Global Alliance
for Genomics & Health

GA4GH Structural Variation - Special Interest Group

Minutes and Actions 2023-2024

Others:

These are the minutes for the SV-SIG meetings, a subgroup of the GA4GH Genomic Knowledge Standards Work Stream. For further information, please visit the GA4GH Work Stream page (will link to the page on the website when it is ready) at https://ga4gh.org.

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.
- GitHub Project board is here.

Supplementary Information

- VCF v4.4 (PDF)
- CNV annotation standards @ ELIXIR hCNV (cnvar.org)
- Beacon "non-SNV" query documentation

Table of Contents

Meeting Protocols
Supplementary Information
Table of Contents
- 2023-02-01: Final meeting
- 2023-12-07
- 2023-11-02
- 2023-10-05
- 2023-09-07
- 2023-08-03
- 2023-07-06
- 2023-06-01
- 2023-##-##: TEMPLATE

Global Alliance
for Genomics & Health

2023-02-01: Final meeting

Hosting: Daniel Cameron, Mrinal Thomas
Attendees "Name (Affiliation)": Beatrice Amos, Larry Babb, Tristan Nelson, Alex Wagner, Javi Lopez, Greg Sharpe, Jeremy Arbesfeld Jie Liu, Kori Kuzma, Mrinal Thomas, Pille Hallast, Tim Hefferton

	Agenda Item	Speaker/Leader	Time
	Welcome, reminders	BA	2 mins
	VRS 2.0 integration, roadmap for event annotation. Merging of SV SIG with standard VRS meeting.	DC	15m
	AOB		

Meeting recording:
https://us02web.zoom.us/rec/share/K00nMg9hOaqFcqn3kAgVzLcQfSTDb_nZcawXArhG7_8rYqmjQy6B1xMEWLTWIeQI.R4insONzzHSBiw0G

Please note that this is the final SV-SIG meeting. Further structural variation discussions will happen during VRS meetings. Please email: beatrice.amos@ga4gh.org to be added to the VRS mailing list.

Minutes:
Connect Meeting Session proposal form due in two weeks. Potential sessions discussed. Please review and put your thoughts/comments in excel planning sheet shared. Need for a session involving SV SIG, VCF, and VA to address the classification of event models.

DC discussed the absence of the actual event model and classification - need a joint discussion between VA and VRS groups to determine the appropriate placement of certain elements.

Discussion about complexities of structural variants and the challenges of representing them accurately: haplotype models, ordering, and the difficulties in capturing real-world complexities in existing models.
DC - important to gather community feedback; perhaps we need dedicated sessions on specific topics, such as implementer experiences, developer tutorials, and discussions around SV data classes.

Larry - how do we maintain momentum? Could consider combining SV discussions within regular VRS meetings? Maybe designate specific meetings for SV-related topics. Daniel says yes, but poss not at 3am!

MB discussion on query models, clinical applications, and the challenges of standardizing classifications across different tools and models. DC/MB agree need for collaborative efforts to bridge the gap between representation standards and practical applications in the clinical setting.

2023-12-07

Hosting: Daniel Cameron, Mrinal Thomas (Epic)

Attendees "Name (Affiliation)": Beatrice Amos, Larry Babb, Tristan Nelson, Wes Goar, Pille Hallast, Greg Sharpe, Vivek Nuthalapati, Michael Baudis, Joanna, Jie Liu, Kevin Riehle, Jeremy Arbesfeld, Mihai Todor, Rodrigo Martin, Daniel Puthawala, Fotios Tsetsos

Please review and provide feedback: GitHub pull request.

	Agenda Item	Speaker/Leader
	Welcome, Housekeeping	Beatrice
	SV-VRS 2.0 Refactor	Daniel
	Placement of Event: VRS or VA?	Daniel

Link to meeting recording:
https://us02web.zoom.us/rec/share/bLSVw9Oo-2xHFyqBZZ6jyM5u_Ca-uNkPw4bN3BVtkrS4Nhkj9HYMNSRzr5qddmtm.vt9FupUeRiTGoUfe?startTime=1701950192000
Passcode: 6%CveQQ.

Minutes:
Review of Variation Representation Specification (VRS) and its alignment with the SV (Structural Variation) model for VRS Version 2.0 Alpha. The agenda included a review of terminology changes and addressing the positioning of events within the model.
Larry mentions the need to finalize the SV changes in alignment with Version 2.0 Alpha and integrate precise information on sequence locations, alleles, and haplotypes, aiming for a Beta Version by Connect in April.
Daniel summarises the model changes, which are mostly practical, refining terminology and not significantly altering conceptual aspects. There is also a focus on defining clear sequences, junctions, and linked sequences, aiming for precise representation while considering implications for imprecise breakpoints.

Larry (in chat):
Just as a reminder, Alex has merged the 2.0-alpha-sv branch that Daniel is showing with the 2.0-alpha branch. Please make all subsequent PRs off of the 2.0-alpha branch.

Daniel describes flagging breakpoints as imprecise or precise, leaning toward an imprecise flag for better representation, given most breakpoints are known precisely. Examples shared also from Alex - more needed. He reviewed a PR highlighting SV changes in the Version 2.0 Alpha branch. Larry: Please delete the branch after merging changes.

Daniel went through the encoding of phasing within haplotypes, describing complex genomic structures involving alleles or adjacencies. He explained how a **haplotype could define a sequence traversal across derivative chromosomes**, navigating through various adjacencies and alleles from different chromosomes in a structured order. This traversal outlined the specific sequence, including regions like telomeres, and pointed out implicit details like chromosome orientation.
Larry question in chat:
would it be reasonable to use the term ‘connections’ for the attribute we are calling ‘connectedSequences’ or ‘sequenceTerminals’? Or is that not as accurate?
The term "connected sequences," reconsidering it as "connections" for clearer terminology related to describing connections rather than sequences.
Daniel: VCF (Variant Call Format) 4.5 changes - improving scalability and adding support for methylation, but these changes are probs not directly relevant to the VRS2 model at this point. Please contact reggan.thomas@ga4gh.org to be included in VCF meetings! (But the GKS meetings are more fun!)

Michael says there’s a need for an upgraded Beacon terminology aligned with VRS2 proposals, especially concerning structural variants. There should be considerations about modeling events like chromothripsis or chromoplexy within the VRS2 specification and how these might fall outside the scope of VRS2's primary purpose, suggesting a separate event classification step downstream.

Michael (in chat): For general information to re-iterate our alignment documentation for CNVs https://cnvar.org/resources/CNV-annotation-standards/#cnv-term-use-comparison-in-computational-file schema-formats

Daniel goes into the complexities of representing genomic structures and the challenges in creating a standardized model for classifying these events across different genomic data formats and tools. (Listen to the recording!)

Michael says it’s important to align and prepare for the Beacon system - we need community feedback and potential joint sessions for refinement, perhaps at Connect. BA - Add to Connect prep.

There's a distinction between low-level annotations within the VRS2 and more complex systemic changes or events that might require a separate structure or annotation system.
There's a recognition of the necessity for events within genomic data representations but also a concern about the current maturity and standardization of tools and terminology for handling these events across the community. Lack of standardization poses challenges in defining and classifying specific event types accurately.

Vivek (in chat):
If you have an adjacency with a linker of type RLE, which of the 2 sequence terminals reference sequence should be used to get the repeated sub sequence?
The current data model allows the linker to be of type RLE but, I am not sure if it makes sense to use an RLE for the linker in practice.

Michel shares paper in chat:
https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-82

Daniel illustrates classification complexities in genomic events, specifically focusing on breakpoints, genomic rearrangements, and the challenges with different tools' classifications.
The different tools handle various scenarios differently. Lists examples of genomic rearrangements like translocations, inversions, and duplications, emphasizing discrepancies between the idealized models represented in tools like VCF and the actual complexities observed in base pair sequencing.
There are difficulties in modeling such intricate genomic rearrangements within existing data structures, particularly highlighting the need for more sophisticated models to handle partial reconstructions and ambiguous classifications.
Daniel says that the current state of tools and models is insufficient in handling the intricacies of genomic rearrangements effectively.

Chromosomes, at the moment of specific events, are more like topologically organized DNA bundles. How do we label and understand these events while acknowledging that detailed reconstructions of entire sequences might not always be necessary for practical purposes?
Maybe understanding event classes and their parameters is more critical, especially when it comes to identifying fusion genes or specific sequences associated with certain tumors.
There are limitations in reconstructing entire sequences due to complexities, especially in cases of heavily amplified regions or breakage-fusion-bridge events with multiple breakpoints.

Daniel is keen to hold a Connect session to address the complexities of defining events across genomic projects and platforms, maybe over a beer, whilst having a walk along the lake.

2023-11-02

Hosting: Daniel Cameron, Mrinal Thomas (Epic)

Global Alliance
for Genomics & Health

Attendees "Name (Affiliation)": Beatrice Amos, Gabriel Bucci (San Raffaele Hospital in Milan), Rodrigo Martin, Sarah Hunt, Alex Wagner, Marco Morelli, Daniel, Nicolás Gaitán, Larry Babb, Tristan Nelson, Wes Goar, Daniel Puthawala

Apologies: Pille Hast

Please review and provide feedback: GitHub pull request.

	Agenda Item	Speaker/Leader	Time
	Welcome, Housekeeping	Beatrice	5
	SV Examples	Mrinal	15
	Use case discussions	All	

Link to slides:
https://docs.google.com/presentation/d/11Hm-_IvKGFUpdsqh_LiIrxjpvQazTxg1/edit?usp=share_link&ouid=110738840763072518407&rtpof=true&sd=true

Meeting recording:
https://us02web.zoom.us/rec/share/nWtvyU320-lpOwDe3K9ni2t__ZVMkex2yHZt35KZvLDvHZSFsmG5HBZG2IpT_aCT.zf32dpKhkPgKy-1A

Minutes:

Mrinal presentation shows the structure and representation of structural variants (SVs) in the context of the Epic project. Example demonstrates an SV involving a breakpoint between two genomic locations. This SV was represented using a structure that included break ends with sequence locations, orientation, and reference sequence accessions. He discussed the coordinate systems and the conversion of VCF-formatted SVs into the new representation. The definitions of "diverges before" and "diverges after" in the context of reference sequences are clarified. There was also a suggestion to visually represent SVs with both break ends on one side, which could make it easier to understand the "diverges before" and "diverges after" concepts. Mrinal showed a YAMM representation of the SV structure - please share the slides.

The representation of structural variants (SVs) have different notations and can lead to variations in how SVs are interpreted. Daniel highlighted issues with the VCF format, including redundancy and the lack of requirements for data consistency.
Daniel - How do we match SV calls, especially when using different callers with varying representations of the same event. It's important to represent SVs in a way that can be easily communicated and digested by different tools and databases.

Use cases:
Marco focused on cancer patient data and genomic data exchange using the Beacon infrastructure, where the challenge was to include structural variants in a standardized format. Suggest Marco - maybe join a Beacon meeting? Another use case involved comparing SV calls from different callers and aligning breakpoints.
Need for clear and comprehensive conventions for representing haplotypes and variants to address the challenges of matching and merging variant calls. Important to balance usability and expressibility in the conventions.

Data transformations are complex and it's important to maintain usability. There are multiple possible representations within the proposed VRS format, and issues related to defining the size and reference context of genetic variants. Daniel and Mrinal are discussing the need for clarity in specifying the relationship between references and samples - which Daniel says is a VCF limitation. It's also a challenge to define variants as deltas with respect to a reference, especially in the context of genome graphs. It's going to get even messier, says Daniel, eek! Still in discussion with VRS folks - the significance of phasing support for accurately matching variations and haplotype predictions within the VRS format.

Question - model SVs explicitly as active breakpoints or implicitly as flanking regions connected in haplotypes? There is redundancy in specifying the reference multiple times. The use of ontologies for SV representation - if anyone has feedback on specific use cases and ontologies in use, please give feedback to Daniel.
Meetings are in progress to align SV with VRS2.0.
Next meeting is in December, where Alex Wagner will give a presentation.

2023-10-05

Hosting: Daniel Cameron, Mrinal Thomas (Epic)
Attendees "Name (Affiliation)": Beatrice Amos, Catherine Procknow, Daniel Puthawala (Nationwide Children's Hospital in Columbus Ohio), Greg Sharpe, Likhitha Surapaneni, Gabriel Bucci (San Raffaele Hospital in Milan), Giorgia Gandolfi (San Raffaele Hospital in Milan), Vivek Nuthalapati (Epic), Peter Audano, Tim Hefferon (NCBI/dbVar), Drew Ellershaw (Medical Genomics Research dept at Illumina), Giorgia Gandolfi, Pillie Hallast, Rodrigo Martin, Kori Kuzma, Wes Goar, Michael Baudis, Tristan Nelson, Larry Babb, Fotios Tsetsos

Please review and provide feedback: GitHub pull request.

	Agenda Item	Speaker/Leader	Time
	Community use cases	Everyone!	15

	Changes to SV model based on community feedback	Daniel Cameron	5
	Align SV model with VRS 2.0 changes	Daniel Cameron	25
	Classification model	Daniel Cameron	15

Link to slides: daniel.l.cameron@gmail.com

Meeting recording:
https://us02web.zoom.us/rec/share/NAGeywHBun6V03N4e-UUdo7wO16Wiu3un_TAr4OWFrQd7huxVtZ-tpxtO2i69874.t8fW0_z9oFKHwi95?startTime=1696503070000
Passcode: Ep%?MY%1

Minutes:
Daniel summarizes updates and changes made to the schema based on feedback from the previous meeting. Further feedback is always welcome.
"Sequence and terminal" have been moved to "breakpoint," simplifying the model. The concept of an "event object" that contains variations and classifications, emphasizing the need for an ontology to describe these events is introduced.
Different classification schemes from VCF, DBVAR, etc have their own complexities.

Classification schemes used in tools like VCF, DBVAR, gnomAD, LINKS, Picard have their own approach and terminology for describing events. There is a need for flexibility in classification schemes to accommodate different use cases. Important to consider biological factors and experimental evidence when defining classification schemes for complex genomic events.
There are challenges of representing events with varying levels of precision and how to handle different representations of the same event.
BA to follow up: Use cases for the Genomic Data Integration (GDI) framework, particularly in the context of synthetic dynamic data and representing events like BCL, BL, FugU.
There is a need for a standardized approach to represent complex genomic events while allowing for flexibility in classification and reporting.

Wes Goar (in chat): I think having concrete definitions of simple events and using those as building blocks for more complex events is important.

Michael Baudis (in chat): Old Beacon structural variants discussion document which we'll start from probably, with fusions, translocations ... the next to finalize. This is not the same as the descriptive data model discussed here. Genomic Variants - Beacon
... see e.g. for CNVs (aligned w/ VRS) https://cnvar.org/resources/CNV-annotation-standards/

Wes mentioned the need for clear and well-defined building blocks for representing complex structural genomic events. We should define simple, standardized building blocks that can be used as the foundation for more complex structures. Providing guidance on how to model

complex events using these building blocks, while avoiding excessive ambiguity in the standards would be advised/preferred.
Should we allow users to define their own custom ontologies or annotations for specific events? Care should be taken to avoid fragmentation and ensure interoperability!
Important to build upon well-defined primitives and provide guidance for structuring more complex events, with the goal of achieving a balance between flexibility and standardization in genomic data representation.

How do we model genomic events, especially the representation of structural variations using sequence ontology and the creation of an event model within VRS? VRS should include an explicit event model to represent complex structural genomic events, with attributes like classification and sequence ontology being part of this dedicated model rather than optional fields in the variation set.
Question asked about how to normalize sequence orientation and infer event orientations.
Structural variations, such as chromothripsis, are complex, and there is a need for proper event classification.
Larry, Alex, Daniel and Mrinal to meet to discuss alignment with VRS 2.

Recommended paper: GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing | Genome Biology

2023-09-07

Hosting: Daniel Cameron, Mrinal Thomas
Attendees "Name (Affiliation)": Beatrice Amos, Sarah Hunt, Kevin Riehle, Greg Sharpe, Larry Bapp, Mindy Shi, Pille Hast, Rodrigo Martin, Rodrigo Martin, BSC, Vivek Nuthalapati

Please review and provide feedback: GitHub pull request.

	Agenda Item	Speaker/Leader	Time
	Demonstration of examples	Daniel Cameron	

Link to slides: **daniel.l.cameron@gmail.com**

Meeting recording:
https://us02web.zoom.us/rec/share/ZZFS7nIFGONAmU9dnYVZcK2az_fJtTKh1yP8XZkL_Mu09Bv4-4nd09Bn5yeeGviL.ZcOpLespoCx_bI3T?startTime=1694083833000
Passcode: p42dSA1!

Minutes:
Demonstration of examples and syntax. It aimed to summarize the content, focusing on key details.

LB: question about the use of location objects instead of just start points for defining breaks. There is a need for ambiguity handling in specifying break locations. Intervals discussed, where a range or a number could represent break positions. Clarification needed - a break point is made up of both left and right break ends.
DC: explained structural variation breakpoints, their definitions, and the need for precise notation.

The concept of a break point and its components.A break point consists of a pair of break ends. Break ends represent the points where the reference sequence diverges to another sequence. Break ends can be thought of as ranges, and they may include ambiguity about their exact location.
RM: how do you represent scenarios where only one break end is known, for instance, when reaching the end of a molecule or when the partner break end is uncertain.
DC gave examples with inserted sequence within a structural variation and the association of sequence with break ends rather than break points. Considered cases where break ends have known flanking sequences but are not uniquely positionable in the reference.

Handling Unpaired Break Ends: how do we represent scenarios where only one break end is known? Agree to always maintain two break ends for each break point, even if one of them represents an unknown location. Simplifies the model and avoids the need for special placeholders or unpaired break ends.

Placement of Sequence Information: Pushing sequence data up to the break point level instead of associating it with break ends. This is a way to standardize and normalize the model. It simplifies the representation of sequences within structural variations.

Encoding haplotype ordering

Tandem Duplication: DC shows example of tandem duplication within a ref sequence, represented by orange, blue, and green blocks (see the slides).
Translocation: one break end can come from one ref sequence, and the other from a different ref sequence.
When traversing a reference sequence, a structural variation (e.g., duplication or deletion) is essentially a change in the sequence. Instead of thinking in terms of blocks, think of it as transitioning from one sequence to another at a specific location.

Variant Classification: Differentiating between variations (e.g., deletion vs. duplication) might require additional analysis, especially in complex regions with multiple rearrangements.

Concept of translocations - involve jumping from one genomic location to another (independent of ref sequences).

Handling Viral Insertions: the model accommodates viral insertions. Typically, a single break end represents the jump to the viral sequence, and a second break point may represent the return jump to the host genome if all sequences are sequenced. Flexibility allows for complex insertion events to be represented.

Complexity in Haplotype Ordering: The complexity of defining haplotype ordering was acknowledged, especially when dealing with variants that are both cis (on the same chromosome) and trans (on different chromosomes).

MT: Backward Compatibility. Subtle changes, such as ordering members in haplotypes, are introduced to accommodate structural variants while not disrupting the handling of non-structural variants.

Impact on Flanking Regions: Variants inherently make claims about the regions they affect, including flanking regions. Current model doesn't explicitly define claims about regions with no listed variants.

Complex Traversal and Ref Compatibility: The discussion highlighted the challenges posed by complex variants like inversions. These variants can result in traversal orders that don't make sense in multiple reference sequences. This underscores the difficulty of defining traversal orders when references differ substantially.

Overall, the goal is to develop a model that supports both simple and sophisticated tools for interpreting structural variants while enabling the classification of events and accommodating phased variants within rearrangements.

2023-08-03

Hosting: Daniel Cameron, Mrinal Thomas
Attendees “Name (Affiliation)”: Beatrice Amos (GA4GH), Alex Wagner, Larry Babb, Rodrigo Martin, Peter Audano, Vivek Nuthalapati (Epic), Heng Yi, Tristan Nelson, Greg Sharpe, Valeriya Mikova, Kevin Riehle, Terry O’Neill, Catherine Procknow, Likhitha Surapaneni, Sarah Hunt, Jie Liu, Nicolás Gaitán (BSC CG group), Andrejs Puckovs, Kori Kuzma

	Actions Arising	Assigned To	Deadline

	LB suggests examples would be helpful. DC to provide for next meeting.	daniel.l.camero…	Next meeting

	Agenda Item	Speaker/Leader	Time
	Welcome, housekeeping: GA4GH Connect 2023: San Francisco - Submit a poster abstract: form - Register for the conference: https://broadinstitute.swoogo.com/ga4gh11thplenary - New hotel booking link for Plenary >> https://book.passkey.com/gt/219333164?gtid=6a35e33de59169ab53000d64ed38ab96		
	The main discussion revolved around definitions for haplotype, genotype, and the new SV fields.		
	Larry Babb, Alex Wagner - alignment of SV with VRS2.0 alpha dev		

Link to slides: daniel.l.cameron@gmail.com

Link to meeting recording:
https://us02web.zoom.us/rec/share/rR520D34RtxWkVFHhCPMKEx1ytdUkq-kXYmuCf3cAbQwheDLC-jvoCbFDBZ0T876.t2ka8ZRVNF_nY8Tw?startTime=1691057204000
Passcode: nW.BmuV1

Minutes:
Part 1 of presentation:
The main discussion revolved around **definitions for haplotype**, **genotype, and the new SV fields**. The presentation focused on scenarios involving chromosomal copies, duplications, and overlapping alleles.
There was a query about whether the **current definition of overlapping alleles** is sufficient or if proximity needs to be considered.
Issue of sample-reference genomic loci mapping, addressing cases of multiple-to-one, one-to-multiple, and many-to-many mappings.
DC clarification of genotype, emphasizing that genotype counts should consider the total count of the genomic locus, not the number of molecules.
SV terminology: SV-VRS and its compatibility with expectations of genotypes and haplotypes.

LB:

Clarifying **haplotype definitions** and complexities.
Questions raised about having nested haplotypes and sample-reference correspondence.
Consideration of phase blocks and potential integration in SVverse.
Discussion on capturing chromosome copy origins within blocks.
Recognized impact of such information on haplotype definitions.

AL:
Discussion about **multiple copies of regions and phasing**.
Mention of using derived sequence expression for insertions.
Need to expand the notion of underlying sequences in the model.
Consideration of composing sequences from components.
Challenge of reconstructing relative references between sample and reference.
Issues with breakpoint calling and reconstruction for liftover.

AL/DC:
Discussion about multiple copies of regions and phasing.
Mention of using derived sequence expression for insertions.
Need to expand the notion of underlying sequences in the model.
Consideration of composing sequences from components.
Challenge of reconstructing relative references between sample and reference.
Issues with breakpoint calling and reconstruction for liftover.

PA:
Challenges in representing structural variation with long read sequencing:
Difficulty in annotating and representing variation within duplicated regions.
Discussion about representing duplications and insertions in BCF.
Need for clear representation of derived reference sequences and variations.
Issues with current standards in capturing complex structural variations.
VCF 4.4's support for defining breakpoints and phase sets.
Versatile solution in SV-VRS by using per haplotype phase sets.
Addressing ancestral sequences without good reference correspondence.
Proposal to use breakpoints to ancestral blocks and define against closest references.
Challenges of representing unique sequences and large-scale deletions.
Importance of finding meaningful reference representation for unique sequences.
Exploration of the size and nature of unique sequence blocks.

PA: Desire to analyze variation in unique sequences across multiple individuals.
Discussion about using VRS to define non-reference sequences.
Using VRS to define blocks and variations against non-reference sequences.
Versatility of VRS in handling variations in non-reference sequences.
Complex! and need for further exploration in this area.

Part 2 of presentation: SV terminology

Variability in SV terminology and the need for agreed-upon terms.

Terms required for SV-VRS: copy number variant (absolute/relative), breakpoints/breakends (SV callers), DNA adjacencies, DNA abundance (biological event signals).
SV-VRS focuses on structural changes causing breakpoints, with DNA adjacency and copy number change signals.
Biological event definition: transition from one stable genomic state to another (including deletions, duplications, inversions, etc.).
Conceptual event definition: set of related biological events, like mobile element reactivations or fragile sites.
Proposed terminology: CNV (absolute/relative), structural variant, genomic rearrangement, and event association.
Challenges in defining copy number change segments within events, considering overlaps and nested segments.
Copy number change definition: absolute or relative, with potential tooling flexibility.

DC would like feedback and thoughts on the terminology model. LB suggests examples would be helpful. DC to provide for next meeting.

See and provide feedback here: https://github.com/ga4gh/vrs/issues

2023-07-06

Hosting: Daniel Cameron, Mrinal Thomas
Attendees "Name (Affiliation)": Beatrice Amos (GA4GH), Greg Sharpe, Jie Liu, Kendra Scheer (university of Buffalo), Michael Baudis, Kevin Riehle, Larry Babb, Likhitha Surapaneni, Nuno Agostinho, Peter Audano, Pille Hallast, Rodrigo Martin, Timothee Cezard, Tobias Marschall, Tristan Nelson, Nicolás Gaitán (BSC CG group), Charikleia Karageorgiou, Jia Zhang, Sarah Hunt

	Actions Arising	Assigned To	Deadline
	Share real world examples of variations	DC	
	Alex, Larry and Daniel/Mrinal meeting	BA to convene	

	Agenda Item	Speaker/Leader	Time
	VRS Pull Request #428: minimal schema	Daniel	30
	VRS GitHub Issue #429	Mrinal	10

Link to slides: daniel.l.cameron@gmail.com

Meeting recording:
https://us02web.zoom.us/rec/share/ydymXruskYj_We5Q5nwngqQMOFA9vLrbWlLjzTsE-jBQOwuEvG2uYHeKd6B4jpSd.0g29U1qh6bvx7rwJ?startTime=1688638937000
Passcode: 9^2n#6#*

Minutes:
Look at pull request on github and comment.

LB: It would be very helpful for me (and I assume others) if we could show some real world examples of variations that represent the models and cases we are covering (in chat).

Encoding phasing in haplotype:
LB: seeks clarification of end goal of SV, usecases
TM: need for representing/encoding SVs, e.g. in GWAS studies
MT: Breakpoint and breakend notation clarification
P: diff between large and small variants are artificial. With tech methods today, we're picking up small and large. This info helps downstream studies. Annotating ambiguity is helpful.

Seeking feedback on VCF SVLEN and CILEN. Missing from current proposal. Discussing ambiguities in breakpoint definitions.

Haplotype ordering issue:

MT: Github issue #429

2023-06-01

Hosting: Daniel Cameron, Mrinal Thomas
Attendees "Name (Affiliation)": Rodrigo Martin,Michael Baudis,Beatrice Amos,Catherine Procknow,Pille Hallast,David Torrents,Juniper Lake,Xinghua Mindy Shi,Kevin R,Tim Hefferon,Charles Lee,Likhitha S,Greg Sharpe,Nuno A,Sarah Hunt,Vivek N,Tristan N,Kori K,Alex Wagner,Eugene Miller,Reggan Thomas (GA4GH)
Apologies : Tobias Marschall

	Actions Arising	Assigned To	Deadline
1.	Draft document on data structure and notation	DC	6 weeks
2.	Review issue 425 and provide thoughts/feedback/solutions	All	4 weeks

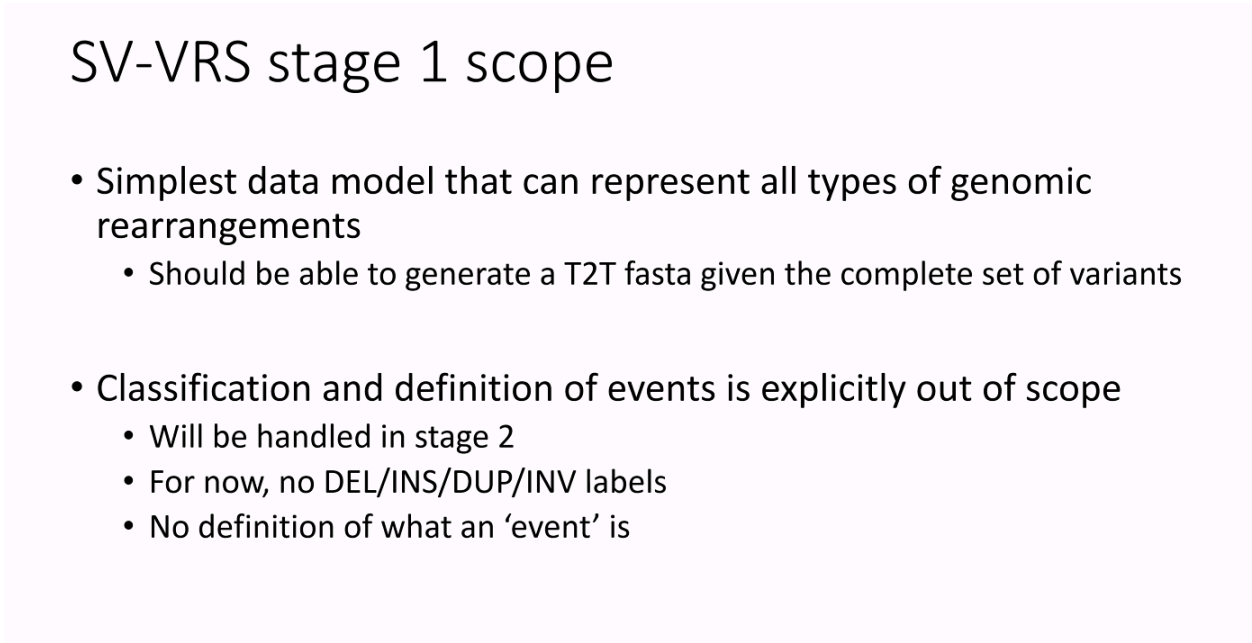
Global Alliance
for Genomics & Health

	Agenda Item	Speaker/Leader	Time
1.	Indicative progress timeline	DC	10m
2.	Explanation of intended concepts See https://github.com/ga4gh/vrs/issues/425 for background	DC	30m
3.			

Meeting recording:
https://us02web.zoom.us/rec/share/kRGvBINti4yE7zota5mEiwHcrUzGLW1CRMt2K5_V5Ped8t0nw8WXlisafKIkzcRc.Vi6tCUaf-hEKJ9Y2
Passcode: R2x*D&a#

Minutes:

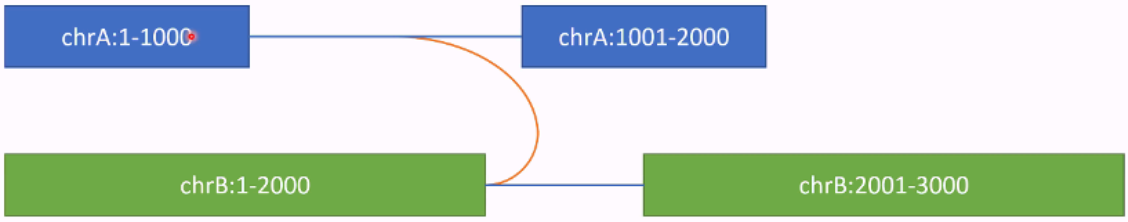
Scope 1



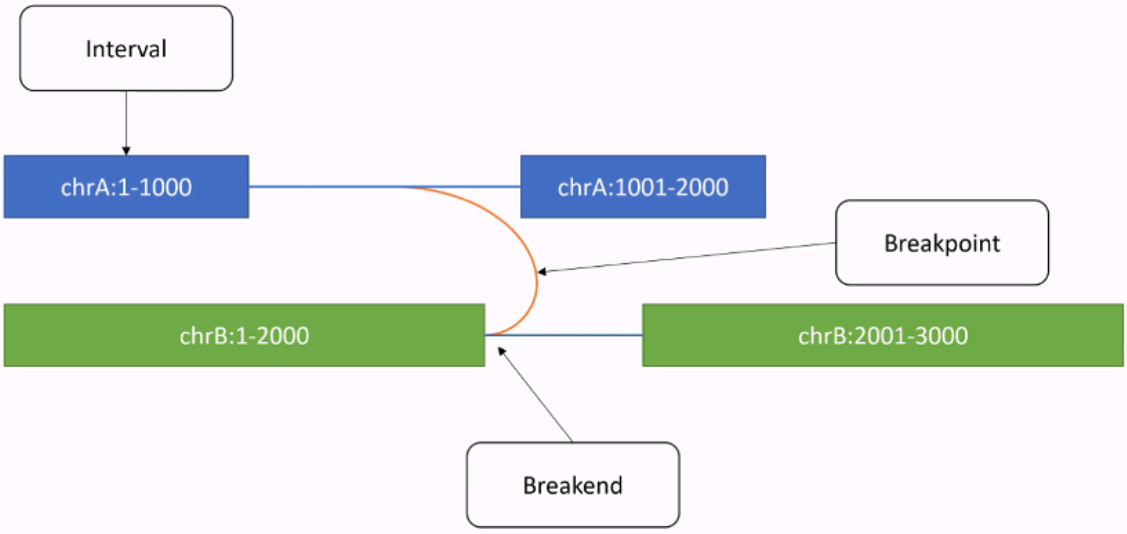
Break point graph decomposition

Breakpoint graph decomposition

- Genome is decomposed into contiguous constant-CN segments
- Edges between segment start/end define genomic adjacency structure



Terminology



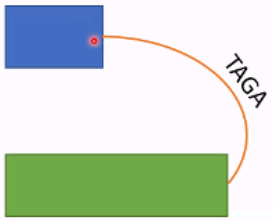
Breakend

- The presence of a non-reference genomic structure at the start or end of a genomic interval
- May signal end of chromosome
- Location of other side may be ambiguous (c.f. VCF single breakends)



Breakpoint

- Defines a genomic adjacency between intervals
  - 2 unordered breakends
- May contain sequence
- May contain gap
- May be ambiguous:
  - Imprecise: caller unable to determine exact location
  - Precise: flanking interval sequence is homologous



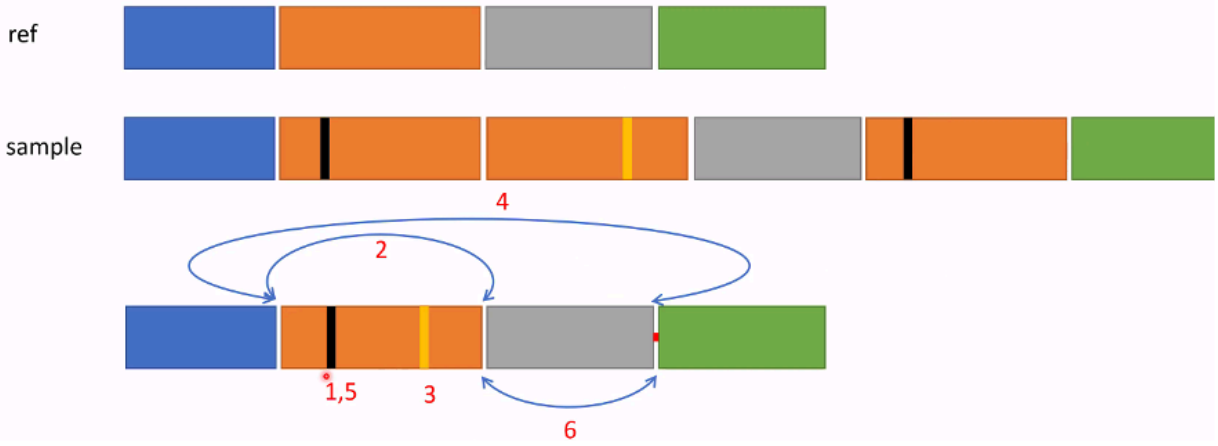
Phasing

- Existing terminology break down for SVs



Proposed phasing model

- Define partial phase set and a relative ordering



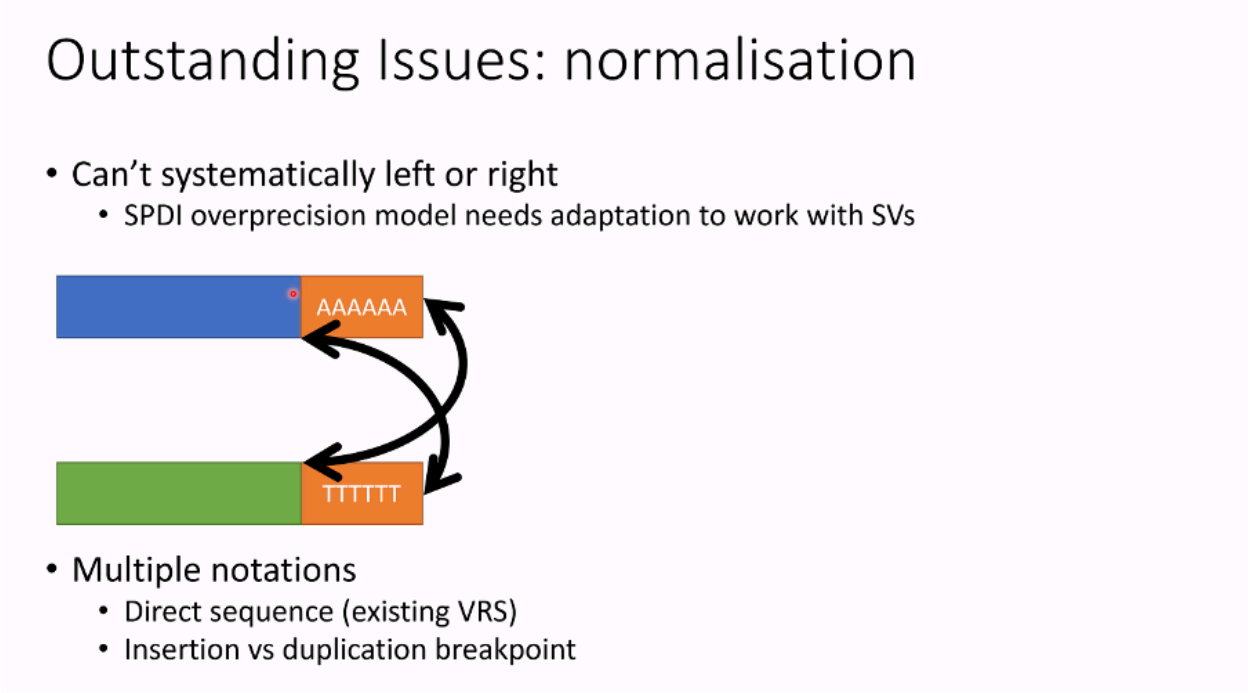
TH – Directionality ?
DC – Recommend to 6 -1

DC - this is the order.

DT – Diploid is a nightmare. Q – how will connect with the entry with the input and output part? Does this classified from the info from the variant caller? Or imposing any new classification? Classification is a mess
DC – Wave 1 – no classification Wave 2 – built on top of that.
DC – duplication – CNV gain. Deletion or duplication
DT – Terminology – you can arrive from different practise.
What about the genotype? –

DC – Snips in blue. – haploid genotype, suddenly become triploid genotype. We need to define the terminology of the genotype means.

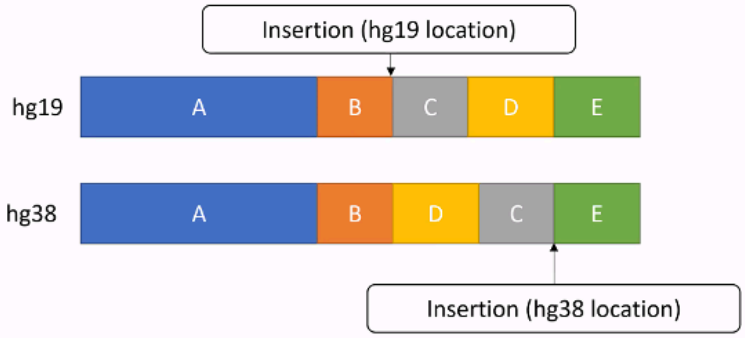
DC- this is example of copy number boundaries.



DC – bigger issues is with notation. You can represent as direct sequence
Or you can have 2 different insertion. Do we want to have a preferred notation of that. W
To put a break point, we need SPDI.

Outstanding Issues: multi-reference ambiguity

- Mixing reference introduces ambiguity



- Breakpoints within inversions are even worse

DC – this model defines the single reference . hg19 path doesn't make the same for hg38

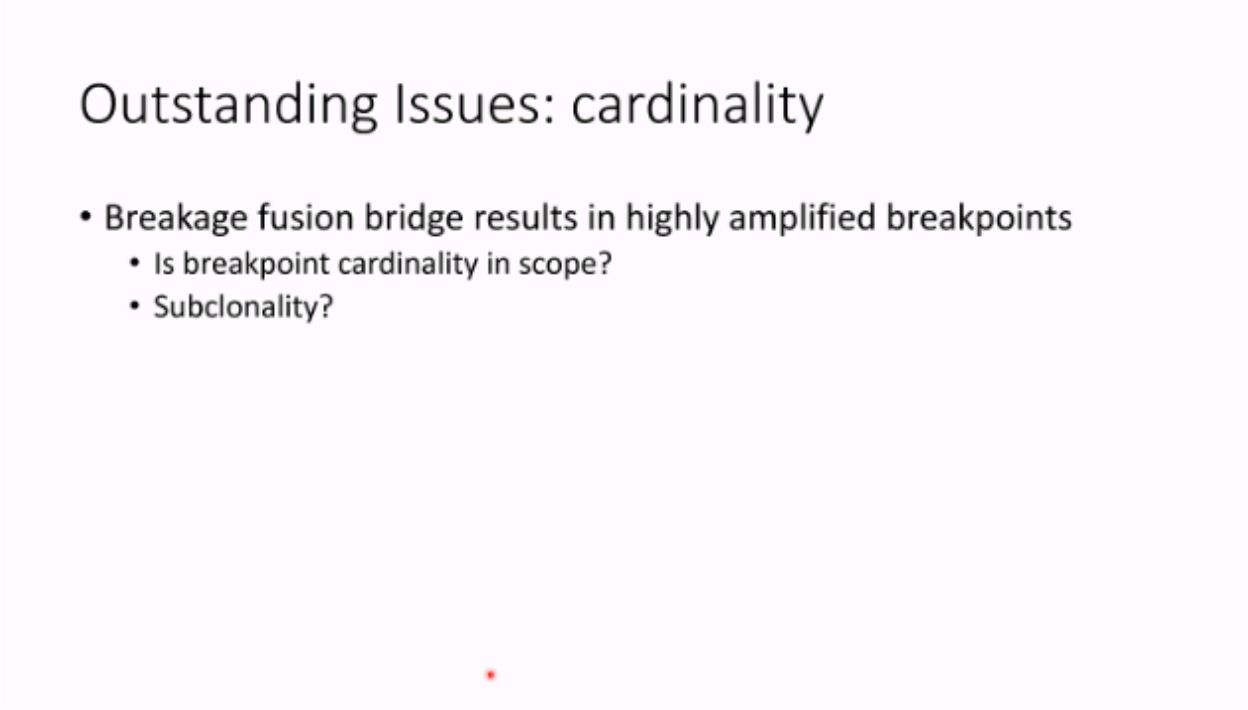
Outstanding Issues: alignment

- Multiple notations possible
  - Existing VRS insertion notation
  - Insertion breakpoint
  - Duplication breakpoint



- Not restricted to just tandem duplications
  - Mobile element insertions: insertion or translocations to/from donor site?

DC – rearranged grey to green . Explicit put the reference to Jason. Should we be explicit? The implied breakpoint may not exist. We need to explicitly state that. As the preferred option.



Outstanding issues:
– Multiple meaning

2023-##-##: TEMPLATE

Hosting: Daniel Cameron, Mrinal Thomas
Attendees "Name (Affiliation)":

	Actions Arising	Assigned To	Deadline

	Agenda Item	Speaker/Leader	Time

Minutes:
Meeting recording: