Information Modeling

HL7 WGM
Sept 23, 2020

Robert R. Freimuth
CG Co-Chair

Agenda

- Why develop an IM?
- Review model
  - Definitional Variant!
- Roadmap

IM Group: Purpose

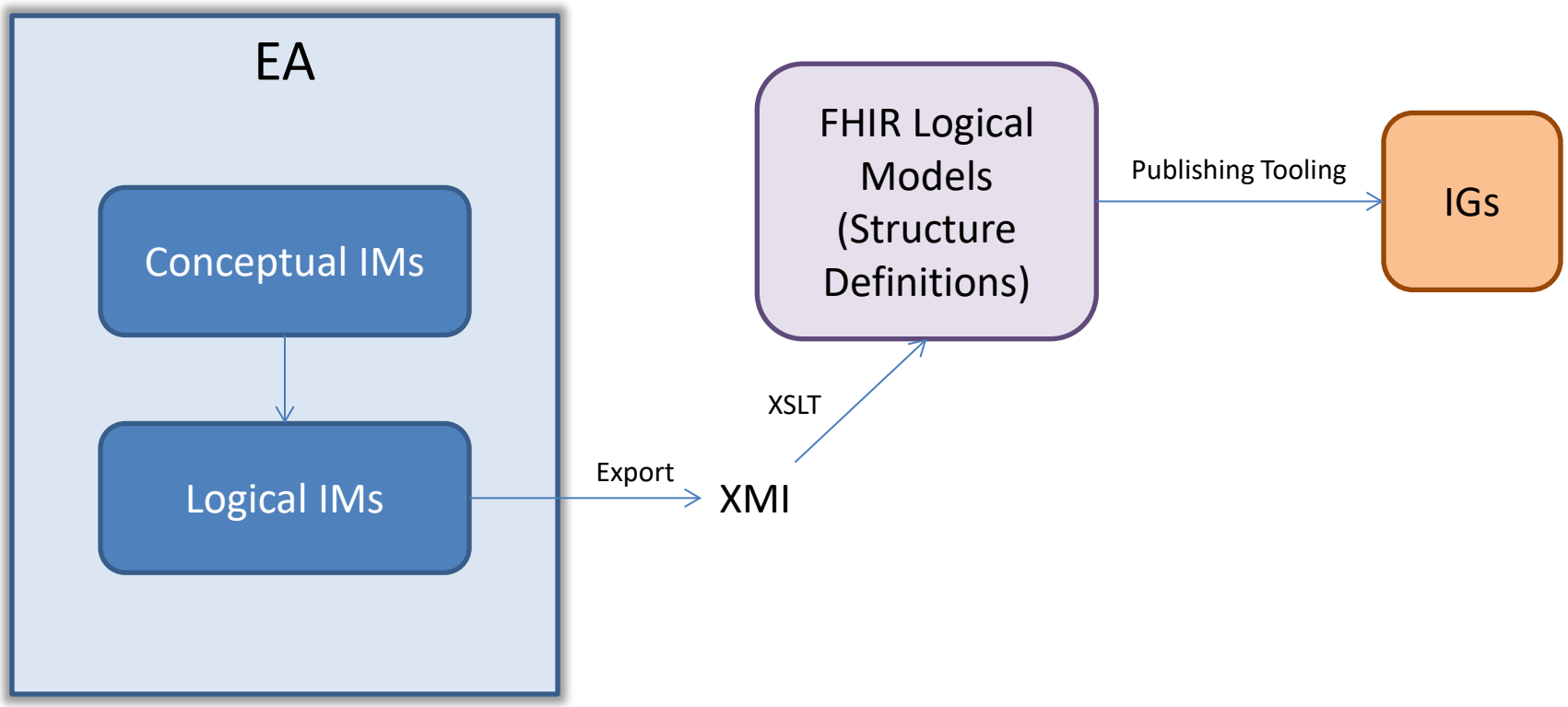
A subgroup of the CG WG will be formed to focus on the development of an information model to represent the clinical genomics domain. The creation of this model is consistent with current CG WG project scope statements and it is a necessary component of the future consolidated standard, which will include the current Domain Analysis Model and Domain Information Model, as described in the CG WG DAM PSS (Q1/2016).

- A model is needed to unify our specifications

CG Information Model (IM)

- Technology agnostic
- Conceptual information model
  - Core concepts that are represented in draft standards produced by the WG
- Promote harmonization and inter-artifact consistency among all CG WG standards
- Incorporate relevant concepts requested by other groups working in the CG domain
  - GA4GH
  - ClinGen
  - NCBI
  - etc

Big Picture

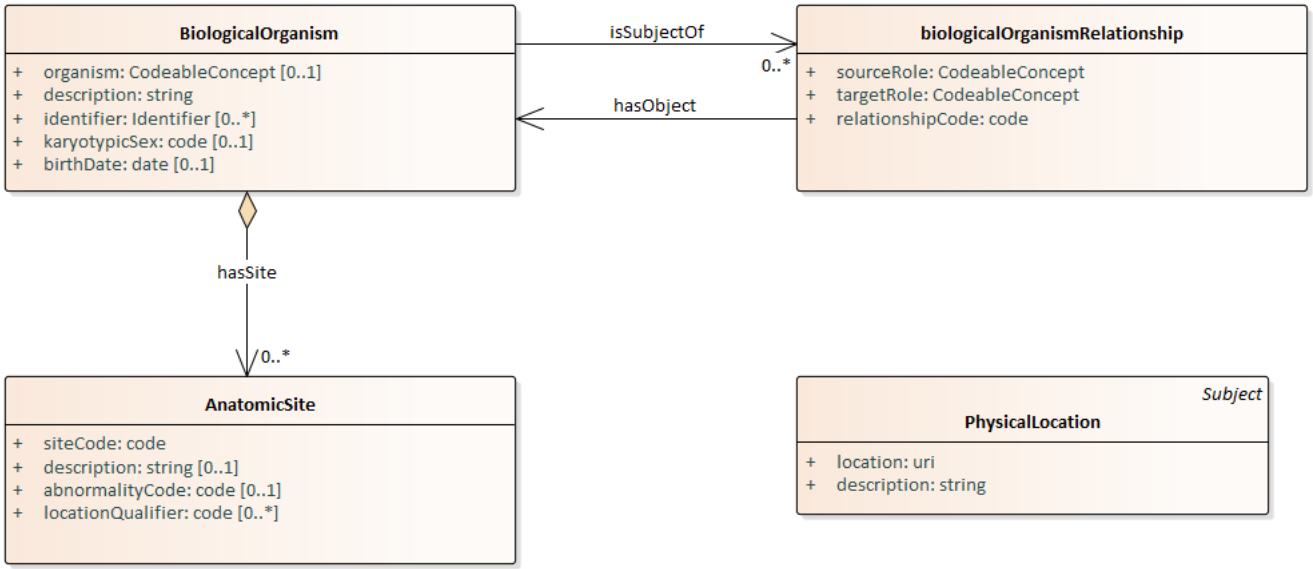

EA
Conceptual IMs
Logical IMs
Export
XMI
XSLT
FHIR Logical Models (Structure Definitions)
Publishing Tooling
IGs

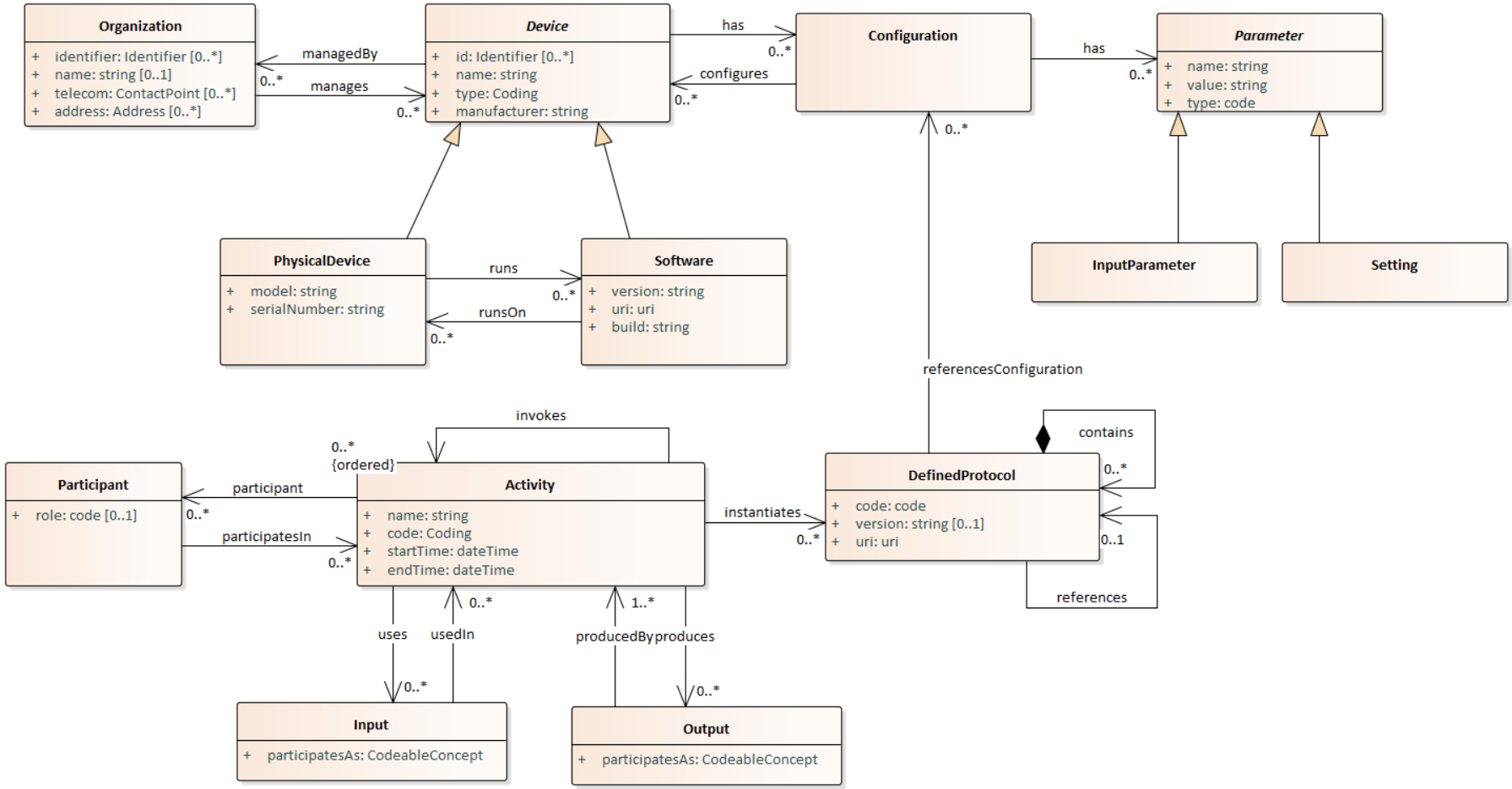
Things to Keep in Mind

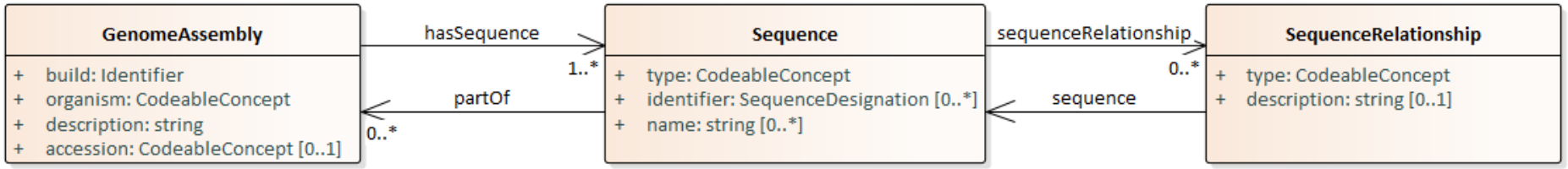
- Domain Analysis Model (DAM)
  - Captures semantics
  - Concepts, relationships
- Intended for SME, not (only) IT
- This is not an implementation

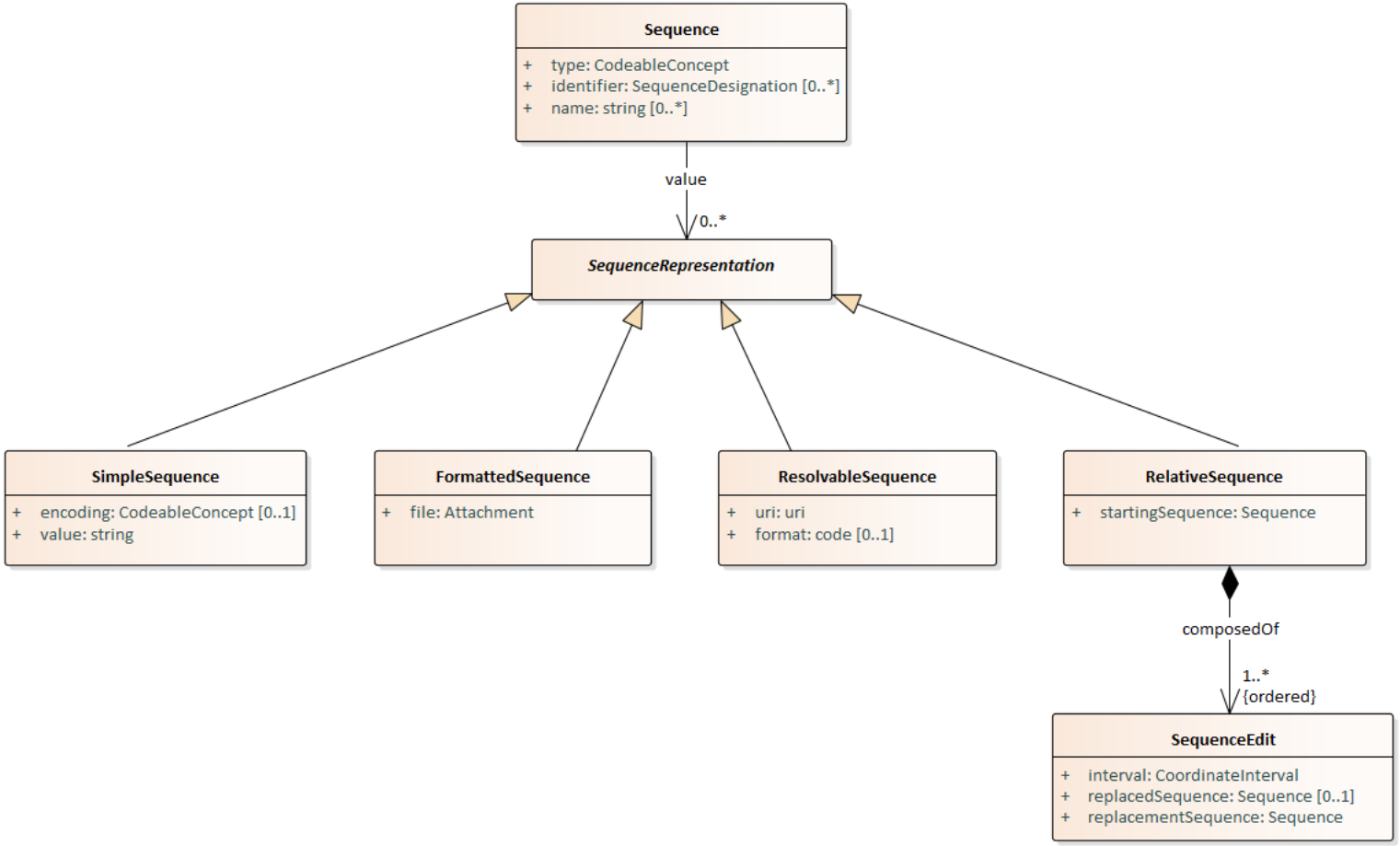
Things to Keep in Mind

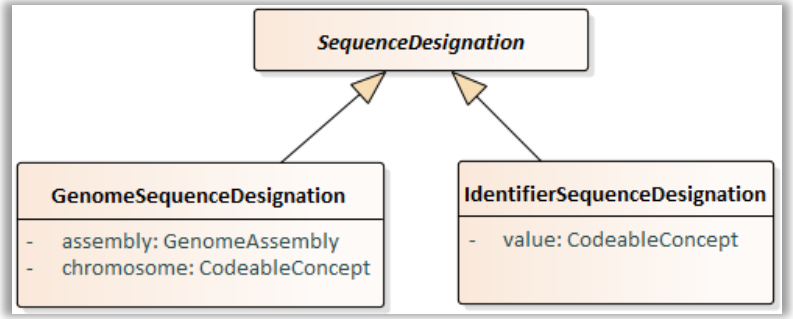
- Builds on existing work
  - Reuse when possible
  - Extend/remodel when necessary
  - Models, ontologies, implementations (FHIR, v2)
- Aligns to other work
  - GA4GH
  - SEPIO/Monarch
  - Specimen DAM
- This is not an implementation (FHIR)

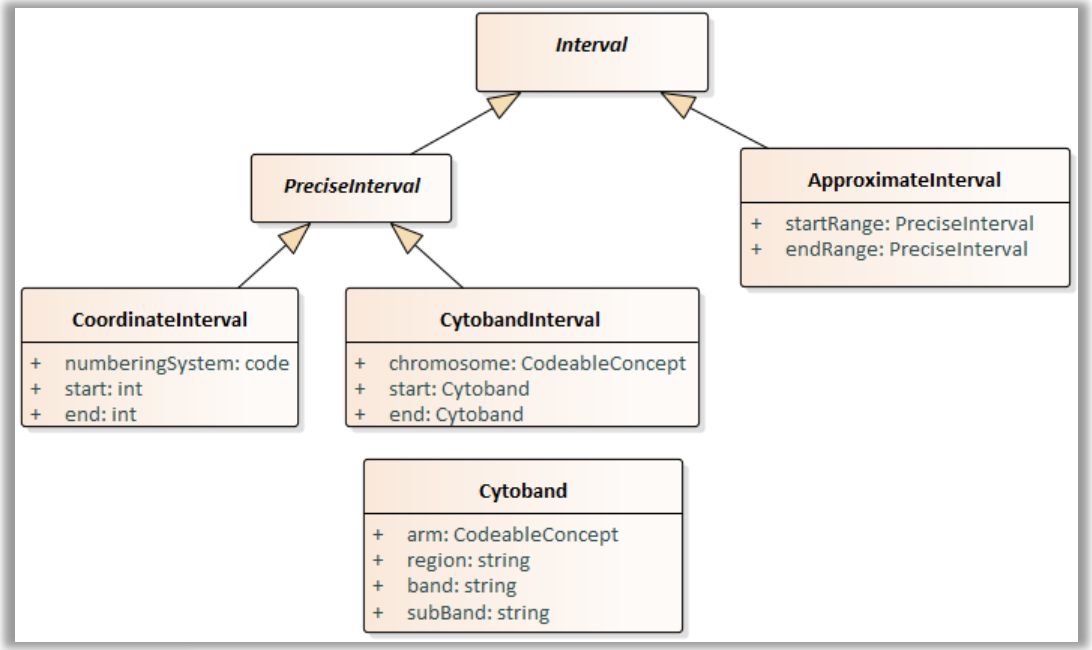

BiologicalOrganism
+ organism: CodeableConcept [0..1]
+ description: string
+ identifier: Identifier [0..*]
+ karyotypicSex: code [0..1]
+ birthDate: date [0..1]
isSubjectOf
0..*
hasObject
biologicalOrganismRelationship
+ sourceRole: CodeableConcept
+ targetRole: CodeableConcept
+ relationshipCode: code
hasSite
0..*
AnatomicSite
+ siteCode: code
+ description: string [0..1]
+ abnormalityCode: code [0..1]
+ locationQualifier: code [0..*]
Subject
PhysicalLocation
+ location: uri
+ description: string

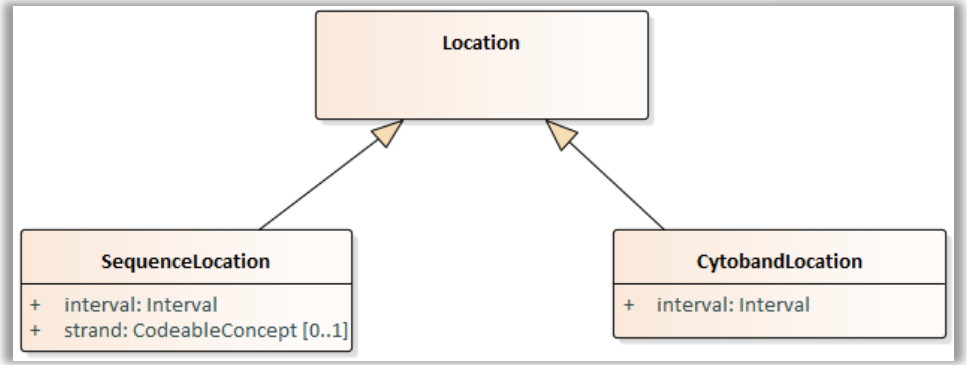

Organization
+ identifier: Identifier [0..*]
+ name: string [0..1]
+ telecom: ContactPoint [0..*]
+ address: Address [0..*]
managedBy
0..*
manages
0..*
Device
+ id: Identifier [0..*]
+ name: string
+ type: Coding
+ manufacturer: string
has
0..*
configures
0..*
Configuration
has
0..*
Parameter
+ name: string
+ value: string
+ type: code
0..*
PhysicalDevice
+ model: string
+ serialNumber: string
runs
0..*
runsOn
0..*
Software
+ version: string
+ uri: uri
+ build: string
InputParameter
Setting
referencesConfiguration
invokes
0..*
{ordered}
Participant
+ role: code [0..1]
participant
0..*
participatesIn
0..*
Activity
+ name: string
+ code: Coding
+ startTime: dateTime
+ endTime: dateTime
instantiates
0..*
DefinedProtocol
+ code: code
+ version: string [0..1]
+ uri: uri
contains
0..*
0..1
references
0..*
uses
usedIn
1..*
producedBy
produces
0..*
Input
+ participatesAs: CodeableConcept
Output
+ participatesAs: CodeableConcept

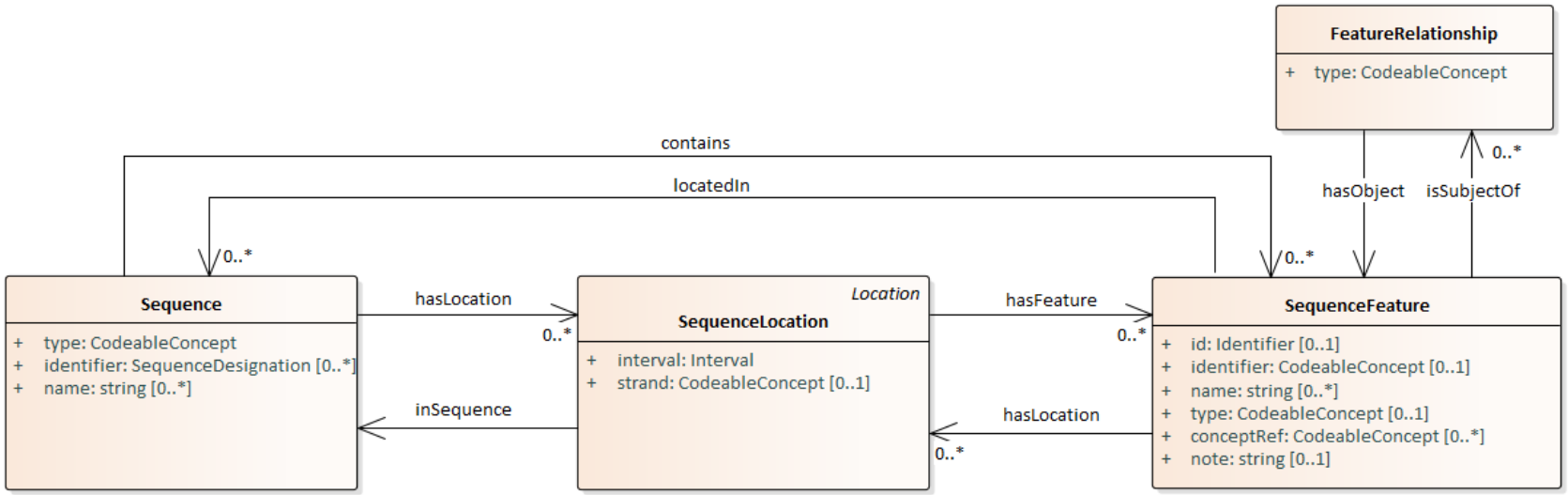

GenomeAssembly
+ build: Identifier
+ organism: CodeableConcept
+ description: string
+ accession: CodeableConcept [0..1]
hasSequence
1..*
partOf
0..*
Sequence
+ type: CodeableConcept
+ identifier: SequenceDesignation [0..*]
+ name: string [0..*]
sequenceRelationship
0..*
sequence
SequenceRelationship
+ type: CodeableConcept
+ description: string [0..1]

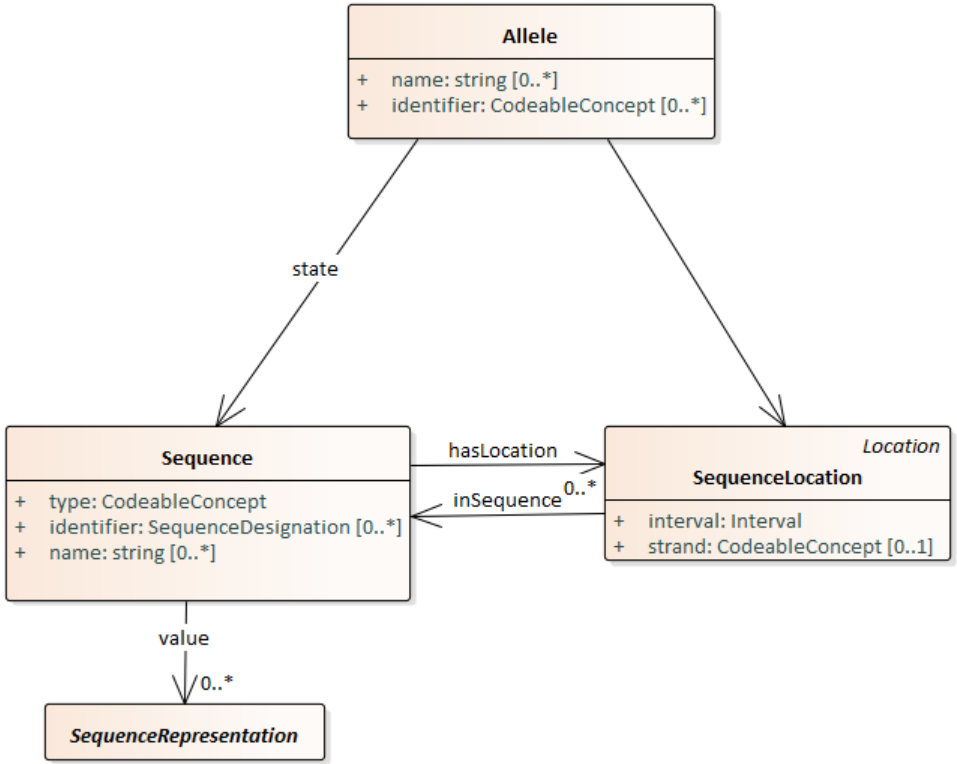

Sequence
+ type: CodeableConcept
+ identifier: SequenceDesignation [0..*]
+ name: string [0..*]
value
0..*
SequenceRepresentation
SimpleSequence
+ encoding: CodeableConcept [0..1]
+ value: string
FormattedSequence
+ file: Attachment
ResolvableSequence
+ uri: uri
+ format: code [0..1]
RelativeSequence
+ startingSequence: Sequence
composedOf
1..*
{ordered}
SequenceEdit
+ interval: CoordinateInterval
+ replacedSequence: Sequence [0..1]
+ replacementSequence: Sequence


SequenceDesignation
GenomeSequenceDesignation
- assembly: GenomeAssembly
- chromosome: CodeableConcept
IdentifierSequenceDesignation
- value: CodeableConcept


Interval
PreciseInterval
ApproximateInterval
+ startRange: PreciseInterval
+ endRange: PreciseInterval
CoordinateInterval
+ numberingSystem: code
+ start: int
+ end: int
CytobandInterval
+ chromosome: CodeableConcept
+ start: Cytoband
+ end: Cytoband
Cytoband
+ arm: CodeableConcept
+ region: string
+ band: string
+ subBand: string

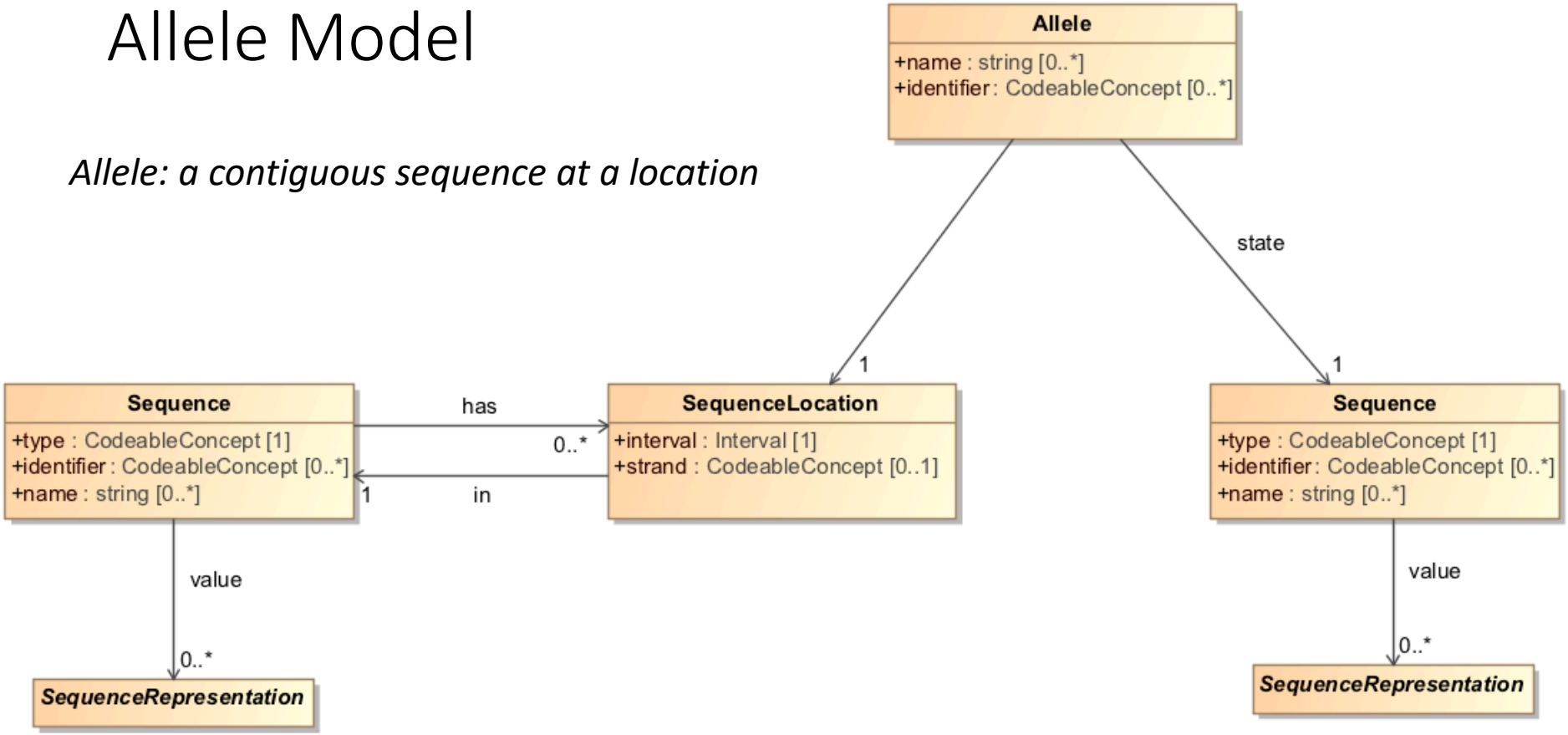

Location
SequenceLocation
+ interval: Interval
+ strand: CodeableConcept [0..1]
CytobandLocation
+ interval: Interval


FeatureRelationship
+ type: CodeableConcept
contains
locatedIn
0..*
hasObject
isSubjectOf
0..*
0..*
Sequence
+ type: CodeableConcept
+ identifier: SequenceDesignation [0..*]
+ name: string [0..*]
hasLocation
0..*
inSequence
Location
SequenceLocation
+ interval: Interval
+ strand: CodeableConcept [0..1]
hasFeature
0..*
hasLocation
0..*
SequenceFeature
+ id: Identifier [0..1]
+ identifier: CodeableConcept [0..1]
+ name: string [0..*]
+ type: CodeableConcept [0..1]
+ conceptRef: CodeableConcept [0..*]
+ note: string [0..1]


Allele
+ name: string [0..*]
+ identifier: CodeableConcept [0..*]
state
Sequence
+ type: CodeableConcept
+ identifier: SequenceDesignation [0..*]
+ name: string [0..*]
hasLocation
0..*
inSequence
Location
SequenceLocation
+ interval: Interval
+ strand: CodeableConcept [0..1]
value
0..*
SequenceRepresentation

Allele Model

Allele: a contiguous sequence at a location



Where is the allele located?

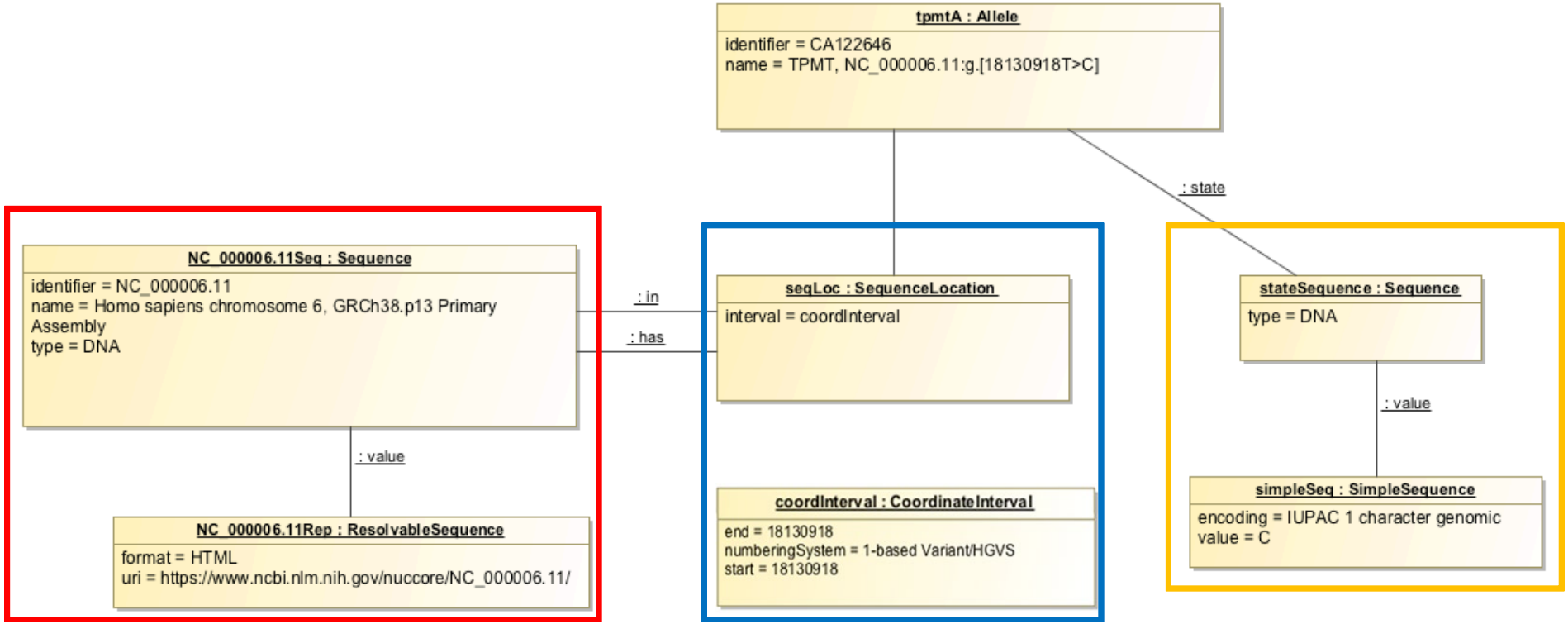
What was found at this location?

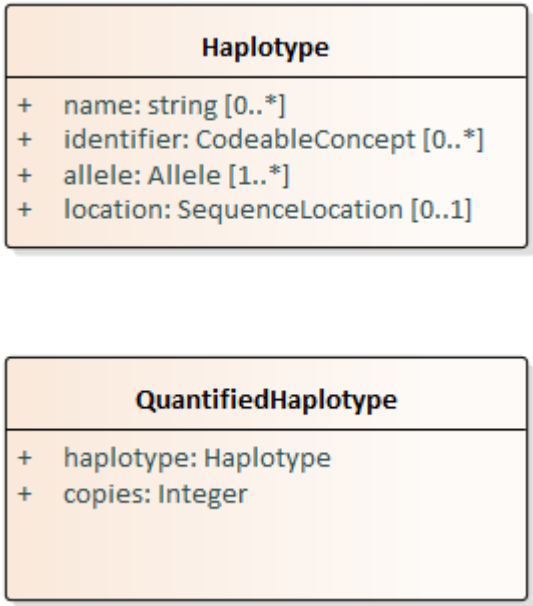
Example Allele: SNV

Allele: a contiguous sequence at a location

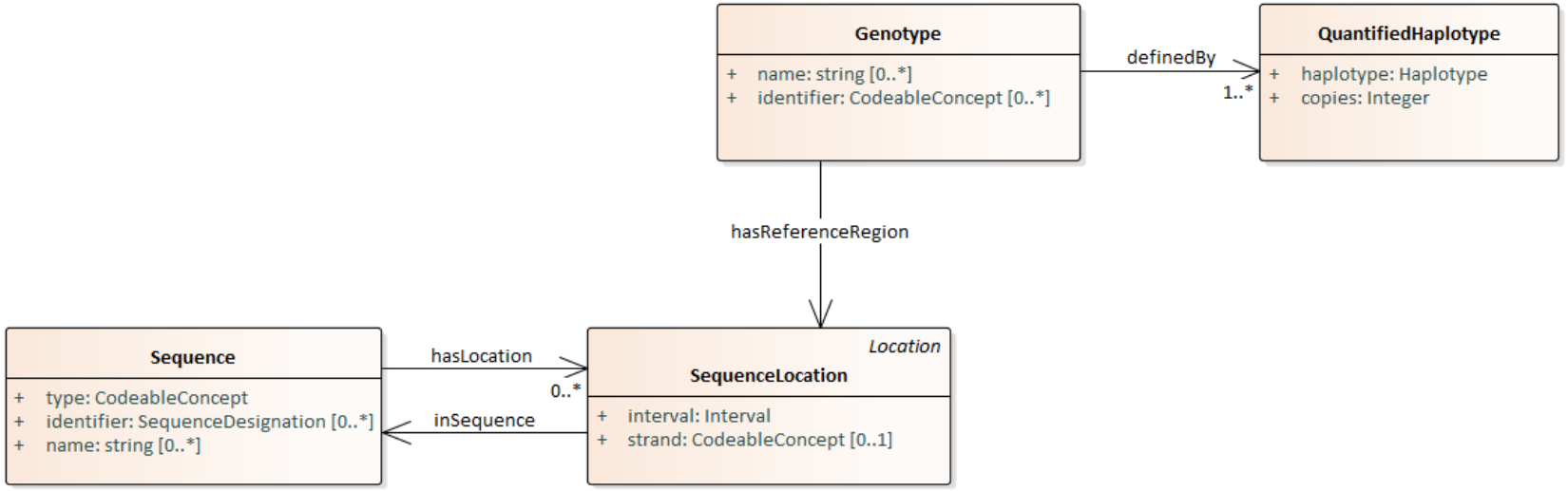
NC_000006.11:g.18130918T>C

Sequence | Location | State




Haplotype
+ name: string [0..*]
+ identifier: CodeableConcept [0..*]
+ allele: Allele [1..*]
+ location: SequenceLocation [0..1]
QuantifiedHaplotype
+ haplotype: Haplotype
+ copies: Integer

Draft


Genotype
+ name: string [0..*]
+ identifier: CodeableConcept [0..*]
definedBy
1..*
QuantifiedHaplotype
+ haplotype: Haplotype
+ copies: Integer
hasReferenceRegion
Sequence
+ type: CodeableConcept
+ identifier: SequenceDesignation [0..*]
+ name: string [0..*]
hasLocation
0..*
inSequence
Location
SequenceLocation
+ interval: Interval
+ strand: CodeableConcept [0..1]

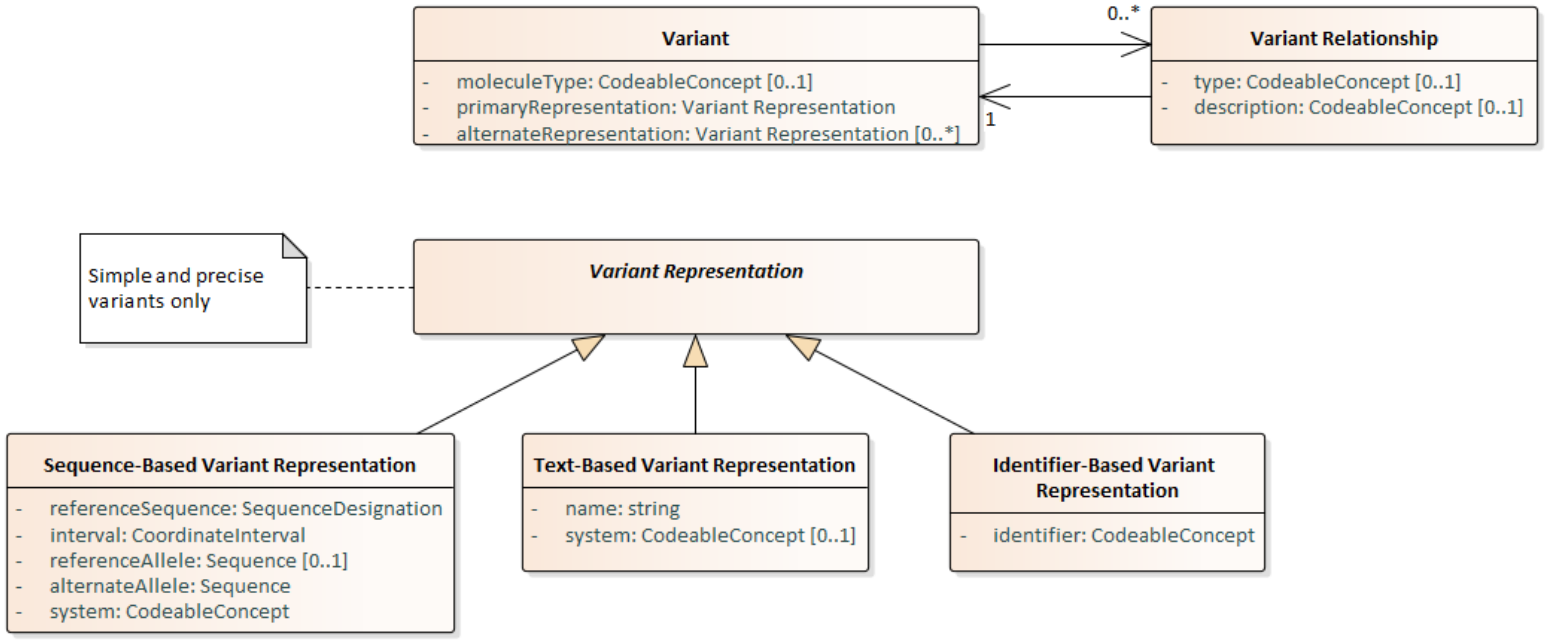
Definitional Variation

In scope

- Genomic
- Known location
- Coordinate-based
- Known change

Out of scope (for now)

- Non-genomic
- Unknown location
- Non-coordinate based
  - Cytobands
  - Features
- Unknown or inexact change
- Categorical/variation classes
- Structural (unless it can be represented as an "in scope")

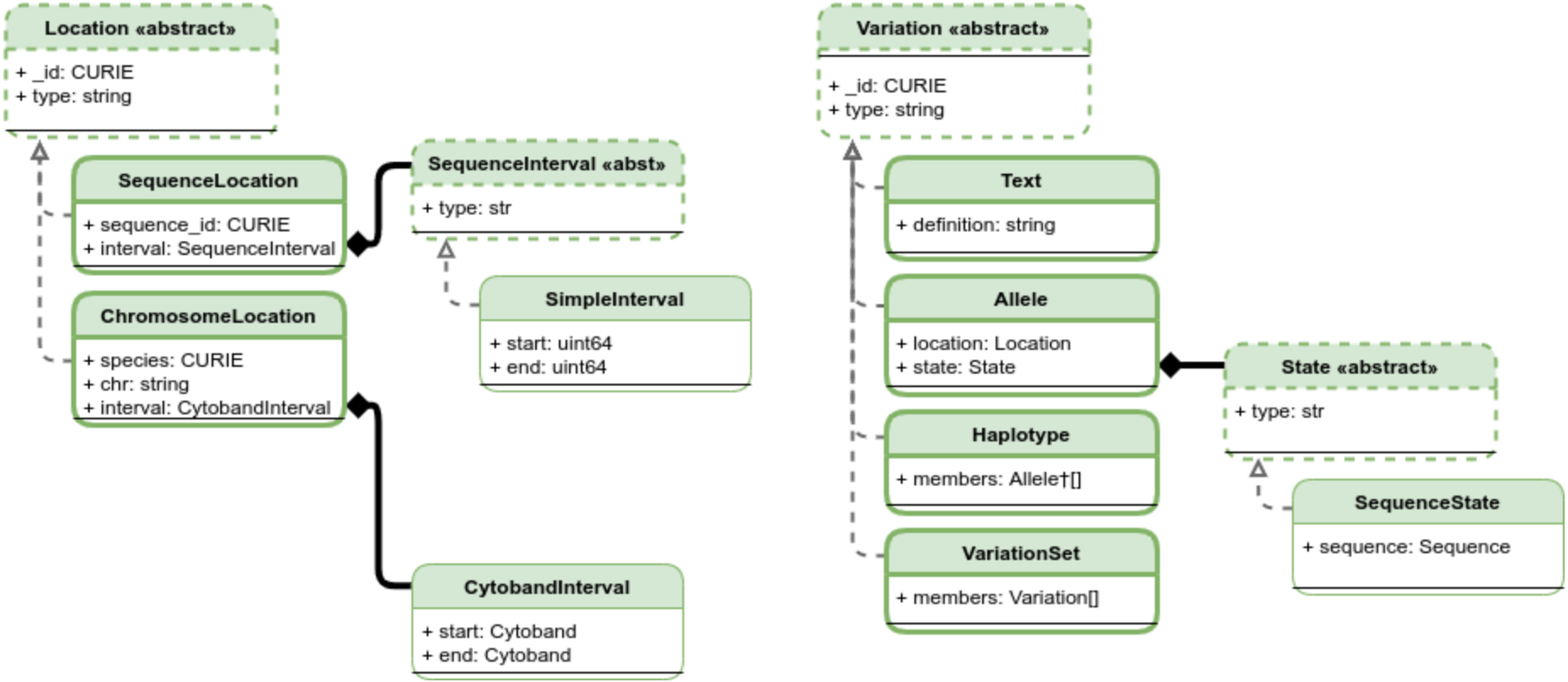


VCF Style
HGVS Style
SPDI Style
gnomAD
VRS

HGVS Nomenclature
Named (e.g., star alleles)
ISCN Nomenclature
HLA Nomenclature
Legacy names

Examples of codes for use in system

GA4GH VRS


Location «abstract»
+ _id: CURIE
+ type: string
SequenceLocation
+ sequence_id: CURIE
+ interval: SequenceInterval
ChromosomeLocation
+ species: CURIE
+ chr: string
+ interval: CytobandInterval
SequenceInterval «abst»
+ type: str
SimpleInterval
+ start: uint64
+ end: uint64
CytobandInterval
+ start: Cytoband
+ end: Cytoband
Variation «abstract»
+ _id: CURIE
+ type: string
Text
+ definition: string
Allele
+ location: Location
+ state: State
Haplotype
+ members: Allele†[]
VariationSet
+ members: Variation[]
State «abstract»
+ type: str
SequenceState
+ sequence: Sequence

Definitional Variant in Context

- Gene
- Disease/condition
- Test (including Panel)
- Observed variant (haplotype, genotype)
- Genomic source class
- Specimen (including subject and metadata)

Inferred/computed (w/ annot)

- Variant length
- DNA change (HGVS)
- DNA change type
- AA change (HGVS)
- AA change type
- Cytogenomic nomenclature

Measured data

- Allelic read depth
- Sample allelic frequency
- Array CGH ratio

Interpretations

- value (e.g., present/absent)
- Allelic state (e.g., heterozygous)
- Copy number
- Chrom CN change type

Cross-Refs <= *are these definitional?*

- Variation code
- dbSNP ID

Related/metadata

- Genomic source class (e.g., germline)
- Gene studied (HGNC/NCBI ID) <= *used as Location?*
- Method
- Ref seq assembly <= *captured in IM via GenomeAssembly*

Other

- Transcript ref seq
- Complex var type (e.g., compound het)

Prioritizing Next Steps

- Other types of Variation
- Information related to Variation
- Measured, computed, inferred, interpreted information
- Observations of <things>

- Develop a logical model for Def Var, Seq, and related classes
  - Identify boundaries between resources
- Transform logical model into FHIR logical models
  - Facilitate review and feedback
- Integrate into IG, propose changes to resource(s)

Bob's Crystal Ball

- CG Resources (Definitional)
  - MolecularSequence
    - All types of sequences (DNA, RNA, protein)
    - Representations and relationships
    - GenomeAssembly
    - Annotations and Features
  - MolecularVariation
    - All types of variation (SNVs, CNVs, structural; alleles; haplotypes; genotypes)
    - Representations and relationships
    - Annotations
- Generic Resources
  - Assertion
    - Based on SEPIO
    - Profiled for Genomic Implications and inferences related to sequences and variations