GA4GH Sequence Annotation

Minutes and Actions 2022

2020 | 2021 | 2022 | 2023 | 2024 |

These are the minutes for the GA4GH Sequence Annotation meetings. This group is initially a joint project between the RNASeq and Variant Annotation teams.

For further information, please visit the GA4GH website at https://ga4gh.org.

Table of Contents

Meeting Protocols

Minutes/Agendas
- 2022-11-03:
- 2022-10-06: Modelling activity
- 2022-09-15: Testing Model
- 2022-09-01: VRS & VOD Objects
- 2022-08-04: Transcript and Gene model
- 2022-07-21: CANCELLED
- 2022-07-07: Transcript and gene models
- 2022-06-16: Regulatory model V1
- 2022-06-02: Sequence Ontology
- 2022-05-19: CANCELLED
- 2022-05-05: Coding & Non coding transcript
- 2022-04-07: CANCELED
- 2022-03-17: Transcript types
- 2022-03-03: Evidence Representation
- 2022-02-17: Feature,Transcript,MANE and Polymerases
- 2022-02-03: UML diagram
- 2022-01-20: Location model for VA and VR
- 2022-XX-XX: Template

Meeting Protocols

- Please note that by participating in meetings, attendees agree to adhere to the GA4GH Standards of Professional Conduct.
- Meetings may be recorded for note-taking purposes. Recordings will be deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd

Global Alliance
for Genomics & Health

Minutes/Agendas

2022-11-03:

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: Evan C, Meghan Savage ,Emilior , Salem Bajjali ,Reggan Thomas , , , , , ,

	Actions Arising	Assigned To	Deadline
1	Modelling exercise - To be continued	All	
2			

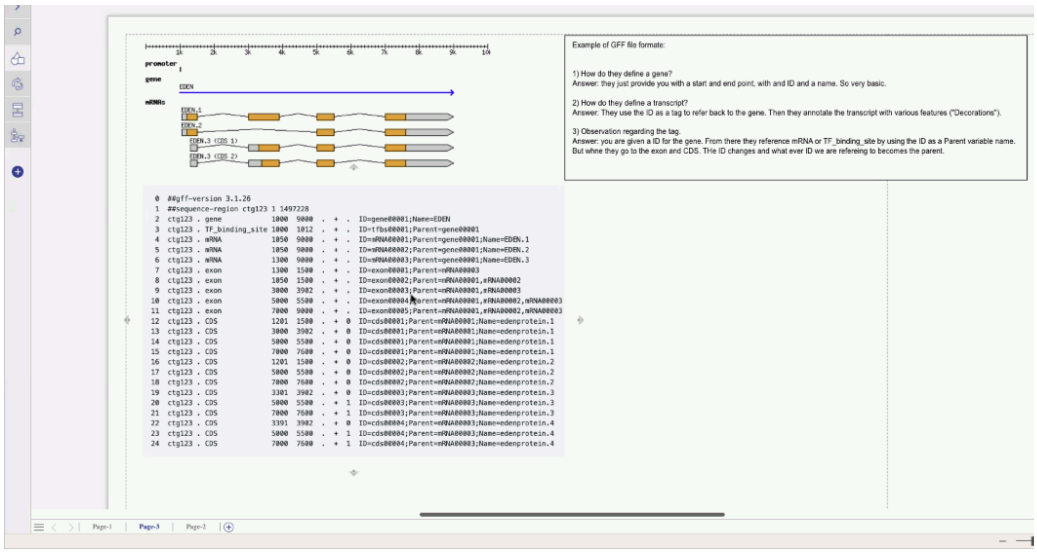
	Agenda Item	Person/Time
1.	Modeling exercise reminder	Evan (5 mins)
2.	Planning for connect meeting (scheduled for Wed Nov 30th at 10:30 - 12:00 EST)	All (45 mins)
3.		
4.		

Meeting recording:
https://us02web.zoom.us/rec/play/m-oRhGoA0TNqxdRQ9RqunkWDzD8_GVVv3C4KfaKTGlZ98zVG1n1chyB3aNbBYFZGBIV87oaOicIGYnPj.6I4ijPpXflT0xnU6?continueMode=true&_x_zm_rtaid=A1aNPHaETaiAGmjPXAvL9A.1667553237476.a35d053f6670ffd959eb034dd5e59901&_x_zm_rhtaid=329

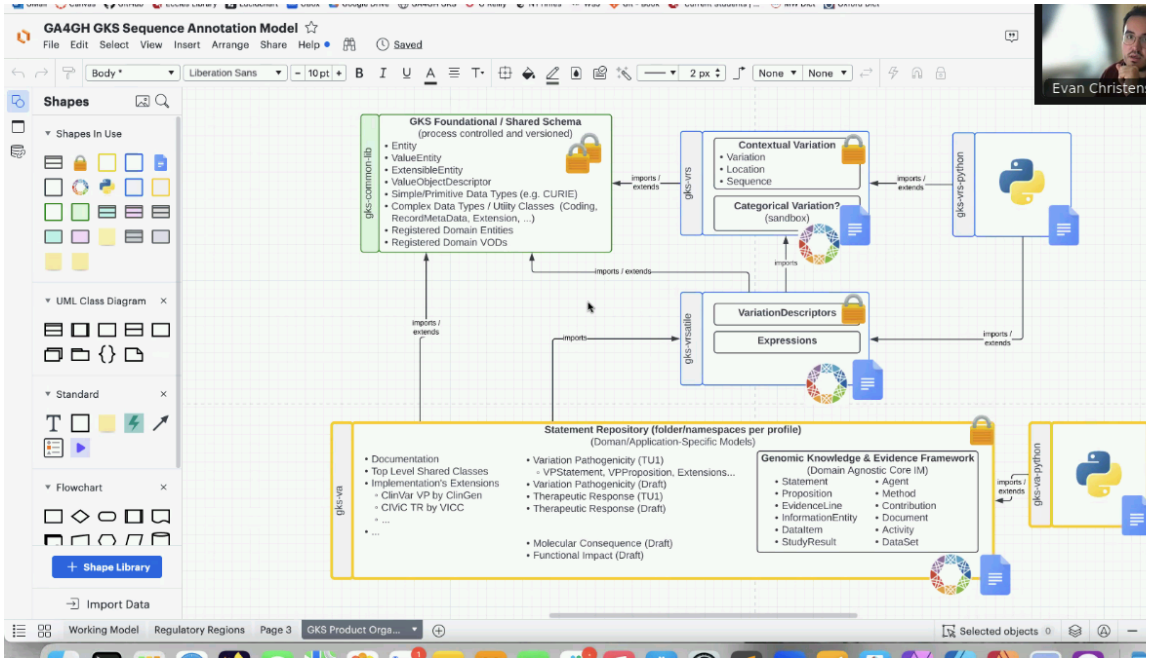
Meeting minutes:
Meeting transcript -
https://otter.ai/u/4eVUHriFmJWeD0hBtYxs6H2sIlQ?r=&i=0&f=oln&u=true&utm_source=va_chat_link_1

Emilior - Working on bio genome project is like this very big international effort to sequence all the eukaryotic species
EC - when you're making this model, modeling the biology versus modeling the information contained in the biological structure gets very tricky. So if we try to map everything back to the molecule, the you know, the, the polymer that's that that's going to be impossible.
SB - representing a gene is pretty much based solely on like coordinates.
EC - focused on transcript Gene.
 the basic feature the really as basic as you can get with like exon and alternative transcript

EC- Meghan to look into the modelling activity
EC - we're not modeling all of biology here. we're just giving people a place to model their own little sections of biology. So we're creating kind of a backbone, kind of a central hub that other groups can then model and attach to.



EC - we're really just looking at updating the GFS file. because that's really what the community uses.
When it comes to sequence sanitation. So the first eight columns, to be clear, like no one has an issue with the sequence Annotation is really good when it comes to the basic features of a sequence, an exon, a coding sequence, or regulatory region.



Dear colleagues,

Global Alliance
for Genomics & Health

A draft Genotype model for VRS is open for community review per our standard development process. Please review at your earliest convenience, and no later than **end of day July 26th**. The following resources may be considered in your review:

- The Pull Request containing the new JSON Schema and documentation code
- The readthedocs build rendering the documentation
- The static Jupyter Notebook implementing the schema to build a CPIC diplotype from ClinVar
- The online interactive notebooks implementing this schema
    - NOTE: (please select the “VRS_Python” kernel when using online notebooks)

Please do not hesitate to reach out if you have any questions.

Best regards,

Alex

EC - https://vrsatile.readthedocs.io/en/latest/value_object_descriptor/index.html

EC - idea of omics research?external sequence, sequence isn't necessarily human that affects health.

EC - how are we modeling that ?

I would say for genomics but then there's also like expose omics there's been a genomics meta genomics all this other extra stuff looking down there

2022-10-06: Modelling activity
Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: Salem Bajjali ,Francis Ouellette , Reggan Thomas , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time

1.	Simple modeling activity Bring: ● Yourself ● Blank paper & pen or your preferred charting software ● A way to snap a picture or scan a page if you're using paper Activity instructions: ● Construct a model/diagram of the simple transcript found here representing entities, attributes, and inheritance. ● No need to adhere to any specific modeling standard/framework/language as long as your thought process is written somewhere. ● This activity is meant to measure consensus when it comes to core entities and relationships so feel free to take liberties with any extra information that is not defined in the slides themselves. ● Send a picture/image/PDF of your model to evan.christensen@utah.edu. ● We'll compare everyone's models and discuss similarities & differences in our next meeting.	20 mins (5 mins explanation/15 mins individual work)
2.	Strategizing a survey of GFF files - Column 9	15 mins
3.	What would a FHIR-first approach to Sequence Annotation look like?	15 mins
4.	Meeting ends 5 mins before the end of the hour	

Meeting recording:https://us02web.zoom.us/rec/play/1oza0Xt894TmqH4Egf7nE6XW4SfXFk0mA-V-F57cbUDig5uF9M_sgansmloOOx901l84OhLL68SnzVEh.13juEQg4Y1Zub1yQ?autoplay=true&startTime=1665064648000

Meeting minutes:
Transcript - https://otter.ai/u/yLZ95ytwUeiYymeBAJY8gBjWfNw?utm_source=va_chat_link

FO - drawing and a list of metadata ?
EC - if it's just as simple as entities or inherit attributes
EC - Current focus is only human genetics
EC - we'll move on to my second pathogenic sequence, okay. And then and then model organisms
EC - we're still actually going back through the user stories to determine the treatment of genome annotation

FO- https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG1

EC - Are these codes unique to GenBank files?

FO - Its worldwide

EC - Would this just be kind of in the header of a GFI file ?
FO - need to do bit of a research

EC - The blue exon should the UTR be 1200. Technically, it's like 1001?
EC - we should probably change that to 1001 and then and 1010

Strategizing a survey of GFF files - Column 9

FHIR-first approach to Sequence Annotation look like?

http://hl7.org/fhir/us/core/STU5/StructureDefinition-us-core-patient.html

2022-09-15: Testing Model

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: , Omar Khan (NHS Digital) , Francis Ouellette, Larry Babb , Bob F , Salem Bajjali , , , , ,

	Actions Arising	Assigned To	Deadline
1	Send out modeling exercise	Sean/Evan	Next Tuesday Sep 20
2			

	Agenda Item	Person/Time
1.	Working session: testing the model against simple transcripts	Sean & Evan
2.		
3.		
4.		

Meeting recording:https://us02web.zoom.us/rec/share/VsslNFg__ySSZrqVKy8MBck0AdZxapcAYA0fhi7Xx3quhPSW6E2cWBqM81wYv59p.-5JGSefh7k6za3jO?startTime=1663250318000
Passcode: hrRnX19^

Global Alliance for Genomics & Health

Meeting minutes:https://otter.ai/u/LKQRRJuh3WyA9jwdec5tSnhW4oc?utm_source=va_chat_link

SU - Walkthrough of SA annotation.

LB - if a "gene" is defined by the list of unique exon/intron/utr regions on a single reference sequence, then is it possible/likely for new “regions" within an existing gene to be discovered, added, removed, changed after the fact?

LB - should we defining physical gene definitions (associated with a unique reference sequence by a specific group) and the conceptual gene that is the canonical representation of all physical manifestations of the "same" gene?

BF - The concept of a “gene” can be defined functionally or based on its physical location (e.g., on a sequence), but I think it is often defined using BOTH. Furthermore, both of those underlying concepts (function, location) can be fuzzy, so there is some ambiguity in the definition.

BF - FWIW, I think like Larry (this is probably scary for both of us). I think about the smallest components first, then build up from there. That would give the person creating an instance of Gene the flexibility to define it using the components that are meaningful to their use case.

BF - I think we’re struggling with level of granularity. We are both trying to capture fuzzy concepts (gene) AND highly precise concepts simultaneously. Maybe we should consider splitting them.

SU - Are you thinking perhaps that the model is a set of regions with associated types (i.e. exon, promotor …) and the implementation of the model is where genes and transcripts are built from these regions? Maybe?

BF - I think so. A challenge is that any given location can be part of multiple annotated regions, so the model must support this and compositional objects built from those regions must be explicitly defined with this in mind.

BF - Does the definition of a Gene depend on the label (symbol) that we use as its name?
FO - No
SB - Is the sequence annotation model thats on lucid available on google docs?
EC - https://lucid.app/lucidchart/aeaabf97-a0cd-482e-9297-6a2f18f80ab4/edit?viewport_loc=-473%2C-1093%2C2963%2C1432%2CMV~PZNcMt~4A&invitationId=inv_a5041b50-7dc8-401e-9241-9013a5d27270#

2022-09-01: VRS & VOD Objects

Chair:Sean Upchurch, Karen Eilbeck

Global Alliance for Genomics & Health

Attendees “Name (Affiliation)”: ,Sean Upchurch ,Bimal Chaudhari , Salem Bajjali ,Reggan Thomas , Andy Yates ,Juliane Schneider , , , ,

	Actions Arising	Assigned To	Deadline
1	Come up with the proposal	Andy / Evan	2 weeks
2			

	Agenda Item	Person/Time
1.	Are VRS & VOD objects being included correctly?	Evan & Sean
2.	Model walk-through	Evan & Sean
3.		
4.		

Meeting recording :https://us02web.zoom.us/rec/play/pTFB8dniUwL6E7D4TDmP3EoYduqUyq2s35vKl6jnvNJa-5jbw4BHiBlHUYqCo1o1XnHW2lJcj4Ri3ysD.LQ8msTjAGxNj22YS?continueMode=true&_x_zm_rtaid=GGwjyPSNTLisgfW3lXgNqw.1662052555919.bc569e3f697dbd5a61e685a13171d411&_x_zm_rhtaid=947

Meeting minutes :

Transcript table - https://docs.google.com/spreadsheets/d/1VQe1rNrzsP1dwLNhtIDB_9DZ6AxJ7DNGQcqr9kMOzk4/edit#gid=0

Model : https://docs.google.com/document/d/17jsuJk_9WSDrowPPXkbOsVdjOXkLvUtFR_IVV3RCCV0/edit#heading=h.u2vxnfkq209rUser

Stories : https://docs.google.com/document/d/1yHhVK-8TdCHOTftBv3cqFg6o-0FkjPn0tG3nXOqwkb8/edit#heading=h.vo7w4rw4ya22

Definition and Model :

https://docs.google.com/document/d/17jsuJk_9WSDrowPPXkbOsVdjOXkLvUtFR_IVV3RCCV0/edit#heading=h.u2vxnfkq209r

Global Alliance for Genomics & Health

AY : https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/1101

2022-08-04: Transcript and Gene model

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	“Transcript” and “Gene” models	Group
2.	Evidence code - Evidence Ontology or a GA4GH managed list?	Group
3.	CDS as “the most relevant unit”	Group
4.		

Meeting recording:https://us02web.zoom.us/rec/share/E2PbzsIXDcCIT-EhhobVgUxqrPj0y5tpjzHOANj8vdJkn-Y0DReVvm4R4TV5bwf2.GntKkQvsDHEVYbDd?startTime=1659621581000
Passcode: 44g&W.RM

Meeting minutes:

2022-07-21: CANCELLED

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time

1.		
2.		
3.		
4.		

Meeting recording:

Meeting minutes:

2022-07-07: Transcript and gene models

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Discussion of September plenary meeting roadmap: ● “Transcript” and “Gene” models completed by September ○ Deliverables? ■ UML or other standard? ○ Checkpoints? ● “Regulatory Region” model completed by end of calendar year ● Potential driver projects at the beginning of next year	Group
2.	Update on “mRNA” attributes and metadata	Evan
3.		
4.		

Meeting recording:https://us02web.zoom.us/rec/share/pmbzVCxGSVMnzGQULULe5AUMKh35rKAH1Vj6TobLgJHhE673AsD7EG9TkmJ5Jcbp.IuB0haAyHjVR4rfI?startTime=1657202266000

Global Alliance for Genomics & Health

Passcode: #s65?CKq

Meeting minutes:

2022-06-16: Regulatory model V1

Chair:Sean Upchurch, Karen Eilbeck,Evan Christensen
Attendees “Name (Affiliation)”: ,Dmitriy Sonkin , Juliane Schneider ,Reggan Thomas , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Roadmap Document	
2.	Regulatory Features for V1 Model	
3.	Next Steps	
4.	Steering Committee on Jun 22, 2022, Deadline to complete SA slide is on Jun 21, 2022 -	Reggan / 5 mins

Meeting recording:
https://us02web.zoom.us/rec/share/zva5YRU15TCYiGIBxbQmNHefDOUb7Mnj89ouR1aCBFJ1m0L3a8DXVSSg5x6jEUae.g296DIjdjPLo0clo?startTime=1655387828000
Passcode: Fe5QjXa$

Meeting minutes:

https://lucid.app/lucidchart/aeaabf97-a0cd-482e-9297-6a2f18f80ab4/edit?invitationId=inv_a5041b50-7dc8-401e-9241-9013a5d27270#

EC - Provided the objective of the SA group to Juliane Schneider.

- Purpose is to create a sequence annotation model which is fully extensible and can be applied to different situations. Kind of trying to build a front end model.
- Not on formal notation still working on pseudo model.Still more focussed on attributes and relationships.

Global Alliance for Genomics & Health

2022-06-02: Sequence Ontology

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Working session: ● Paring down Sequence Ontology terms under “regulatory_region”	
2.		
3.		
4.		

Meeting recording:
https://us02web.zoom.us/rec/share/SxN9ylvgBc5eou1G1xXmGB1dh9_hNnR845s5LlSqjjPVjoMImVBpTCA8pV_vHkxN.-Mk6YmMuz9xNJdEh?startTime=1654178148000
Passcode: 3sm^Rzu&

Meeting minutes:
https://lucid.app/lucidspark/9c130692-fbf1-4d70-b31e-8714db810122/edit?viewport_loc=-2882%2C-525%2C4467%2C2192%2C0_0&invitationId=inv_5dd099a3-a679-4b07-b8cb-639c90ef3216#
KE - This is a paper that we worked on about transcription/regulation in bacteria
https://pubmed.ncbi.nlm.nih.gov/32665585/

KE - This is a paper that Dave did in my group on Gene regulation
https://www.sciencedirect.com/science/article/pii/S1874939921000638?via%3Dihub

2022-05-19: CANCELLED

Chair:Sean Upchurch, Karen Eilbec
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.		
2.		
3.		
4.		

Meeting recording:

Meeting minutes:

2022-05-05: Coding & Non coding transcript

Chair:Sean Upchurch, Karen Eilbec
Attendees “Name (Affiliation)”: , Wes Goar ,Evan Christensen , Dmitry Sonken ,Alex Tsai , ,Reggan Thomas ,Juliane Schneider , Andy Yates , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Review of objects and attributes	
2.	Coding & Non-coding transcripts ● Do we need to divide the hierarchy by coding & non-coding or can this just be an attribute?	
3.	Transcript inheriting attributes from Gene ● Are there transcripts not associated with a gene relevant to this model?	

	• Do we need to force transcript to inherit everything from gene? • Do we want to limit our model to transcripts associated with a gene?	
4.	Blank boxes • Do we need to keep them even if they don't have any attributes?	

Meeting recordings:
https://us02web.zoom.us/rec/play/wpy4bK09nzIHdJ204KmKH3N-Kdn4ALDWmnj5MS2Tw4m3dj7T98CJLDvfoH7xBUG3GFmR6Qv-PF9_P7iv.UekAZw2K3Yg3tUi7?autoplay=true&startTime=1651759100000

Meeting minutes:

Variant annotation and variant representation groups realized that they needed a higher level model, but they at least needed a sequence annotation model, so they needed a model for sequence features and Gene and transcript so that they unite their models.

all of these terms and relationships are already modeled in the sequence ontology

So we selected this group of terms and all of these are in the sequence ontology. And the challenge has been kind of how do we connect all these together? How do we also, I mean, what metadata do we need to attach to each of these terms? So that's kind of where we are at so today.

Do we want to divide the hierarchy here by coding and non coding, or is this something that can just be an attribute because so far, we haven't really attached any metadata specific to coding versus non coding.

From the standpoint of a data model, the phasing phase would exist as a coding, you know, feature and be inherited by everything that sits under it.

About a blank box because we are building something that is a framework to be extended. And if there was, you know, a need for for that level of granularity such that another group one that for example, might specialize in a certain type of you know, of RNA or type of transcript that we don't have properties for but that they might want to add and call out specifically, than having it as an entity in the in the model is a desirable thing.

Your genes and then transcript and regulatory regions are kind of the main containers for lack of a better term. We're still working on that but from transcript which is really where we spend most of our work right now, we decided to basically go with transcript types. And then with all the different transcript types here, we have mRNA, Link RNA or RNA, you know, so on.

We put together a list of a bunch of different attributes that we thought apply to all or some of these. And it turns out that really the only ones that all of them share are exon and poly a site of these terms that we chose to include in this initial model.

There's the sequence feature gene transcript coding, I mean, is there anything that we really need to be adding I guess we're gonna have phased coding, but is there anything that we would want to add? You know, say specifically to mRNA or anything like that.

Are there non coding transcripts? Not associated with a gene?

these pseudo genes, things they can make transcripts, whether they're actually useful or not, is up for debate.
There aren't any non coding transcripts that will not be associated with the gene for the purposes of Genomics and Health.

What do we want to do with coding and non coding? Because I don't like it just as like two separate entities below transcript. That doesn't feel quite right to me, but maybe not.
We can do the easy ones like promoters to enhance those kinds of things, but I think at some point, we're gonna have to go look at what other databases are capturing and what's what's in the literature.

you have your main for your promoters, enhancers, silencers and insulators.

regulatory regions are going to be part of the v1 model, right

the treatment of regulatory regions is that really going to have any bearing on the transcript types?

the next stage then is going to be beginning to model regulatory regions. so I guess as a kind of framework for moving forward.
Look at the sequence ontology, right, look at the major databases.

Is there any other kind of forum or repository I need to kind of look at? for regulatory reasons you're gonna want to look at what ENCODE has done, and ensemble

2022-04-07: CANCELED

Chair:
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

Global Alliance
for Genomics & Health

	Agenda Item	Person/Time
1.		
2.		
3.		
4.		

2022-03-17: Transcript types

Chair:Sean Upchurch, Karen Eilbec
Attendees “Name (Affiliation)”: Evan Christensen, Sean Upchurch, Tim Hefferon (NCBI),

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time

1.	• What Evidence/Validation code would we be interested in using?
	  ○ Does RefSeq record some sort of evidence code?
	  ○ Candidates:
	  ■ http://geneontology.org/docs/guide-go-evidence-codes/
	  ■ https://evidenceontology.org
	  ○ Ensembl coding for human data:
	  ■ APPRIS (looks for isoforms through mass-spec)
	  ■ TSL (a measurement of confidence based on EST evidence)
	  ○ Naming a computational method vs naming statistical method or algorithm in the model?
		Evan

2.	• Organization of transcript types	Evan
3.	• Application of MANE attribute in “transcript”	Evan
4.		

Global Alliance for Genomics & Health

Meeting recordings:
https://us02web.zoom.us/rec/share/JU3qpCdufcGfE0_jS70cinHRRatYf4X6KP5g7OqBI7jY1Masp_HvKfaIwYRN4Jc.3_Agv-K8EWfVkV64?startTime=1647525606000 (Passcode: 2f%+?ha9)

Minutes:
SU - Gene, transcript and regulatory region are pretty much are our top level entities in the model
EC - a feature is essentially any piece of knowledge that you want to attach to either a point in a sequence or stretch of sequence

feature is just found, you know, at one point, so like, say, between two different precedent, then that she we call that a junction and then if it's a stretch of sequence, and that's a region but in essence feature includes both of those terms, and can encompasses really anything that you would want to annotate a sequence with
We did have the question of location, we just decided to adopt VRS location model verbatim.

we haven't really looked at it beyond just kind of putting it here. But when it comes to location, we're just kind of going with what the larger group has decided to use.

we're looking at the usage of sequence annotation, especially as the research starts to transition over to clinical applications

so this next line here with the different transcript types, we tried to, we spent quite a bit of time assigning attributes or different, you know, other features, I guess, or properties, properties, sure, to these different transcript types. And we found that really the ones that we the different properties that we had highlighted that we thought would be important for this base model. Of those really, it was just link RNA and mRNA, that had stuff that was unique. So we're still really deciding what to do with these last columns here.

we really want that to change, you know, the the next steps of the model based off which transcript type you select, but that's kind of still an area of development

There is even an ontology for certain types of evidence. there's the geo evidence ontology, but also just like a more abstract evidence ontology as well
Questions - what sorts of things we need to say, you know, where did this sequence come from? Is that a relevant piece of information? Was this computationally derived? Was this observed in an experiment? How I mean, what is the provenance of this information?

we're thinking that might be a helpful piece of information to build into the base model

Tim - in terms of our variation databases, we do accept evidence and ClinVar certainly accepts evidence as it applies to making the clinical assertion between a genotype and a phenotype.

EC - we're looking at the usage of sequence annotation, especially as the research starts to transition over to clinical applications, it would be helpful to know, at what stage are we in kind of our analysis, especially as we start to talk about, alternative splicing and things like that
We were looking at existing ontologies on on the matter, and it seems like the Gene Ontology one captures a lot of them.

SU- if we want to restrict it to just a particular subset, then as Tim said, we make our own, you know, controlled vocabulary for it
this might be one of those those properties that we we leave in the initial model

Organization of transcript types :
we want to have these separate entities that affect say you select mRNA,they affect subsequent areas of the model

Application of MANE attribute
SU - MANE is a boolean. All transcripts are coding

2022-03-03: Evidence Representation

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: Dmitry Sonkin, Evan Christensen,Andy Yates, Emilior , Francis Ouellette (McGill University), Bob Freimuth (Mayo) , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Paper Tables	Karen
2.	Evidence Representation ● GO Evidence Codes? http://geneontology.org/docs/guide-go-evidence-codes/ ● Is this out of scope?	Group
3.		

Agenda Items for next time:
- What Evidence/Validation code would we be interested in using?
  - Does RefSeq record some sort of evidence code?
  - Candidates:

        - http://geneontology.org/docs/guide-go-evidence-codes/
        - https://evidenceontology.org
    - Ensembl coding for human data:
        - APPRIS (looks for isoforms through mass-spec)
        - TSL (a measurement of confidence based on EST evidence)
    - Naming a computational method vs naming statistical method or algorithm in the model?
- Organization of transcript types
- Application of MANE attribute in “transcript”

Meeting recordings - https://us02web.zoom.us/rec/share/UtMqPjV3Ygm1DBSIzuFhDukOLbZljAYuqw9O-wpPo0ZccXNTpmwLjLhSWtDric9-.B4-B8TcCHY5MwXQo?startTime=1646319469000 (Passcode: xu%UN1qe)

2022-02-17: Feature,Transcript,MANE and Polymerases

Chair:Sean Upchurch, Karen Eilbeck
Attendees “Name (Affiliation)”: Dmitriy Sonkin, Reagan Thomas, Andy Yates, Bob Freimuth , Evan Christensen , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Working session: • Feature ○ VERS/Refget Sequence Location? • Transcript Type Enumeration? • MANE (Boolean or Int?) • Polymerases (enum?)	
2.		
3.		
4.		

Global Alliance for Genomics & Health

Meeting recording - https://us02web.zoom.us/rec/share/wLMqacrGjuGurxrLSY4nZCyt6_7MjTrNupBJl9wM9_B5vJLLfpOo5s33Wlwcmiho.ChD7QCoFlstPFDJO?startTime=1645109903000 (Passcode: zAz@xw*6)

2022-02-03: UML diagram

Chair: Sean Upchurch, Karen Eilbeck

Attendees: Evan Christensen ,Andy Yates, Dmitriy Sonkin , Emilior , Reggan Thomas

	Actions Arising	Assigned To	Deadline

	Agenda Item	Person/Time
1	Welcome & Confirm Agenda	
2	Working session to rework our UML diagram incorporating feedback we've received about: • Genomic location • Build • Transcript type	Evan
3		
4		
5	A.O.B	

A generalized model to describe sequence feature annotation

Links from the chat :

Sequence Annotation GKS requirements/status -

https://docs.google.com/presentation/d/1XLzsNH_gSKUP7S2z91ezenEH827XTqM4JWMv6c_ZvMc/edit#slide=id.p

GA4GH VRSATILE ->Value Object Descriptor Specification - https://vrsatile.readthedocs.io/en/latest/value_object_descriptor/index.html#value-object-descriptor

Schema - https://vrs.ga4gh.org/en/stable/schema.html

Meeting recordings : https://us02web.zoom.us/rec/share/B4ojr5XBPlFelJmbhXyV_LEzFOkhpSQtKOZggEkRSjGfSL1Hv5Mr9U3K1EVyMtse.wY5P0S-RvDcx4SCY?startTime=1643900264000 (Passcode: 8u!033$y)

2022-01-20: Location model for VA and VR

Chair: Sean Upchurch, Karen Eilbeck
Attendees: Tristan Nelson (Geisinger/ClinGen),Andy Yates, Evan Christensen,Dmitriy sonkin,Reggan Thomas

	Actions Arising	Assigned To	Deadline

	Agenda Item	Person/Time
1	Welcome & Confirm Agenda	
2	Tying up loose ends and plans for 2022	Sean/Evan (15 Mins)
3	Andy - discuss his findings about the location model for VA and VR	Andy (20 Mins)
4	Possible AMIA paper	15 Mins
5		

Global Alliance
for Genomics & Health

Meeting link :
https://us02web.zoom.us/rec/share/KFCS59AZZy27rpEF9fs15OpZB2i7U0fUg5d34-tXg9N9oefQfs5j4Y-nzMUlrbLs.T0UKc6Em91YJLCpq?startTime=1642690548000

Notes:

KE
I just got a text from Evan saying that he's going to be about five minutes late to the meet him Hey, Shawn, do you want to start without Evan? Yeah, I

SU
think why don't we go on ahead? I mean, there there are some of the wrapping up items that you know that he's already familiar with. Why don't want to go ahead?

KE
Yeah. So, um, I guess Andy didn't show up in there. Um, so I can

SU
there, he's now been summoned. Oh,

KE
oh, that's nice. I nd Good, good job. I wish I could do that every time just say someone's name and they would appear. Okay, so we've got like a

AY
change. I'm just wondering, who else would you choose?

KE
Who would I choose to come to this meet and

AY
I seen if you can call anyone forth from the

KE
let me let me go back to the agenda. So um, we were going to use today as the day to kind of tie up all of the loose ends and make a plan for moving forward. And that sort of ties into my idea for putting together an Emir paper that describes the work that we've done so far, and outlines a pretty simple model. So I've started to write that. And I'm hoping that that's sort of going to be the product for Evan as he finishes up his master's. So I did want to sort of talk about that today, get some feedback, and maybe even hand out some tasks to other people. So for those of you who aren't familiar, Aimia is the medical informatics Association, American medical informatics Association, I think it stands for, and they have this meeting in November every year. And it's like the place that informatics people go and schmooze and present their work. So it seemed like a pretty good place for us to advertise this work, especially since like the HL seven people all go there as well. So if anyone has ideas for a different forum for writing up this work, I mean, I'd be happy to take any advice. So if you have any ideas, Speaker Hi, everyone, I just see you popped in.

EC
Hey, sorry about that. I need to leave earlier from the house these days, traps picking up again.

KE
I'm really in the middle of our search, there's more traffic than normal. That's great.

EC
Well, it might have just been this morning. I don't know if it was just one driver going real slow, but it felt like there are just more cars on the road now.

KE
It just never leave my house. So yeah, I was just floating the idea of the Emir paper. So the deadline for that is it's usually around March 12. And I haven't looked for specifics, but it's around about the second week of March that we would have to get there. And I like that idea because it's a very hard deadline that we sort of have to work to. So Shawn, and Evan, any ideas about that? You think it's feasible? That's

SU
a I think that's an excellent idea. It's interesting that the deadline is in March.

KE
Yeah. And the conference isn't until November.

SU
Yeah. Regan, is there any work done on whether or not we're going to be having the march Connect meeting?

Unknown Speaker
Oh, should I don't have any information on that? Maybe I can check with the Secretary and get back to you.

SU
Okay, cuz yeah, that sounds like that's coming in at the time that this paper deadline is And it might, it might work out for us if there's any, you know, polishing that last minute polishing we need or anything like that. But it also means that whatever plan we come up with today, in terms of getting, you know, a model or the process together, aiming towards March, for looks like both of the paper and potentially the Connect meeting, March, it looks like our date.

EC
Great, actually.

KE
So I have like a really rough skeleton draft, I'm gonna see if I can share that. And, I mean, it's, I hate to use the word, but it's kind of brain farts right now. But I'm hoping that we can all work together and make it a real paper. Okay, so can you guys see that? Yes. Okay. So, like, I'm not married to any of this, like the title like anything, right? We'll put this in a Google Doc. And we'll all work on it, I hope, right? But the idea is just like this generalized model to describe sequence, feature annotation, I put the word feature in, I think that sort of helps, and, you know, with the scoping of what we're doing. So what I want to do, and I think Shawn used to help me with this, is really to articulate the gap, like, why was this group created? Why? Why are we working on this? What's the gap? And then talk a little bit about, you know, the role of GFR GH in biomedical informatics standards. The methods section is basically like, what we've done, right? So we did use case development with stakeholders, we've had these iterative group meetings, and we do every year, and we do this wider community feedback. And so the results are just gonna be like, what? Basically what, like what we've discussed in these meetings, right? So we have the user stories, we've padded out use cases. And then we've looked at all of these, like different parts of the problem, right, so we've looked at like the terminologies that exist, we've looked at the currency, a little bit of data exchange of sequence annotations. And then we're looking out and we've had people come in who described like, what they do and what they need is annotation providers. So this is where it gets a bit fuzzy. But I think we also have to sort of articulate where we are within GFR GH itself and talk about the other groups that are our neighbors, so the variant annotation variant representation group. And I'm hoping that Andy can talk more about this today. And then our discussion is just going to be the things that we like, finalized in these meetings, right, like our containers are going to be genes, transcripts, regulatory regions, we've talked a lot about features for the first version, which ones are important. And, you know, the relationships between them. And that I mean, this is all all i All I've got two so far. Let me stop sharing any thoughts? from anyone?

DS
So this is mitre. Um, should we concentrate on RMA related things? Because, of course, we have conversations about vertical regions. But I thought we decided we're going to limit them on Yeah,

KE

no, absolutely. Like, we're gonna talk about that in the scoping that these like three containers are really important. And then we're going to focus on the transcript. So yeah, like, that's part of the scoping discussion. So thanks for bringing that up.

AY
Yes, a very good start. Yeah, I completely agree as well. The the bed that comes to mind is there's going to be a tightrope for us to walk here between presenting what we have as being latest developments around this rather than a kind of like a pointer towards what may become a standard from Global Alliance versus proposing this is actually a standard just because it will, it will be impossible for it to go through a production approval process by March. Because the timelines definitely won't match up there. So,

KE
right now, so we have to be a little vague about that. Right. But we've got another six months before it gets accepted, which, you know, give it 5050. Right, like Robert Freeman's had one of these papers accepted. He presented it last November. So some GFR gauge work gets through that.

AY
Yeah, it is more of a case of making sure that wherever the stone, wherever the standard is, in terms of Global Alliance is processed. The paper doesn't kind of misplace it, if that makes sense.

KE
Right. Right.

AY
Yeah. So it's just, it's just that we want to be pushy here. But we also want to make sure that we're not so pushy, that basically we make it sound like oh, well, this is not only approved, but it's rubber stamped by global lights, and it's an official standard or anything like that. So yeah, it's beyond that concern. I think. I think the general structure of what you've got is really, really good. And actually concentrating a lot on the process, I think is good, because let's face it, that's where the majority of the time is being spent.

KE
Right. Right. And it's not it's not a normal science paper. Right? It's, I mean, we, we basically like, right now, it's more like, I don't know, like a psychology paper or something. Right. It's like people talking in consensus building. So it's different from, you know, a discovery.

SU
It's, it's kind of more of a materials and methods. Right. Right. And I think that's good. I think Andy made a very excellent point, I think we need to make sure that as far as the sequence annotation standard is concerned, we need to make sure that we're very clear. Right. We don't get any further than a save the date for some future actual release.

KE

Yeah, yeah. I mean, I think we have like in the discussion, it's like next steps. And in the next step section, we'll talk about the actual process going forward. Next step, actually do this. Right, like, right, next step, right? This isn't actually like a standard. Yeah. And we have to do blah, blah, blah, to make it official.

Evan Christensen
Oh, sounds good to me. I mean, so I guess the next step for me is just to continue pulling citations and top to draft that up.

KE
Yeah, I mean, I think we need to put it on Google Doc. And I didn't want to put it in my personal space. Because that is, I know, that's a bit weird. So hopefully, like everyone, Reagan today can, we can send the draft and throw it up on Google and start assigning tasks, right? Like, there's a couple of people I want, I want them to fill in the table with their things. Like I want, I want him half around to contribute some of his immense knowledge to this. I think this stuff Dimitri can add, I want to talk to Tristan about, you know, the clinical side for what features are needed. So if we have it as a Google Doc, we can start like pulling people and

TN
Yeah, happy to offer whatever I can and to loop in whoever you think could be helpful in that respect. So

KE
yeah, one, one of the comments that just stayed with me from the last connect me and we had, or maybe it was even the one before that there was a lady from Australia, and she's just like, I need everything. ICMJE. Right. So I think that's like an important thing to keep in mind as we make this model. And it's something that I think can go in this paper, but I just want to be careful about that. So I think Tristan can help.

Unknown Speaker
The comment was I need everything from ACM G.

KE
Yeah, like every feature and every piece of metadata metadata that could be used, you know, in the SEMG classification of variants.

EC
They got it.

SU
Okay, so reggan is there a space in which you can carve out the GA for GH Google space for this?

Unknown Speaker
Yes, Shawn, so we have a dedicated space for G fo GH GKS drive Google Drive, I can actually post that into Google Drive.

EC
Great. I'm gonna afford her unless Karen, the file that you shared with me is the most up to date version?

KE
I think I probably tinkered with it. I'll send it to Regan.

Evan Christensen
Okay. Great. So the next item in our agenda. Let's see, what is it? So tying up loose ends plans for 2022. I guess we kind of started talking about that.

KE
We're hoping. Can we do Andy first before we tie up loose ends? Sure. Yeah. Is that okay, Andy? Yeah, close, it is. Sorry, just let me get up

AY
the presentation, as I did go back to the so I did paste this up into the Slack channel as well. So the slides are linked out from there, so don't worry about that. So the idea of this was, I'll start with a video that didn't realize any of my videos, and give them some light as well big skinned up. So yeah, I think it was last week, we had a little powwow about re-evaluating where we are. One of the points came out of where the other GKS groups are with respect to the requirements and status about SQL sanitation. And this I, I would agree as well, sequence feature annotation is important concepts here. But actually, the concept seems to cover far more the conceptual entities that may be located again, sequences of sanitation, and also particular instances of that. This is basically a summary of a set of likes, tennis or emails or went flying between different people on that. But just to move on. Yeah, this is basically repeating myself here, the thing that I do want to point out is that the groups have already gone and created some of the data structures that they needed, because there was a factory gap. So when they were trying to annotate the subject of various states, based on a piece of sequence sanitation, they didn't have a data structure that could represent that. So in order for them to comply with the bill to do that, they had to create it. So basically, I'm going to go through at least for VR, what's existing at the moment, there's a little bit on va as well. So these are some of these parts of specs already. And standards, others on our part of specs, but they've not yet been approved, or anything like that, and others are clear indicators of the direction travel of worship groups is heading. So the first one is verse, They have two concepts at the moment, Gene and sideband. So actually, if you go into the presentation, anything that's a link is basically there to give you a jump into where the spec is. So basically, that tricky strain to the jean spec that will put you to the site plan spec. I think really that even though the site plan stuff is kind of relevant, it's not completely the area of interest that we're that we've scoped ourselves down to. So really, the only thing of interest is jeans here. And really, there's the definition that they have from their own spec, that they have explicitly linked to the CIO term Prejean. But really, it's a namespace query identifier. That's the only thing they consider a gene to be because versus intended as encrypted be the bare minimum information required to transmit a piece of information. So if you're trying to transmit the idea that 613 is deleted from NCBI gene, well, you don't need anything else other than to

say it says the big 6135 over the wire. And then you would end up with being able to represent that using either lookups or the semantic reasoning on your side and then understand a little bit more what's going on there. They actually recommend to use Heisha CIDs as the main one because effectively just allows you to normalize between multiple providers, which is kind of sensible, really because if you do a one that's dealing in ensembles or stealing in SEO gene, you're not going to know you're going to basically have first checksums come out to different values because they fundamentally representing different pieces of data.

KE
Hey, can I ask a quick question? So there's the Shansi IID when it refers to the same gene in the two different sort of provider systems, is the sequence always identical?

AY
No, it's just the conceptual unit of

EC
heritability.

KE
So, so I couldn't, I couldn't give it like the HGN Cid and then say there's a deletion, like 1000 bases in and it'd be the same and both.

AY
So you could formulate a, it is possible to formulate a statement like that for sure. We have a deletion in this gene, there is a deletion of 1000 base pairs into the location of that. But that will depend on what the variation you're trying to model is. By which I mean that if it is a data storage, that is what the, that's what the variant you have recorded is, is correct to represent it as search, if you suddenly want to convert that into genomic coordinates, understand exactly where the start is. That's where it comes up to interpretation on the client side in order to basically deconvolute that into something that makes sense. So if you're using chaincode, somebody else, or genkos, I'm using ref seeker, NCBI gene that's basically on them to understand that they're going to come out with different answers. But fundamentally, the statement can still be true. But you wouldn't, you wouldn't understand, basically, a b, a b, side lassi, but I don't think there's a way around that at all unless you start to transmit more information. If you start to transmit more information, you may not actually choose to represent the variation as a gene based variation, you might actually switch to coordinate based variation because you actually have the genome and then you could transmit that then a bit usefully over the wire.

KE
Right? I mean, I, I guess I'm, I don't know, I'm a little confused. I can see mostly,

AY
mostly these in use for like copy number statements.

KE
Right now. I was about to say that, right? Like, I can imagine like an amplification or like a whole gene deletion, you would just talk about the gene because the actual stuff that's inside that container doesn't really matter. If you're talking about a one base pair, change in that gene, it's lossy, or it can be okay.

AY
Yeah, I it's it's a versus responsibility is to attempt to represent both the data as it stands at rest in the format. But basically, if someone wanted to impute additional semantics over a statement, like you saying there, that'd be, that'd be on the provider to do. But ultimately, if you didn't want to impute those additional items on that and represent a piece of variation as it really is, then it has been recorded perhaps in literature, then you may not do that. But yeah, mostly this is copy number variation stuff, this gets used for service sites back stuff as well, that's also basically copy number. For the most part.

KE
Right. Okay. And then so then with, sorry, to go back to this, it's like, I worked with Reese a lot at the beginning of this, and then I had to drop off because of other stuff. But with the very representation group, right, you can also sort of represent the entire chromosome, and then use that to give actual positions, right.

AY
Yep. That's yeah. So so you can create other statements that say, we have a single base pair change

KE
in this version of the chromosome? Yeah.

AY
Yeah. Yeah. So you can still make those statements very, very, very easily. You don't have to make statements around. You don't have to use genes as part of creating a verse statement. That's you don't you don't have to include it as part of the variation payload. Okay.

DS
So this is the reason why because the definition of a gene is very fuzzy.

AY
Yeah, I believe so. Okay. And, and you can tell that Alex is somewhere along the way, as well, because it also starts to include regulatory elements. I believe Alex always feels like it does include regulatory elements as well. But yeah, it's fuzzy because by its very definition, the concept that they're representing here is also quite a fuzzy concept. So moving on. The other thing that has come out from this framework called versatile, versatile is there is a way of collecting additional pieces of data that basically is not relevant towards the well, as it says though superfluous and salient elements, but necessary for specifying a value object. So effectively saying, verse alone is good, if you just want to get down to the bare minimum of what you're talking about, but actually, you may want to provide other additional pieces of data over the wire. And that's why you need to use this sort of framework. Versatile is not a

Global Light standard. It's developed by Alex and Larry, just to say where it is. And searches it is mentioned a lot with first but it isn't officially a standard. But I think that might change depending on some of the other things that came out from variation annotation. So in this seems a bit weird, why would you actually do this? Is because versus how allows you to bundle up other pieces of information as well. So that might be alternative representations of the same variation or alternative, slightly VCF record, or you might want to provide a high HGVs representation. And so just allows you to bundle more of those things together. But actually, what we found out by asking the same by basically asking VA, what are they doing around sigma sanitation, he came out that actually VA have gone and adopted versatile as one of the ways of providing, again, as this says, not superfluous things, superfluous to the salient elements, but necessary for specifying a value object.

KE
So I guess I don't know what a value object is. Value objects.

AY
Yeah, sorry. So I think in how Larry puts this is, first, as in the first packet is transfer elements as in interest in you, I'm going to Larry here in a second. If I get this wrong, you're very good. So data transfer being that you are just using that as an encapsulation for firing salient information over the wire, whereas value objects are actually associating that additional piece of data that is not necessary to the transmission of the major packet of information? Is that right, Tristan?

TN
Um, I think it may be the inverse. Yeah, the value object is kind of the minimally defined entity that you're describing, and usually described with just the fields that are essential to the definition of it. So variants would be location, sequence, change, transcript, or sequence, etc. And then versatile, kind of gives the ability to annotate one of these things with additional metadata.

Dmitriy sonkin
I'll be honest, I think, again, it's a little bit of philosophy production. More I agree with the way how nd defined it. Because in my mind, what happens is, let's say viewers, deals with pure arithmetic, definition of change. And very object allows to couple information which in practice are critical for example AGVs change, of link to location on with different genome builds. So this is my personal understanding sort of goes more along with how Andy defined it. But I think what Tristan mentioned it's a little bit tricky because there are still arguments with invariance you have people exactly how do you find itself maybe respiration why people sometimes try to reverse it depends on how exactly we're going to put on the model.

AY
I think ultimately for me it is so scaring. I think that the easiest way of saying this is that there's confusion after hearing interested as well, I'm confused whether I've said the right thing or not here but ultimately it's it's it's not. It's the stuff that surrounds by first house. everything that's around is something right. So what we have here is the NC diagram that is currently in the VA draft spec at the moment link at the top and also link at the bottom as well just jump you in there. But what you can see is that there are

statements so that the language here is pretty important in terms of what comes after things, but where you have a molecular consequence statement, that is what VA will consider to be the statement of what has happened. And then the molecular consequence result being the results of what's actually been as happened in that. And you can see as well that that references multiple different descriptors at the top and to my understanding is where you see the word descriptor appearing, that is basically saying, Okay, we are actually dealing now with a versatile container, an extension of that versatile container, I found it really interesting to to have both transcript descriptor and protein descriptor at the top there very clearly. So basically, the way that VA is approaching how to handle concepts such as transcripts and proteins, is they have gone wholesale into how the versatile framework works. And I've created extensions to that which would represent the various different elements that they feel is required, in order to make a really functional lethal consequence. statement. I think you can see this as well from when you go back to this, the basic description here, we have ID type level description extra, alternatively, alternate labels and extensions, you have ID type level descriptions, extensions, it's basically it's that the exact same things are there. So I think that this is, in fact, 100%. Sure, these, these are the same things that we're talking about here. And this is how VA has approached that statement. Now, the other bit I'd point out as well is that there are elements here like Gene ID, and gene symbol, but biotypes only appear as well in the middle for a transcript. So already, there are elements that we have picked out in some of the work, some of the stuff that we've done, that we've deemed as important. And actually you saw that importance reflected over here in the models that VA has put together, there was something else that I did want to point out as well, which I am a little bit disappointed by. I can't remember if this is true or not. But if you look at gene ID here, that effectively this is versus container for Gene ID. And basically, semantics and the usage of the not the semantics, the data model that's been provided that if that is just a string, then that's effectively different from how this actually represents this, but from looking through some of the documentation and comments from Larry, in the document this link that we will see, sorry, I can see this charts. Okay, sorry, currently, fingers crossed, that actually holds out. Your internet connection holds out. I believe that the idea is that that actually expands to the verse gene container. So then that would be equality to those things would be the same as that right, Justin?

TN
I think you're right to call this out and hope that the VA does. God hears a mistake, and that should expand to a VR gene concept. I think that's the right way to do it. Like you're saying,

AY
I yeah, I think this reflects how Matt and Javi and the VA group approached it originally. And it seems that there's a tendency to do this. And so but I think this is the challenge that you have with groups, when the groups are working on related entities, and there is no kind of underlying glue between them all that effectively. You get that kind of diversions happening. Anyway. Oh, just really quickly,

SU
Andy, that statement that you just made, is in fact, pretty much the essay mission statement. It's specifically really for cases like this, because we have these potentially two disconnects and how genes are defined by these two groups. And that that's, that is that is our reason to exist.

AY
Yeah, absolutely. And that's why I kind of pointed out that it's not the case that its challenges happened.

DS
So so one thing I want to also to briefly mention, because I was involved in what can this Can I was trying to find out in this representation where information is so actually the key so think about this molecular consequences, right? So let's say to defy it, let's say simple nucleotide change in DNA level. Now we're trying to figure out what does it mean? And really, the key information is molecular consequence result is basically provides information what is what exactly is the consequences, right, the change to the potential change the transcript in what's pattern, potential change to the protein level. So wherever because a lot of conversation and that is my perspective and corporate perspective, majority people This is really what restatement is about. Now what is trying to say. So, but sort of various dimensions, because they essentially provides you a way to mathematically show what changed and make it sort of consistent. But it's not sufficient by itself, you need to provide additional information to basically say, what are we trying to say in this statement?

TN
Yeah, I wholeheartedly agree with what Dimitri just said, I hope that helps to illustrate, I think, what was intended by the concept of a value object. And that that's really intended to be kind of the immutable mathematical description of a an entity. And any, any annotation, any, anything you're trying to say about that thing, even to the extent of applying a human readable Label Label to it kind of needs to sit outside, because that's not really part of the objects definition, per se.

DS
Yeah, right. So but it's very tricky, because one thing is define what is the change? But another keep in mind with title is its consequences. Right? So it's this two concept, one where it happened? And second, what does it mean? And this particular statement is mostly about what it's mean. But you still have costs to provide to your steps in what changed. So that's the tricky part where sometimes why well, meaning people flip it around, and that still sort of in discussion. And as I'm dimension, it's not really kind of been approved service is still fluid.

AY
Still fluid? I think it's fair enough. I'm Tristan, I can see you come in next, just because I want to go into the next slides. So go for it.

Unknown Speaker
Okay. Dimitri, is correct. I think the versatile concept is intended to help fill that gap. But it seems like like we're observing the VA, and you know, secret sanitation, there are a lot different overlapping groups trying to create solutions in that part of the space, even if I think the VR part is, I think, relatively well accepted and defined at this point. So happy to hear what you guys have to say and get on to the next slide in that case, cool.

AY
So the next few slides are just some comments that I've extracted out from the emails as well that that came from that and Javi was just basically try to at least eight some of the things that they are saying, and a lot of this is actually just moving over back through what has been said before, but I think actually, it's probably the middle bit that it's important here as well, that, that basically, if you get to a entity, or a piece of evidence, or data or provenance that warrants basically is fundamental to knowledge statements, VA would include that as part of the VA Scipio framework, you would create a statement object. So again, you would create this thing on the left hand side, and you'd have it as a field directly in that, which suggests that you can have fields basically borrowed from other objects, perhaps or you would have other objects integrated into here as and when you want to make those statements. And there's also basically a similar thing here about what would happen if the molecular consequence statement wants to understand that something hit three prime UTR, a frameshift, during translation.

KE
So don't thinking about like the VA doing that, right? Like the molecular consequences. I looked at their model, and they come from so breaks, we develop those with the API. Yeah, right. But do they have like three prime UTR as a feature? Or they just use the consequence? Because if they're if they have features already, and they're talking about three prime UTR, then they've modelled sequence features?

DS
Oh, I don't think of molecular consequence model, any type of feature.

KE
Okay. All right, it's good.

AY
I think I could create that statement as well, where the so if I was to run back over a VCF file, I could definitely call things out as being three prime UTR. As a consequence, link it to a transcript, but then leave any other kind of information out of that statement, because by having the transcripts recording the transcript ID and version and then also recording the consequence type. You could infer that back so long as you could get back to the original annotation.

DS
Right? Right. So we just basically Apart from molecular consequence, I think was the same thing pretended just mentioned. If it hits three prime UTR, if you look at the molecular consequence result, it essentially if it's three parts, which is part of official there, thank you, if free prime which happens to be part of a transcript, then yes, in the transcript, you potentially can have AGVs, which basically will tell you what happened. But of course, there will be no associated where protein change. But I guess the key thing is, the molecular consequence by itself does not really model anything, it basically just say, okay, is very sort of AGVs representation of this change. And if it is, it will, one way or another, we will go to the transcript allele, so to speak. So, if that's my least understanding, doesn't really explicitly modelling.

KE

Right, I think that's good. Cuz I didn't want to be stepping on their toes as well, if they did model the features as well. But it seems like there's a gap. Yeah,

AY
I think there's a gap. And if someone if you wanted to actually fire that over the wire to say, as well, oh, it was in this fight. It was in this transcript at this position so far. So the nice thing is our bounding would also be another descriptor that would surround us maybe very interesting.

TN
No, I just wanted to point to Karen, It's spot on.

EC
So sorry for interrupting, I would like to make sure that Andy is able to complete his presentation. And we do need to discuss a roadmap before we finish this meeting. Thanks.

AY
So I'm going to skip to the last one here. Because this talk that I don't think is that is useful, please read it offline. Is this bottom bit here have Matt say we will suggest that essay, use the versatile framework to define the value object representation, a sequence feature and sees it as transcripts exons. And also define specialised descriptors for each of those entity types that hold on to the types of additional non essential metadata relevant for each, such as the gene and transcript ones that were previously that would be analogous to the variation descriptor. So this is actually is I'm going to just stop, share, stop that go back to my thing. So basically, the the outcome of this that became clear at the end is that VA have already gone and put a lot of effort and a lot of resources behind the idea of using this tool and using these descriptors. So one output of the of the work from sa will either in order to fulfil the requirements of what VA now has, is either to recommend what those descriptors should look like. In addition to another model, or actually, in some cases, provide what those descriptors are. And that is effectively some of the that is sort of the model. Now in that descriptors, there is no fields yet for describing what a location is or how you refer to a sequence. Well, you might want to reuse verse locations. There's still a huge amount of freedom within that statement. But it appears clear that there's a desire to basically carry on using those VA those versatile descriptors. And if we do not provide those, then it seems like VA probably would not use what use what comes out for this group. I think that that's that's my take home on that.

SU
Yeah, a plus one on that. I think from from looking at the the definition that you presented, you know, it's very high level, and it's a lot of stuff that we're going to want to put in any kind of an object in the way so I think Evan make sure you take a look at that when when working on the the draft of the bottle, and make sure there's no last stumbling blocks there.

EC
But uh, yeah, I agree. Sounds good. Well, we have five minutes left. before we'd like to end the meeting today. We do want to talk about what kind of thing we want this to quote Sean v one A alpha male.

Andy, you're fine. No, this is I really enjoyed this video presentation. Where are my camera? We are thinking, you know, what do we want this model to look like? Are we looking for like just the transcript or We're looking for gene and transcript or we'd like so we're thinking we have kind of this containerized analogy that we're working through, we also have kind of a almost like a filmstrip analogy that we're using. But where do we want to cut things for this kind of initial stage?

AY
So the one thing, sorry, Karen, no Yuka. So I was gonna say the one thing I didn't actually get out of mats and heavy. And the verse guys are where their roadmaps are taking them at the moment. So I, before describing potential other features or entities, we would want to attack maybe also, I should go back to them and just say, what are the next things on the radar, you know, Gene and transcripts has been pointed out there is exon on the radar as well, because then that would, that would, perhaps, that might give more of an indication as well were to actually sliced that,

EC
Perhaps it sounds good. So maybe we can make that our first agenda item for next week. In the meantime, I'll continue gathering citations and doing what I can feel free to hop into the working Doc, I did make some edits to the working doc. So things are formatted slightly differently than what we have seen in the past, I did it because we were trying to do everything in a spreadsheet. And it just wasn't really working for me. But essentially, we have the definitions that we're working on right terms and definitions, I put a the existing definition, the g4 Gh definition we've developed, and then just any discussion on the topic, and we would like to have some relevant citations here as well. If you happen to have something and you want to throw it in, please feel free to throw it in there. So that's kind of where we're at. And this is where I'm focusing most of my effort right now. So maybe once we get back together and our next meeting, we can discuss a little bit more how the sequence segmentation model will fit in with other models. And then that'll give us a better idea where our stopping point will be.

DS
To me, I think for this paperwork, Karen was describing the table on your forks, the wtitten pro gramme, you're looking different types of, in describing different properties. To me personally with Vicki Fink to show how we're thinking about this, and how we're trying to deal with this.

EC
Well, we'll definitely need a table and will need some sort of figure for our paper. But for now, for developing these individual terms. I think this is what right is

DS
going to face in my mind this table with you folks, we're putting together when you least mRNA tRNA. Other types of RNAs to put together I think that's the key thing with this vet would basically convey what are we trying to model right? To collect information to model it?

Evan Christensen
Yeah, that's in the working doc. It's on page 48.

KE
And so we were also, like, offline from this, we're talking with the RNA central people to make sure that our terminology for the pieces of those different types of transcripts is the same as what they use. Right. Right. And so I mean, I, it's one of the things that has been bothering me, like to the non coding people use the word exon, and it looks like that's what they're doing, even if it's not technically an intron that separates the functional pieces. But we meet a guy Steve Meyer whose goal is helping us with the definition ends.

DS
Yeah, no, I did. I personally also don't really like the way the storms use it, but that's how we use it. So we don't have.

EC
Right. Alright, folks, well, in case anyone has a meeting at the top of the hour, we'll give everyone five minutes back. So thank you very much. And we will see you in two weeks. Well, no, because in two weeks. I don't have a calendar in front of me. Whenever our next meeting is it looks like Reagan sent out the invites already. All right. Thank you for Bye, everyone.

Global Alliance
for Genomics & Health

2022-XX-XX: Template

Chair:
Attendees “Name (Affiliation)”: , , , , , , , , , ,

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.		
2.		
3.		
4.		