Showing your work:

Combining genetic variant interpretations with evidence to enable reanalysis and reuse

Bradford C. Powell MD, PhD[1,2], Matthew Brush PhD[3], Marina DiStefano PhD[4], Robert R. Freimuth PhD[5], Rajarshi Ghosh PhD[6], Melissa A. Haendel PhD[3], Steven Harrison PhD[7], Tristan Nelson[8], Heidi Rehm, PhD[4,7], Shawn Rynearson[9], Natasha Strande PhD[1], Chris Bizon PhD[2], Larry Babb[4,10]

[1]University of North Carolina at Chapel Hill, Department of Genetics; [2]Renaissance Computing Institute; [3]Oregon Health and Science University, Department of Medical Informatics and Clinical Epidemiology; [4]Laboratory for Molecular Medicine, Partners Health Care Personalized Medicine; [5]Mayo Clinic, Department of Health Sciences Research; [6]Baylor College of Medicine, Department of Pediatrics-Oncology; [7]Harvard Medical School; [8]Geisinger Autism and Developmental Medicine Institute; [9]University of Utah, Department of Biomedical Informatics; [10]Sunquest Information Systems, Inc.

Objective

Facilitate clinical molecular genetic interpretation by creating a data model to collect and exchange analyses of genetic effect along with the evidence and provenance of those analyses.

Background

- The ACMG-AMP guidelines provide a framework for a more systematic evaluation of pathogenicity
  - These guidelines establish a set of **rules** for **combining a set of criteria** with different **strengths of evidence** to classify variants as Benign, Likely Benign, Uncertain Significance, Likely Pathogenic or Pathogenic for Mendelian conditions.
  - However, laboratories differ in how these criteria may be applied (e.g., as described in Amendola et al 2016)
- Re-evaluation of pathogenicity assertions (in light of new data or to reassess discordant assertions) requires fine-grained sharing of the evidence underlying these assertions

Concrete variant assessments guide model development

- Specific examples of the application of the ACMG-AMP criteria were used to:
  - demonstrate the data types to be captured, and
  - inform development of the model
- Textual descriptions were generated by four molecular analysts, drawing from sources including clinical variants, research variants, and published descriptions of variant interpretations (many derived from the variants interpretations compared among CSER sites as described in Amendola et al 2016)
- A data modeler then independently produced a provisional model to represent the data in a provisional structural form.
- Provisional models were iteratively refined throughout model development.

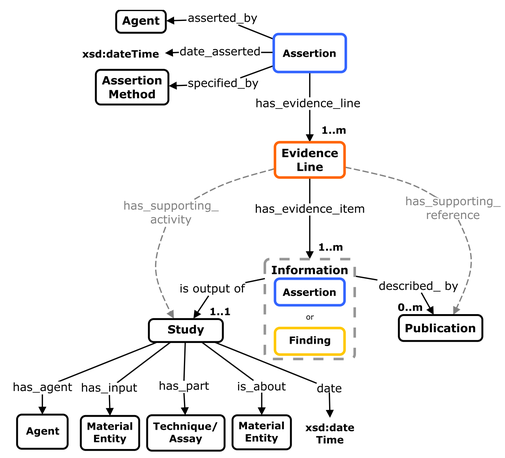
Criterion	# of Examples	Data Types Used
PVS1	4	ConditionMechanism, MolecularConsequence, NullAllele
PS1	3	VariantInterpretation
PS2	3	FamilyCondition, IndividualAlleleInheritance, IndividualCondition
PS3	4	AlleleFunctionalImpact
PS4	3	CaseControl
PM1	3	BenignMissenseVariationRate, RegionAlleles, RegionAnnotation, VariantInterpretation
PM2	7	AlleleFrequency
PM3	3	IndividualAllele, IndividualCondition, VariantInterpretation
PM4	3	ExcludedRegionType, MolecularConsequence, RegionType
PM5	4	VariantInterpretation
PM6	4	FamilyCondition, IndividualAlleleInheritance, IndividualCondition
PP1	4	AggregateSegregation, FamilySegregation
PP2	2	BenignMissenseVariationRate, ConditionMechanism, MolecularConsequence
PP3	3	InSilicoPrediction
PP4	2	LocusHeterogeneity
PP5	3	VariantInterpretation
BP1	4	ConditionMechanism, MolecularConsequence, NullAllele
BP2	3	ConditionPenetrance, IndividualAllele, IndividualCondition, VariantInterpretation
BP3	3	ExcludedRegionType, MolecularConsequence, NullAllele, RegionAlleles, RegionType
BP4	3	AlleleConservation, InSilicoPrediction, MolecularConsequence, NullAllele
BP5	5	IndividualAllele, IndividualCondition, VariantInterpretation
BP6	3	VariantInterpretation
BP7	3	Conservation, InSilicoPrediction, MolecularConsequence
BS1	5	AlleleFrequency, ConditionPenetrance, ConditionPrevelance
BS2	3	ConditionPenetrance, IndividualAllele, IndividualCondition, VariantInterpretation
BS3	2	AlleleFunctionalImpact
BS4	4	FamilySegregation
BA1	2	AlleleFrequency

Table 1: The number of examples generated for each of the ACMG-AMP variant interpretation criteria (Richards et al 2015) and the data types that comprise lines of evidence for each of these criteria.

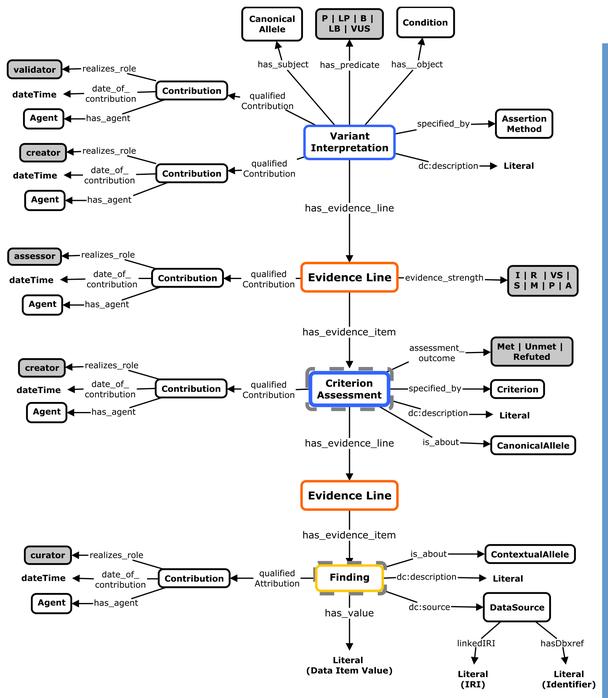
Interpretations modeled together with evidence and provenance

Figure 2: Representing variant interpretations according to the Scientific Evidence and Provenance Information Ontology (SEPIO)

(a)



(b)



(a) *High-level overview of SEPIO.*

- **SEPIO** (a project of the Monarch Initiative) is being developed to support the description of evidence and provenance relevant to scientific claims.
- Within the generalized SEPIO model, scientific **Assertions** may be informed by multiple lines of evidence (**EvidenceLines**).
- Each **EvidenceLine** in turn groups together supporting fine-grained statements of **Information** (primary **Findings** or prior **Assertions**) that together can be used to help support or refute an **Assertion**, independent of other lines of evidence.
- Additional links between objects in the model may be inferred (e.g., *has_supporting_activity* and *has_supporting_reference*) or may be used to represent imported data for without available provenance

(b) *A SEPIO representation of variant interpretation according to the ACMG-AMP guidelines.*

- The ACMG-AMP variant interpretation guidelines are mapped to SEPIO using at least two levels of **Assertions** and **EvidenceLines** to represent the overall **VariantInterpretations** and individual **CriterionAssessments**
- The strength of evidence for the assessment of a criterion is determined by the analyst and recorded as part of the **EvidenceLine**
- The model is intended to be flexible enough to be able to capture as much information as possible from external sources where metadata may be incomplete.
- Our data representation allows for flexible coding of data attributes (e.g., those shaded in grey), with ongoing work in establishing preferred ontologies for these attributes.

ClinGen Allele Registry	"Canonicalizes" allele representations across different transcripts and genome assemblies
DO Disease Ontology	Semantically integrates disease vocabularies across MeSH, ICD, SNOMED, OMIM
GENO	Provides terms relevant to description of genotypes and relation to phenotypes, such as description of sequence features and modes of inheritance
HGNC HUGO Gene Nomenclature Committee	Provides standard and stable identifiers for human genes
LOINC Logical Observation Identifiers Names and Codes	Defines pathogenicity assertion codes
OBI Ontology for Biomedical Investigations	Describes types of evidence used to evaluate allele function under experimental conditions
SEPIO Scientific Evidence and Provenance Information Ontology	Defines overall structure of evidence, grouping into individual lines of evidence and supporting information, provides additional domain-specific terms such as a coding of the ACMG variant interpretation guidelines
SO Sequence Ontology	Provides terms for description of the predicted molecular consequence of alleles at the protein sequence level

Table 2: A selection of ontologies and data sources used in the Variant Interpretation Model

- Use of established ontologies facilitates data exchange
- Linked-data principles allow for incorporation of additional terms and ontologies
- JSON-LD context file is used to map between human-readable "shorthand" versions of terms and Internationalized Resource Identifiers (IRIs)

A structured example

HS Asserts that CA090919 meets the criterion PM3 for Primary Ciliary Dyskinesia

BECAUSE

Another allele (CA273615) has been previously interpreted by SD as Pathogenic for Primary Ciliary Dyskinesia

AND

An individual (Ind004) has been clinically reported (e.g., on requisition form) to have Primary Ciliary Dyskinesia

AND

Individual Ind004 has both the variant in question (CA090919) and the pathogenic variant (CA273615) in trans

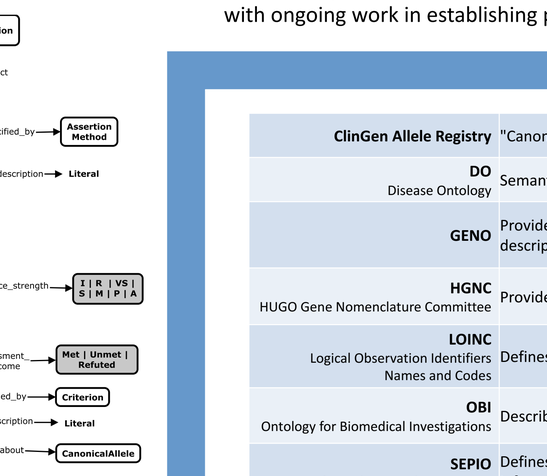


Figure 3: An example Criterion Assessment according to the ClinGen Interpretation Model

- Assessment of ACMG-AMP criterion PM3 (variant in *trans* with a pathogenic variant in a condition known to have recessive mode of inheritance) is supported by an evidence line that combines three pieces of information (one prior **Assertion** and two **Findings**)
- **Canonical Alleles** (in green) are represented according to the ClinGen Allele Registry, which facilitates linking to ClinVar records and HGVS-formatted representations of the variants
- A **Variant Interpretation** would include one or more **Criterion Assessments** as individual lines of evidence.

Products & Milestones

- Description of a semantically-rich set of concrete interpretations using the ACMG-AMP variant interpretation guidelines (individual criteria and in combination for full interpretations)
- Definition of a formal data model for assertions of variant pathogenicity, with both human-readable and computer-readable specifications
- Specification of a JSON-LD-formatted message structure for interoperability
- Contribution to the further development of SEPIO by application to the domain of variant interpretation
- Development of an OWL-based ontology (ongoing) JSON-Schema (planned) representations
- Coordination with the ClinGen Variant Curation and Gene Curation Interfaces (ongoing)
- Development of export capability for deposition of assertions in ClinVar (in process)
- Alignment to additional related ontologies and data sources (ongoing)

datamodel.clinicalgenome.org/interpretation

Supported by U01 HG007437 and U41 HG006834