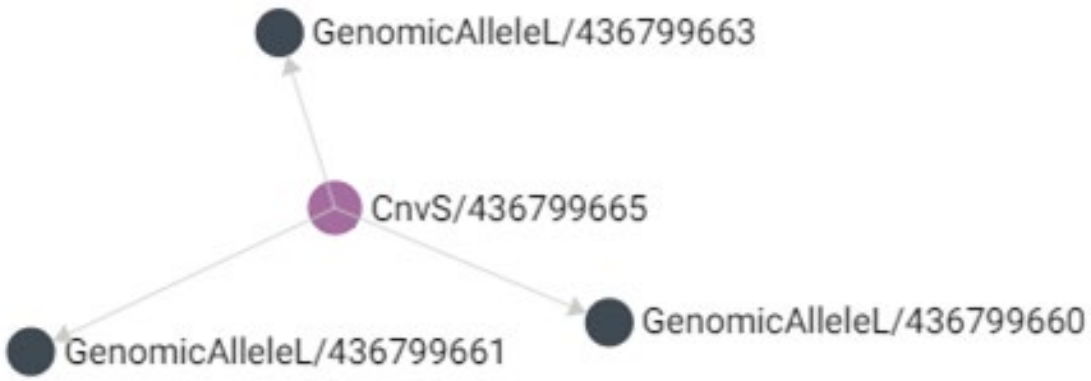




Pawliczek P, et al. ClinGen Allele Registry links information about genetic variants. *Human Mutation*. 2018;39:1690–1701.

CAR CNV Integration

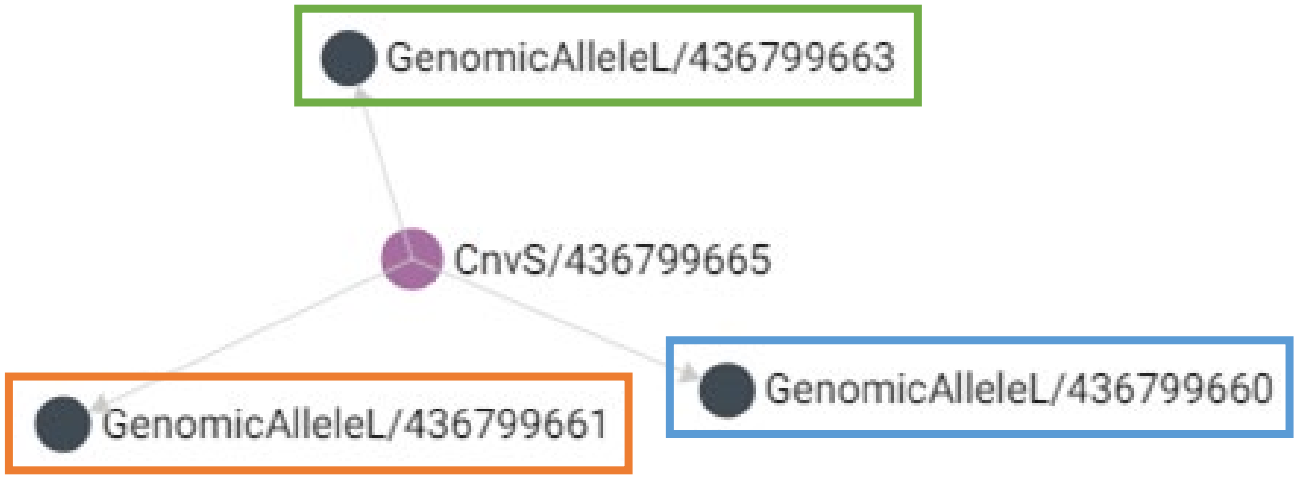
- ArangoDB
  - NoSQL document-based database (hybrid graph/doc)
  - Document collection can only contain one kind of document and map to an entity (self-imposed restriction)
    - Collections (entities)
      - Vertices in graph
    - Edges (relations between collections)
      - Edges in graph




```
"@id": "https://reg.genome.network/allele/CACN436799665",
"CACNid": "CACN436799665",
"genomicAlleles": [ {
        "chromosome": "2",
        "coordinates": {
                "copies": 0,
                "cytoband": "2q33.1",
                "end": 197707187,
                "start": 197705406 },
        "descriptor": "GRCh38 (chr2:197705406-
197707187)x0“,
        "referenceGenome": "GRCh38"
        },
        ...
```

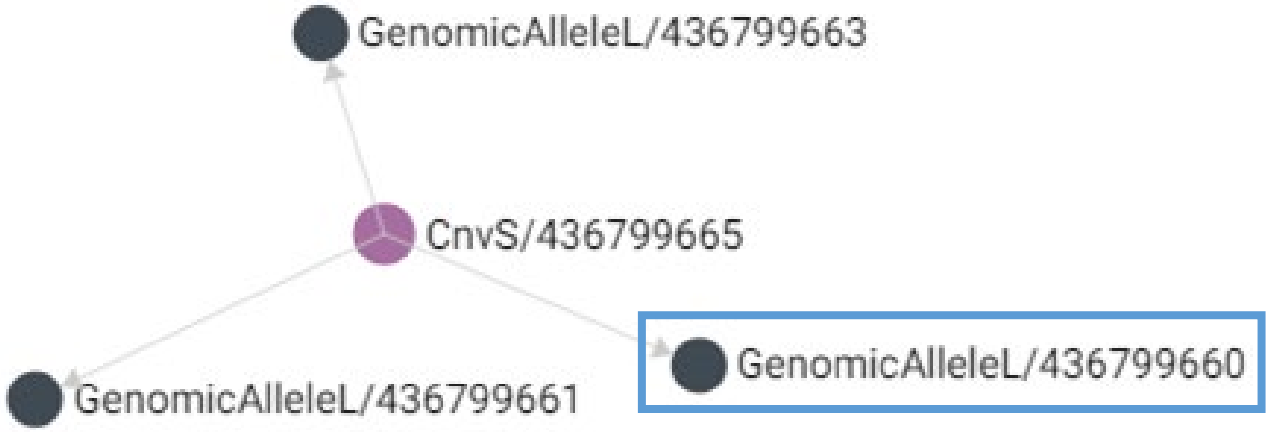
Modeling CNVs in ArangoDB – Collections

- CnvS document
  - {“_id”:”CnvS/436799665} for a CnvS document => 436799665 (subtracting CACN from above ID)
- CnvS_GenomicAlleleL_E edges
  - {"_from":"CnvS/436799665","_to":"GenomicAlleleL/436799660"} => 436799668
  - {"_from":"CnvS/436799665","_to":"GenomicAlleleL/436799661"} => 436799670
  - {"_from":"CnvS/436799665","_to":"GenomicAlleleL/436799663"} => 436799671

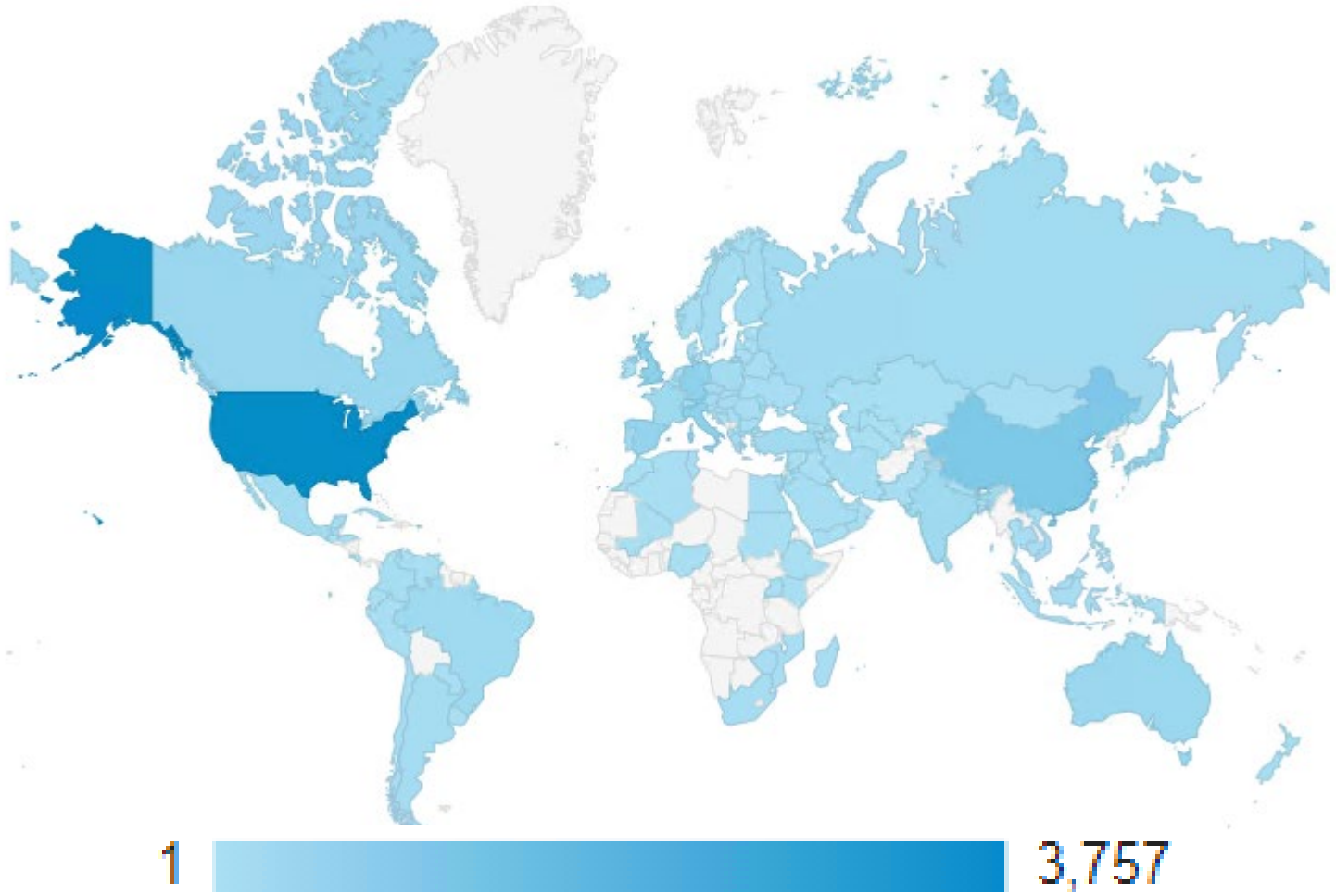


Modeling CNVs in ArangoDB – Collections

- GenomicAlleleL/436799660
  - object {4}
    - chromosome : 2
    - coordinates {5}
      - minCopies : 0
      - maxCopies : 0
      - outerStart : 197705406
      - outerEnd : 197707187
      - cytoband {2}
        - start : 2q33.1
        - end : 2q33.1
    - descriptor : GRCh38 (chr2:197705406-197707187)x0
    - referenceGenome : GRCh38



ClinGen
Clinical Genome Resource

Baylor College of Medicine

Registry Users