FHIR genomics – Overview of comments* on 9/2016 ballot

Amnon Shabo (Shvo)

Co-chair, HL7 Clinical Genomics Work Group

Co-editor, HL7 CDA R2 / CCD / Pedigree / GTR

14Dec2016

To the CGWG FHIR Subgroup,

I've sprinkled in comments (in blue) touching on a number of the things that Amnon covered in his presentation yesterday. s far as they go, I hope my comments are clear and constructive

There's much food for thought in Amnon's document, both his critique and his list of future expansion. My general takeaway is that changes that will — and should — happen to arrive at STU4 should reflect a combination of absorbing internal contributions such as Amnon's and feedback from real-world pilots sites. We'd really be remiss not to take both into account to craft a better STU4 specification!

So onward …

David Kreda
david.kreda@gmail.com

* For the detailed comments, please refer to the submitted document

Should Sequence be a resource?

This issue should be examined by two main criteria sets:

1. Domain requirements and domain information modeling

2. FHIR requirements for creating a new resource

Domain requirements and respective information modeling

- Can a Sequence resource naturally represent non-sequencing data sets, e.g., cytogenetics, expression data, mass spec data for proteomics, etc.?

The initial target of Sequence encompassed DNA, RNA, and AA.
Other types of data, involving sequencing with other dimensions might be a stretch.

- If Sequence is the only base resource in FHIR in the omics domain, how could the other types of omics data be represented?

- Need a more basic & common structure, e.g., genetic/genomics locus
  - Then, any type of omics data could profile that base resource
  - And thus – share a common semantics

A sensible aspiration, though we should want several examples to see what this means in practice.
And we will want to show that we can avoid creating a profusion of new profiles that would tyrannize implementers over small differences!

Notably, EHR vendors have said that profile proliferation is impractical for them to do.

FHIR requirements for creating a base resource (Resource appropriateness*)

Does the resource meet the following characteristics?

Quick review where
✔ = Sequence satisfies characteristic

- *Must*
  - *Represents a well understood, "important" concept in the business of healthcare* ✔
  - *Represents a concept expected to be tracked with distinct, reliable, unique ids* ✔
  - *Reasonable for the resource to be independently created, queried and maintained* ✔
- *Should*
  - *Declared interest in need for standardization of data exchange* ✔
  - *Resource is expected to contain an appropriate number of "core" (non-extension) data elements (in most cases, somewhere in the range of 20-50)* ✔
  - *Have the characteristics of high cohesion & low coupling – need to explore whether coupling is good some places, not elsewhere – layers from Bo's document*

Somewhat challenging to "know" in the abstract.

* http://wiki.hl7.org/index.php?title=Template:FHIR_Resource_Proposal

FHIR requirements for creating a base resource are not met

- The proposed Sequence resource is a mixed-bag of sequences, variants, structure variants and more
  - The criterion of “high cohesion” is not met

A mixed bag is not unique to Sequence.
FHIR is not about perfect normalization.
It aims to satisfy real world needs & looks to implementers to see what is practical and effective!
(Stashing stuff in non-genomic Observation profiles, for example, is very FHIR-like.)

- The proposed Sequence resource includes variants found by NGS
  - The criteria of “reliable” (and “naturally identifiable” per correspondence with Lloyd McKenzie) are not met

Does DNA sequence data not qualify as the most naturally identifiable human data.
the reliable part seems a function of now pretty good lab technology. What am I mis

  - Also because of the complexity of a sequence referenced from a sequence but not reusing the same structure (i.e., Sequence points ReferenceSeq)

A reference sequence should be representable in the same manner as a specimen-observed sequence.
But if this is actually a problem in the STU3 spec, is it insuperable to “fix” with a tweak?

Sequence design principles

- Sequence should hold merely sequence data (observed, reference,…)
- Sequence should not contain any information that is the result of downstream analysis (i.e., beyond assembly of the sequence itself)
- Sequence should include metadata about the sequence, e.g., quality, provenance, pointer to repository holding the full sequence, etc.
- Sequence could encapsulate (inline) a sequence portion if it's key to its association to phenotype and not larger than limits posed by FHIR
  - In which case, native formats (i.e., any bioinformatics format commonly used in the industry to represent sequences) should be used
  - HL7 Sequence should not provide yet another format to represent sequences

ence should represent data in legacy formats that force NGS / WGS results to be represented in computationally inefficient ways, e.g., that would disadvant

Long & short term changes - Sequence

- As aforementioned it is proposed to design a more basic and common resource to all omics data types, so that Sequence can be a profile over that basic resource (e.g., locus)

Per my comment on slide 3, some elementary profiles would be useful to show the bang-for-the-buck of such an abstraction.

- In the short term, if this proposal is not accepted, then it is proposed to make changes in the Sequence resource that are valid even if Sequence is designed as a profile (see next slide)

It IS a change, but not remotely a big change nor calamity for implementers. That is, if the case can be made on the merits, adjusting should not be hard: STU3 Sequence resource now is an STU4 SequenceOmics-profile on an STU4 Omics resource.

Perspective: I believe that pilot implementer of STU3 will request other changes that are more disruptive in fact to better suit their real world needs. But we should welcome that mostly if this allows us to graduate our work to FMM2.

Summary of proposed changes - Sequence

- Remove:
  Move the following elements (including their nesting elements/attributes) from the Sequence resource to the Observation-genetics profile:
  - ReferenceSeq
  - Variant
  - Repository.variantSetId
  - StructureVariant

Step 1. Let's cross-check this with the reconciled STU3 to see which of Amnon's STU3 comments have already been accepted.

Step 2. Then we can review the rest and see what "organically" can occur as we evolve the Current Build.

- Change:
  Change the attribute name "observedSeq" to "sequence"
- Constrain:
  sequence (name changed from observedSeq)

Per my comments on slide 6, dropping string is (IMO) not NGS-friendly.

  - This attribute is currently of type string, but it should be constrained to a common bioinformatics format for sequences as described above
  - A number of common formats could be allowed
  - A bioinformatics format could be constrained in its usage within this attribute
- Add:
  - Add a category attribute to define if a Sequence instance is an observed sequence or a reference sequence
  - Alternatively, this addition can be avoided, by looking at the attributes 'patient' or 'specimen' – if they are populated then this is an observed sequence, otherwise it's a reference sequence of some kind (determined by other attributes)

I don't favor allowing a logical test of "emptiness" to tell us what's in the payload.
Something this fundamental should not be left to downstream implementers (and mistakes)!

Variants are everywhere…

- Variants appear in both
  - Sequence (resource)
    - Variant
    - variantId (in staging site: variantSetId)
    - StructureVariant (removed in staging site)
  - Observation-genetics (profile)
    - DNA change
    - Amino acid change
    - more
- Propose to consolidate all information about a variant in one structure (‘GenomicsObservation’ profile)

Musing: VCF carries low-level “interpretive” information, i.e., they call out variants. If legacy formats mix data and interpretation (and they do by design), legacy bioinformatics formats may be said to violate inherently any rule we would set that would separate data from interpretation. Maybe we can’t serve both simultaneously???

The controversy/history of this topic is as follows:

We did ourselves no favor some while ago in thinking we could (among other things) optimize payload size in Sequence by allowing using variant codes to represent data. This helped create the impression - even reality - that the STU3 resource looks like a “mixed” bag

Given the choice of having to stick to one format with zero confusion, I would favor doing strings for Sequence (resource or profile). Then, IF there is pragmatic griping from implementers, deal with it.

Variants

*** I address what the CGWG call raised during Q&A: the recommendation is to treat Observation as the place for base level "informatio about variants only and move "higher level" clinical interpretation to another resource. It is offered to avoid yet another "mixed bag" (etc. payload. The intent of the idea is a good one. I don't know how Observation is treated elsewhere — other groups may AOK with being " or maybe they simply do not have the challenge of so many layers! Anyway, while I am in favor of "bright line" boundaries as proposed, have to be reliably reproducible by real world implementers. Can such boundaries be reliably seen so that implementers are not flipping

- Variants in the Sequence resource should be removed See relevant comments on slides 6, 8, and 9
  - In particular, per the specification, Sequence variants are meant to represent the sequence and are not intended to represent clinical-related data
  - The above is an attempt to suggest new formats for sequencing, which is out-of-scope for HL7 Clinical Genomics, and in addition adds yet another format to several existing formats in bioinformatics

- All information relating to variants should be held in one placeholder; best is in the Genomics Observation profile See above ***

- In principle, variant's interpretations should not be held as part of the variant, however, since the 'observation-geneticsInterpretation' extension has been restructured as a reference to a related observation, this could stay, assuming:
  - The use of the base Observation.interpretation attribute is explicitly disallowed
  - Extension points to a related GenomicsPhenotype (Observation profile – TBD)

DiagnosticReport-genetics

- Genomics tests are not necessarily diagnostic (e.g., carrier, prenatal, HLA)
  - Therefore, propose to call this profile - “GenomicsTestReport”
  - Propose to stick to ‘genomics’ assuming genetics is included in genomics

- Interpretation of an entire genetic test is held in this profile, as follows:
  - The use of FHIR DiagnosticReport.conclusion & codedDiagnosis attributes should be disallowed
  - The DiagnosticReport-geneticsAnalysis extension attribute holds the ‘integrated’ interpretation of all variants in a genetic test (or any other observations done as part of the test)
  - This extension should point to a related GenomicsPhenotype (profile - TBD)

Future work - adding document & phenotype

favor these suggestions but note that documents are harder to query than atomic elements,, Moreover, reports are (or s/b) "synthetic" combinations of elementary FHIR payloads, so it should be possible for a simple program to make other FHIR API calls and assemble the payloads for any of these proposed reports. We could consider the fallback that a good enough (open source) FHIR-based report writer that post-coordinates report production.

- Introducing a document structure
  - Port the HL7 Genetic Testing Report (CDA-based) to FHIR
  - GTR consists of sections; main section type represents a genomics test
    - by pointing to a GenomicsTestReport profile (currently - DiagnosticReport-Genetics)
  - GTR also has summary, test-background-info sections and more context
  - The summary section consists of an overall interpretation, summarizing several GenomicsTest interpretations in a study (e.g., hearing loss)

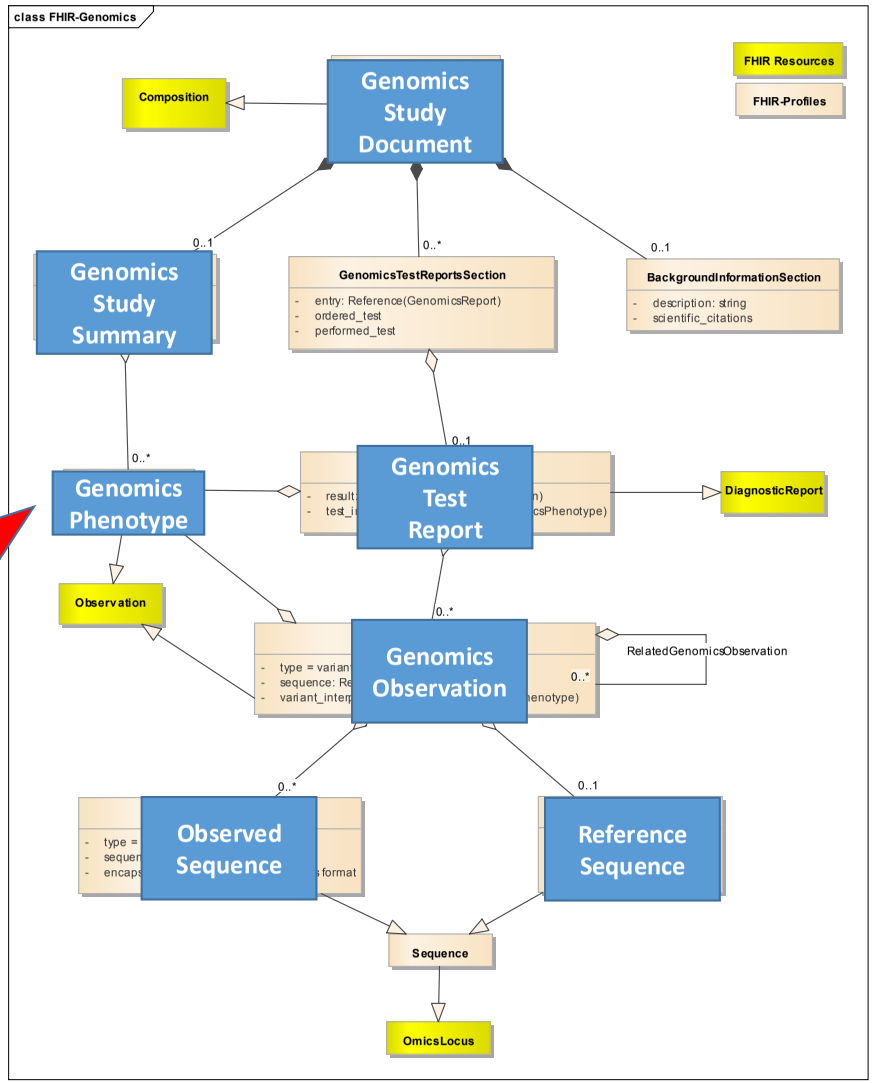
- Develop a more robust and expressive model for phenotypes
  - 'Phenotype statement' involving conditions, medications, etc.
  - Extend the related observation value set to represent 'gen-phen' semantics

The proposed roadmap

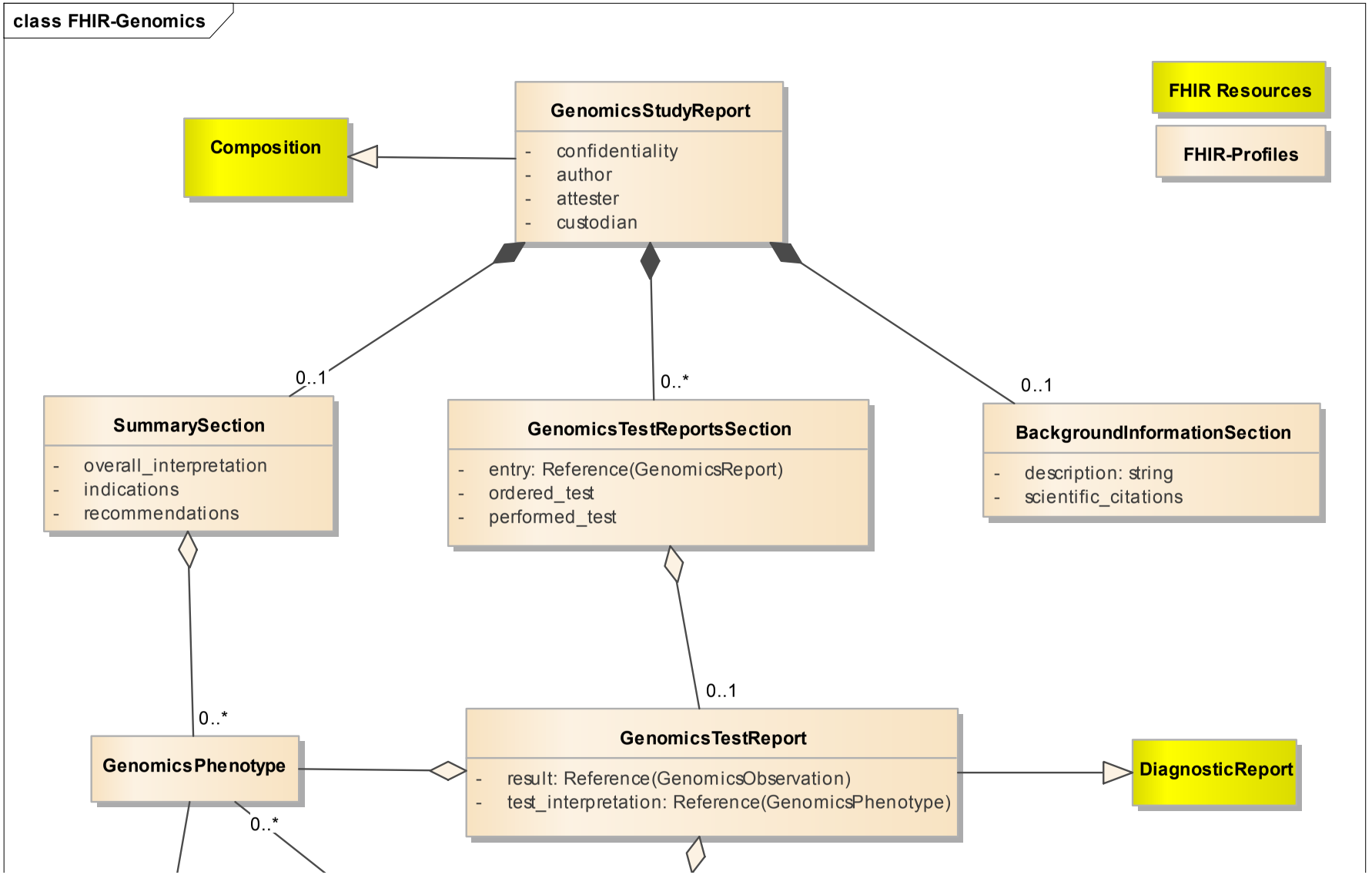
- GenomicsStudyReport document includes multiple genetic tests and summary with overall interpretation
- GenomicsTestReport represents a single genetic testing and holds its interpretation
- Variants reside solely in Genomics Observation, optionally pointing to observed and reference sequences
- Sequence can be both observed or reference, using the same construct

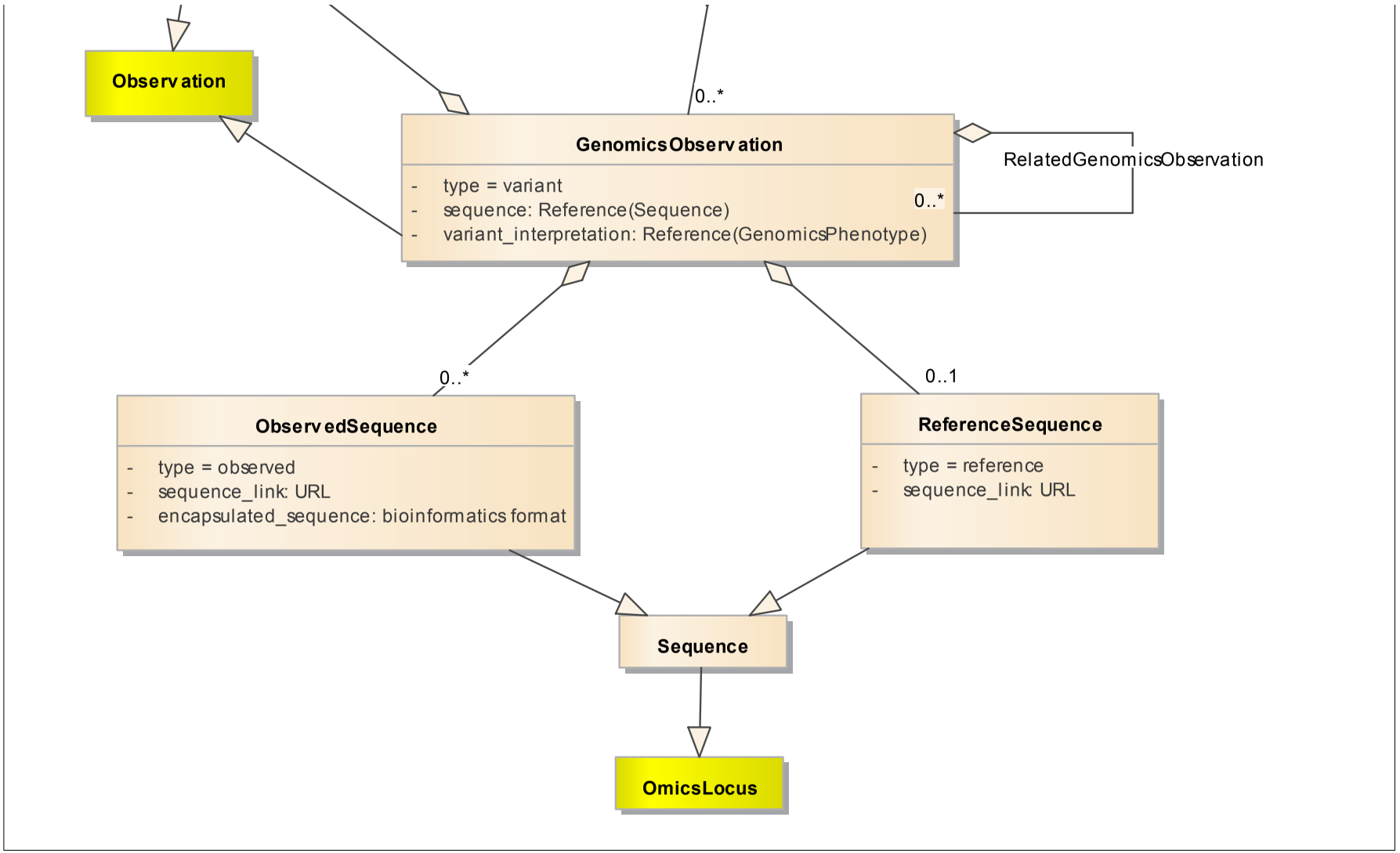
Same Phenotype construct is shared by the three levels of interpretation

This demonstrates how light weight DIM work could guide specifications without having to wait for a complete or perfect model. This would strike the right balance between modelers and GTD ("getting things done")



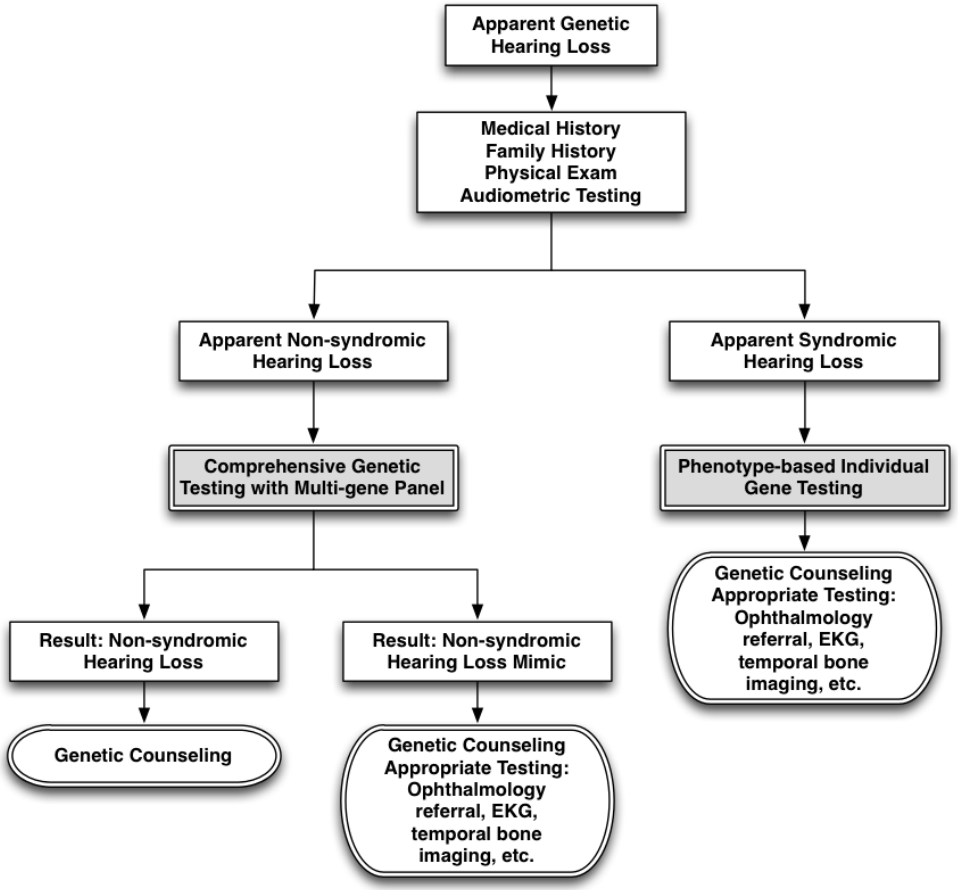
* Skeletal & conceptual model, for illustration only


class FHIR-Genomics
FHIR Resources
FHIR-Profiles
Composition
GenomicsStudyReport
- confidentiality
- author
- attester
- custodian
0..1
0..*
0..1
SummarySection
- overall_interpretation
- indications
- recommendations
GenomicsTestReportsSection
- entry: Reference(GenomicsReport)
- ordered_test
- performed_test
BackgroundInformationSection
- description: string
- scientific_citations
0..*
0..1
GenomicsPhenotype
GenomicsTestReport
- result: Reference(GenomicsObservation)
- test_interpretation: Reference(GenomicsPhenotype)
DiagnosticReport
0..*


Observation
0..*
GenomicsObservation
- type = variant
- sequence: Reference(Sequence)
- variant_interpretation: Reference(GenomicsPhenotype)
0..*
RelatedGenomicsObservation
0..*
0..1
ObservedSequence
- type = observed
- sequence_link: URL
- encapsulated_sequence: bioinformatics format
ReferenceSequence
- type = reference
- sequence_link: URL
Sequence
OmicsLocus

Example:
Hearing Loss Panel

- A panel is actually a study, similarly to the notion of study in medical imaging
- The study document can hold the context in the best way
- A document can also be easily exchanged
- Attestation (& signatures) and other ‘medical records’ prosperities are explicity represented

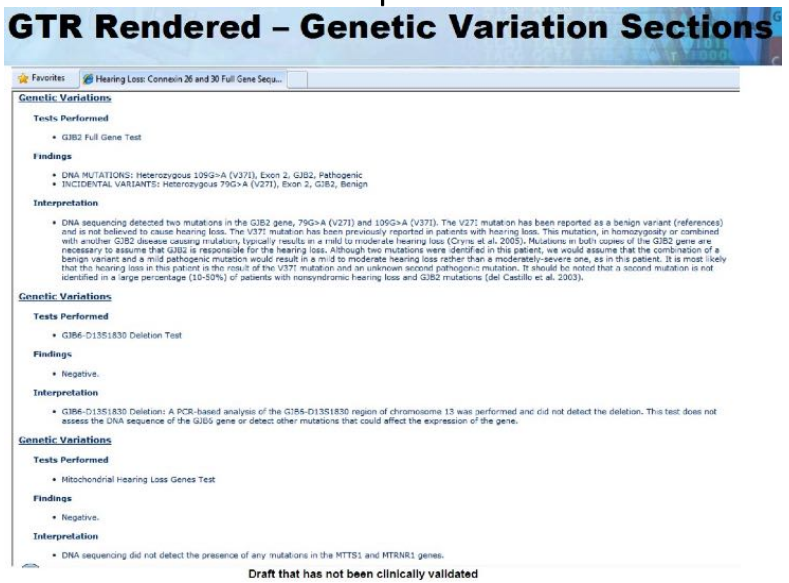


Source: Iowa Head and Neck Protocols

ARUP Hearing Loss Nonsyndromic Panel

1. GJB2 – Sequencing
2. GJB6 - 2 Deletions
3. Mitochondrial DNA - 2 Mutations

HL7 CDA-based Implementation Guide

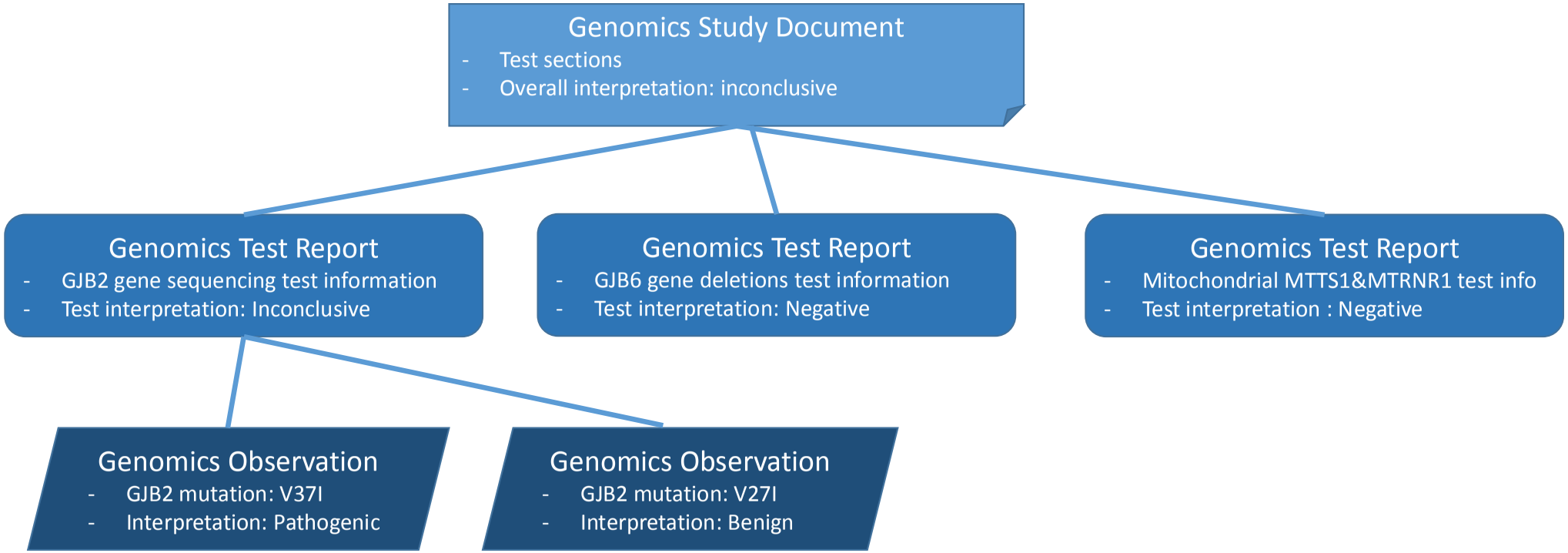


ARUP LABORATORIES | LABORATORY TEST DIRECTORY

Hearing Loss, Nonsyndromic Panel (*GJB2*) Sequencing, (*GJB6*) 2 Deletions and Mitochondrial DNA 2 Mutations

http://ltd.aruplab.com/Tests/Pub/2001992

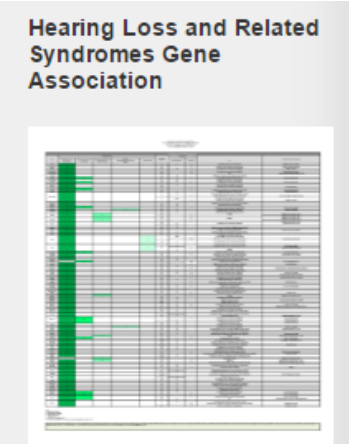
ARUP Hearing Loss Nonsyndromic Panel Example results (as used in HL7 v2 and GTR)



Studies get complex… e.g., OtoGenome™

- The OtoGenome™ Test targets individuals who have a diagnosed hearing loss whose underlying etiology has not yet been identified
- Goals & context expand to hearing loss and related syndromes
- OtoGenome™ Test includes 87 Genes

PARTNERS HEALTHCARE | PERSONALIZED MEDICINE



HLA study example