Agenda – VCF Pull Request #231 Video Call

# Date:

27 September, 2017

# Attendees:

**@d-cameron** (David Cameron)

**@Shuang-broad** (Steve Huang)

**@cyenyxe** (Cristina Yenxye Garcia)

**@thefferon** (Tim Hefferon)

**@lopezjo** (John Lopez)

**@cwhelan** (Chris Whelan)

**@xuefzhao** (Xuefang Zhao)

# Goals of this call:

To achieve resolution to the issues listed in the table at the end of this agenda.

# Background:

PR #231 (<https://github.com/samtools/hts-specs/pull/231>) was motivated by the observation that the reporting of structural variation by different VCF authors is highly inconsistent. This is likely due, at least in part, to the failure of the VCF spec to provide sufficient guidance – in the form of clear definitions and solid examples – to promote consistent reporting. A general disregard for the spec instructions that *do* exist – especially w.r.t. REF/ALT, SVTYPE, SVLEN, etc. – contributes to the problem. The result is inconsistent SV reporting that confounds attempts to consolidate datasets from different sources, which could otherwise result in the fruitful exchange of data to the mutual benefit of all participants. Let’s try to improve this.

# Discussion Topics:

## SVTYPE

A central topic of PR #231 concerns the appropriate use of the SVTYPE info tag. David has suggested SVTYPE can be interpreted in one of two ways:

1. as shorthand for the equivalent BND representation of a variant
2. as an assertion about the type of biological event that produced a variant

Which of interpretation is used has implications for the VCF spec. David has been asked to give a brief explanation of the first interpretation. *(David)*

1. SVTYPE – Are SVTYPE values intended as shorthand for equivalent BND representations (Interpretation #1) or as indications of the types of biological events that produced the variant (Interpretation #2)?
2. Is BND an SVTYPE, or is it something different? Does it belong in the list?
3. Current values include: DEL, INS, DUP, INV, CNV
   1. what about deletion-insertions? do we need to add RPL to the list?
   2. how will translocations be represented, in light of #2?
4. Proposal to be considered: Since SVTYPE is redundant with REF/ALT, deprecate SVTYPE, and us another INFO tag (such as SVCLASS, EVENTTYPE, etc.) to serve the needs of those who subscribe to SVTYPE Interpretation #2.

## BND/TRA

(Any details which have not already been addressed by discussion above)

## IMPRECISE

Tim originally proposed the IMPRECISE tag could be deprecated, in part because it remains undefined in the spec, in part because one could assert it is redundant with CIPOS and CIEND. However, subsequent discussion largely justified the usefulness of keeping the IMPRECISE tag, namely as a way of distinguishing between calls:

1. which cannot be precisely placed despite complete sequence coverage
2. whose breakpoints cannot be precisely determined due to lack of sequence coverage (e.g., paired-end)
3. whose exact breakpoints are known, but whose full sequence content and size are not known (e.g., split-read)

Proposal: The IMPRECISE tag and its intended usage must be carefully spelled out in the spec, so there is no ambiguity in its intended usage. *(Tim nominates David to write up this addition to the spec.)*

## CIPOS, CIEND, CIVALUE, and “hard boundaries”

‘CIPOS’ and ‘CIEND’ are defined in the spec as follows:

CIPOS: Confidence interval around POS for imprecise variants

CIEND: Confidence interval around END for imprecise variants

However the spec provides no definition of “confidence interval”, nor does it provide solid examples that clarify usage of CIPOS and CIEND, so that everyone who encounters them in SV VCF understands exactly how they ought to be.

The PR introduces the tag ‘CIVALUE’:

CIVALUE: Degree of certainty expressed by CIPOS/CIEND: ‘1’ indicates 100% certainty that a breakpoint falls within confidence interval; default value is ‘0.95’

Another possible improvement has been observed in the form of revised definitions:

CIPOS: Confidence interval of the start position listed as the distance to the inner left side breakpoint

CIEND: Confidence interval of the end position listed as the distance to the inner right side breakpoint

Associated variants would then use a value of ‘1’ wherever a hard boundary is indicated.

## EXAMPLES

The VCF spec could more effectively accomplish its purpose with respect to SV if it were to include an example, in both shorthand and BND-type notation, for each type of variant it is meant to accommodate. Furthermore, the exercise of generating these examples will likely uncover inconsistencies and/or weaknesses in the spec. *(Tim proposes we cooperate to generate a robust set of examples in both notations.)*

Vote Topics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proposed Action | Vote | | | Comments |
| Yes | No | Un |
| Shift from SVTYPE to REF/ALT for haplotypes |  |  |  |  |
| If Yes: | | | | |
| Strictly enforce “SVTYPE” controlled vocabulary in ALT |  |  |  |  |
| Disallow colons, subtypes in ALT |  |  |  |  |
|  |  |  |  |  |
| Delete or change the following in the controlled vocabulary: | | | | |
| DEL |  |  |  |  |
| DUP |  |  |  |  |
| INS |  |  |  |  |
| INV |  |  |  |  |
| CNV |  |  |  |  |
| BND |  |  |  |  |
| RPL |  |  |  |  |
| (other?) |  |  |  |  |
| Maintain a list of SVTYPEs (Interpretation #2) with the following tag: | | | | |
| SVTYPE |  |  |  |  |
| SVCLASS |  |  |  |  |
| EVENTTYPE |  |  |  |  |
| (other?) |  |  |  |  |
| Disallow colons in above list (use single terms only) |  |  |  |  |
| Confidence Intervals: | | | | |
| Add definition of “confidence interval” to spec |  |  |  |  |
| Add usage notes for CIPOS and CIEND |  |  |  |  |
| Implement `CIVALUE` INFO tag |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |