LSG Work Stream 2022

LSG / File format

Date: Tuesday, 14 Nov 2022
Time: 16:00 - 17:30 pm EST
Meeting Chair(s) : Daniel Cameroon

VCF 4.4 Release Candidate 2 Public Feedback

Description : The upcoming VCF version 4.4 overhauls how structural variants and copy number variants are handled as well as introducing support for STRs and VNTRs. In this session, an overview and justification for all changes will be presented. This public presentation of the draft VCFv4.4 specifications gives the public a final chance to provide feedback and request changes before the specification is finalized.

	Agenda Item	**Speaker**	**Time**
1.0	Introduction and goals setting	Daniel Cameron	5min
2.0	Conceptual model of genomic rearrangements in VCF 4.4	Daniel Cameron	5min
3.0	Overview of changes	Daniel Cameron	15min
4.0	Impact on SV & CNV callers	Daniel Cameron	5min
5.0	Support for STRs and VNTRs	Daniel Cameron	15min
6.0	Public feedback		30 min

Session: VCF 4.4 Release Candidate 2 Public Feedback
Tuesday, Nov 14: 16:00 - 17:30 am EST

Zoom Link:
https://us02web.zoom.us/meeting/register/tZYocu6orjoiEtBYD0lK1ahaqFpE0vzkL6qr

Best Practices for Virtual Speakers

- Make sure laptops are plugged into power prior to and during your session.
- Check cameras to ensure they are centered, sit in a well-lit area and ensure your background is what you want for your presentation.
- Use an ethernet cord for the best connectivity; if using Wi-Fi, make sure to test your Wi-Fi connection prior to the conference to ensure it works.
- Earbuds or headphones will prevent audio echoes.
- Please stay muted except when speaking.
- All sessions will be recorded, and the chat boxes will be saved.
- Have water or a beverage close by.

Attendees “Name (Affiliation)”:
Oliver Hoffman,Rob Davies,Fabian Klotzl,Takudzwa M, Andrew Patterson,Geraldine V,Jeffery Yuan,James Bonfield,Patrick Cheung,Ray Kransinki,Bob Dolin,Heidi Sofia,Vivek Krishnakumar,Sujai Chari,Mamana M,Erfan Sayyari,Eric Roller,Gordon Krieger,Praveen Nadukkalam,Jonathan L, Marcus K,Marzie Rasekh,Jeffery Yuan,Kent Ho

Zoom recording :
https://us02web.zoom.us/rec/share/eKfGY1kL_AvQSmQZR009KvDqkxhmdOUif-QSW58nL_j2gRhJFgJ0IRuDfEwFyJ9H.g64eW8-1RX03rKvU?startTime=1668458956000

Meeting transcript -
https://otter.ai/u/7s6Kt3Cve9jcIqStewp7yY520tw?f=home&tab=summary

Notes:

DC - At a very high level 4.4 is designed to fix structural variant support in VCF. There's a number of significant issues with 4.3 for representing structural variants that were resolved in 4.4.

DC - Variant detection and interpretation
OH - Did the SVCLAIM fields have to be exact breakpoints or can they be fuzzy?
DC - dont need to be descriptive
JB- I'm reminded of the EMBL "database" file format with gene markup that had start..end notation, or {a.b}..{c.d} for a region starting between a and b and ending between c and d.
JB - define as offset relative to normal position. Haven't defined a confidence interval. One side it can be 100 percentage but on the other

Not suggesting it here, but similar things have been done before.
(https://www.insdc.org/submitting-standards/feature-table/#3.4.3)

OH- And then there are approaches to fuzzy-merge it all back together ;-)
(https://github.com/mkirsche/Jasmine)

DC - Yes. At the spec level not defining it. We define how to do variant interpretation.

OH- Images Daniel are showing are from the excellent Linx
(https://github.com/hartwigmedical/hmftools/tree/master/linx)

Symbolic Structural variant
ALT field Closed vocabulary
Subtype must have identical interpretation
Support for multiple symbolic ALT alleles
Field changed to Number =A or Number =2A
SV type deprecated
DPADJ,CNADJ,CICNADJ deprecated

- Unsed
- Multiple adjacent segment

SVLEN redefined to be positive in 4.4 version
abs(SVLEN) for backward compatibility

SVLEN number = A, But END still Number = 1
 END used for Indexing,NON ref alleles

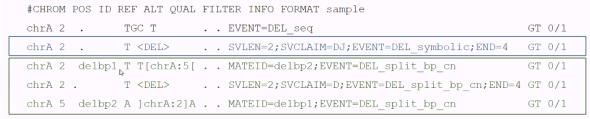
Copy Number Variant:
CI and CICN now Type = float

- Backward compatibility

Info CN
Alleles specific Copy number
Fomat

Examples:

Examples



```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample
chrA 2  .      TGC T        . . EVENT=DEL_seq                                    GT 0/1
chrA 2  .      T <DEL>      . . SVLEN=2;SVCLAIM=DJ;EVENT=DEL_symbolic;END=4      GT 0/1
chrA 2  delbp1_b T T[chrA:5[ . . MATEID=delbp2;EVENT=DEL_split_bp_cn              GT 0/1
chrA 2 .       T <DEL>      . . SVLEN=2;SVCLAIM=D;EVENT=DEL_split_bp_cn;END=4 GT 0/1
chrA 5  delbp2 A ]chrA:2]A . . MATEID=delbp1;EVENT=DEL_split_bp_cn              GT 0/1
```

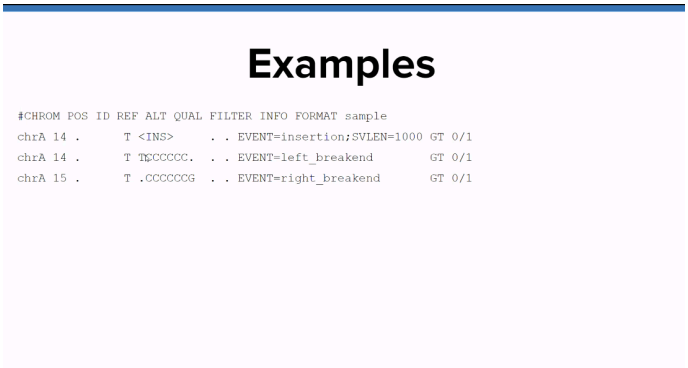
JB : I see no END on the breakpoints. Does anything bother indexing them? If not, should we recommend using simpler forms *where possible*?

DC - Symbolic allele easier to process. Specs not going to make the recommendation.

ER: see no END on the breakpoints. Does anything bother indexing them? If not, should we recommend using simpler forms *where possible*?

DC - 2 possible interpretation
Need to add to the specs - clarify SVCLAIM=D is implicit for copy-number events
JB - Could it just be part of CIPOS? Eg CIPOS=0,5,2? With ,2 being totally optional. That also implies we can retrofit it with ease later on.
DC - it's already a list of pairs. VCF does not support list of list.
Need to have it as a separate field.

Examples



```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample
chrA 14 .      T <INS>      . . EVENT=insertion;SVLEN=1000 GT 0/1
chrA 14 .      T T.CCCCCC.  . . EVENT=left_breakend        GT 0/1
chrA 15 .      T .CCCCCCG   . . EVENT=right_breakend       GT 0/1
```

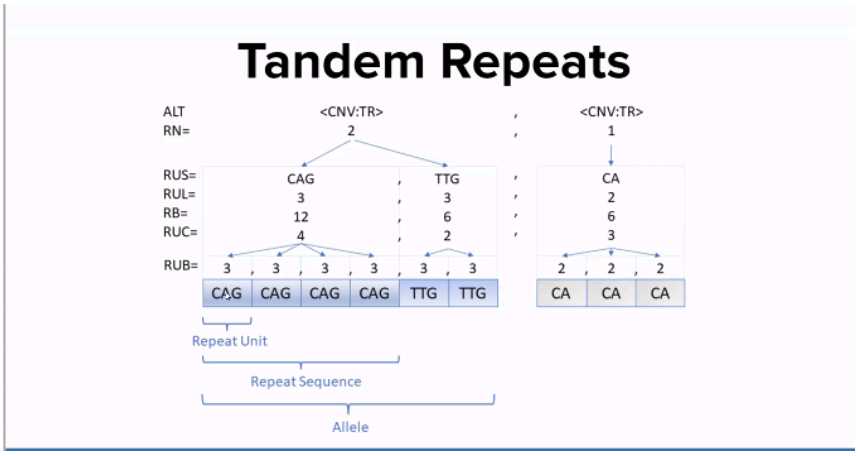
TR - another big change

VCF already has TR.

New CN TR ALT Allele

CNV TR allele composed of one or more

- Repeteaed seq composed of one or more
- Repeated units
- RN, RUS, RUL,RN,CIRB,RUC,CIRUC,RUB



Instead of writing repeat seq, write repeat length. In this case we have 4 copies

JB - So does this copes with overlapping repeats? Eg ACAGACAGACAGAGAGAG - there's an AG there that is part of both ACAG and AG rep. Or do we canonicalize things and put it in the left repeat? Or fudge RB/RUC?

DC - There are multiple possible ways to represent repeats. In specs we are not to define it. You actually split that in CACAG , AG, CAG .. Truncated repeats.

Is it a change of repeats or snips?

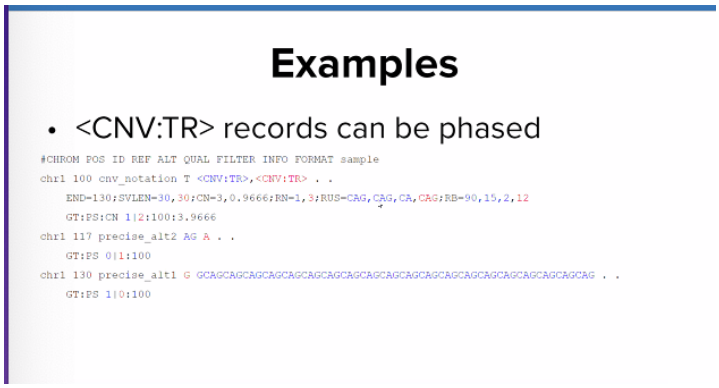
CNV caller to determine. Not worried about canonicalization, dont make sense.

JB - So does this copes with overlapping repeats? Eg ACAGACAGACAGAGAGAG - there's an AG there that is part of both ACAG and AG rep. Or do we canonicalize things and put it in the left repeat? Or fudge RB/RUC?

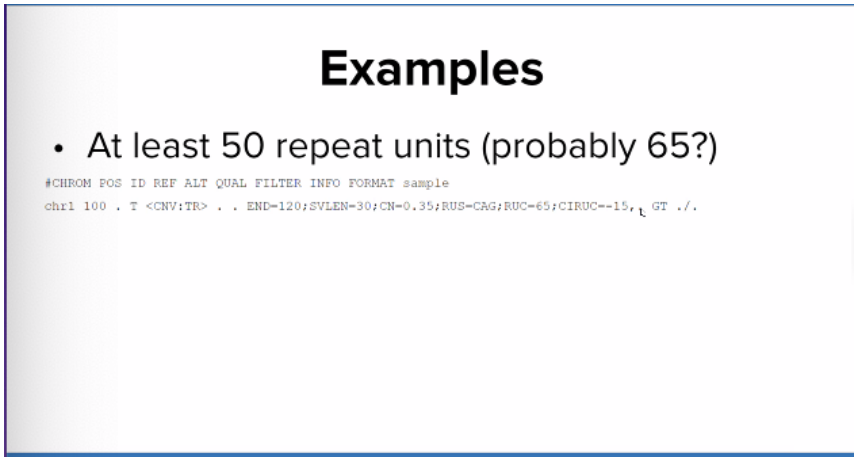
DC - none are overlapping

OH - I'll throw in questions here as well - what does deprecated mean for VCF in general? Does 4.4. recommend parsers throw a warning if they encounter, say, SVTYPE or an error?

DC - They no longer specification defined.



ER - I'll throw in questions here as well - what does deprecated mean for VCF in general? Does 4.4. recommend parsers throw a warning if they encounter, say, SVTYPE or an error?



Reminder / Clarification

- ALT allele field can be sensitive
- Bundles have been removed
- POS is the position before the symbolics SV
- <*> is not symbolic

Meta- info

PSL - slide missing
Snip within SV or ordering within the SV

OH - How can the reference copy number of a VNTR be calculated from this record? In case of VNTRs, with indels inside repeat units, the svlen/rul might be different from the reference copy number.
DC - happy to incorporate them.

MR - How can the reference copy number of a VNTR be calculated from this record? In case of VNTRs, with indels inside repeat units, the svlen/rul might be different from the reference copy number.

DC - If there is a user requirement, DC can add this to specs. All tools write single symbols.
 Existing annotation tool should work fine with 4.4

Outstanding issue - Methylation.
If this is an important issue, please raise an issue in hts specs

One month for final comments.

Key takeaways:

- Final chance for public feedback on VCF 4.4 before finalization

Next actions:

- Need to check with the secretariat if they can publish VCF4.4 for public comment one more time.
- Address public feedback raised on hts-specs before finalization of VCF4.4

- Plan for April Connect
  - x
  - x
- Plan follow-up meetings
  - X
  - x
- Any other business arising
  - X
  - x