**Variant Numbering Approaches**

1 2

0-based 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

1-based 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

Reference Sequence 5’ A T G T C G T T C A C G G T C A T C A C A C A G A C C A A 3’

1 g.1A>T subs **T**

2 g.2delT del A **-**

3 g.4\_5delTC del . . G **- -**

4 g.8\_9delTC del . . . . . G T **- -**

5 g.10\_11insT ins . . . . . . . . . A**T**

6 g.12\_13insTAA ins . . . . . . . . . . . G**TAA**

7 g.16\_17insATT ins . . . . . . . . . . . . G T C**ATT**

8a g.18\_19delCA,g.18\_19[2] del . . . . . . . . . . . . . . . . T **- -** C A C A

8b g.20\_21delCA,g.20\_21[2] del . . . . . . . . . . . . . . . . T C A **- -** C A

8c g.22\_23delCA,g.22\_23[2]\* del . . . . . . . . . . . . . . . . T C A C A **- -**

9 g.24G= match . . . . . . . . . . . . . . . . . . . . . . . **G**

10a g.25\_26insC,g.26dupC ins . . . . . . . . . . . . . . . . . . . . . . . . A**C**C C

10b g.26\_27insC,g.27dupC ins . . . . . . . . . . . . . . . . . . . . . . . . A C**C**C

10c g.27\_28insC,g.28dupC ins . . . . . . . . . . . . . . . . . . . . . . . . A C C**C**

10d g.27[3]\*,g.26[3],g.25[3] rep . . . . . . . . . . . . . . . . . . . . . . . . A C C**C**

|  | HGVS | Alignment Format (1-based) | | | | Variant Format (1-based) | | | | Interval Format (0-based) | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Start | End | Ref Allele | Alt Allele | Start | End | Ref Allele | Alt Allele | Start | End | Ref Allele | Alt Allele |
| 1 | g.1A>T | 1 | 1 | A | T | 1 | 1 | A | T | 0 | 1 | A | T |
| 2 | g.2delT | 1 | 2 | AT | A | 2 | 2 | T | - | 1 | 2 | T | - |
| 3 | g.4\_5delTC | 3 | 5 | GTC | G | 4 | 5 | TC | - | 3 | 5 | TC | - |
| 4 | g.8\_9delTC | 6 | 9 | GTTC | GT | 8 | 9 | TC | - | 7 | 9 | TC | - |
| 5 | g.10\_11insT | 10 | 10 | A | AT | 10 | 11 | - | T | 10 | 10 | - | T |
| 6 | g.12\_13insTAA | 12 | 12 | G | GTAA | 12 | 13 | - | TAA | 12 | 12 | - | TAA |
| 7 | g.16\_17insATT | 13 | 15 | GTC | GTCATT | 15 | 16 | - | ATT | 15 | 15 | - | ATT |
| 8a | g.18\_19delCA,g.18\_19[2] | 17 | 19 | TCA | T | 18 | 19 | CA | - | 17 | 19 | CA | - |
| 8b | g.20\_21delCA,g.20\_21[2] | 17 | 21 | TCACA | TCA | 20 | 21 | CA | - | 19 | 21 | CA | - |
| 8c | g.22\_23delCA,g.22\_23[2]\* | 17 | 23 | TCACACA | TCACA | 22 | 23 | CA | - | 21 | 23 | CA | - |
| 9 | g.24G= | 24 | 24 | G | G | 24 | 24 | G | G | 23 | 24 | G | G |
| 10a | g.25\_26insC,g.26dupC | 25 | 26 | A | AC | 25 | 26 | - | C | 25 | 25 | - | C |
| 10b | g.26\_27insC,g.27dupC | 26 | 27 | C | CC | 26 | 27 | - | C | 26 | 26 | - | C |
| 10c | g.27\_28insC,g.28dupC | 27 | 28 | C | CC | 27 | 28 | - | C | 27 | 27 | - | C |
| 10d | g.27[3]\*,g.26[3],g.25[3] | 27 | 27 | C | CC | 27 | 28 | - | C | 27 | 27 | - | C |

\* These are the HGVS recommended representation for the canonically equivalent representations of item 8 and 10, respectively. These representations may appear in practice, but should be canonicalized so that they are seen as the same. HGVS recommends a right-justified representation and VCF recommends a left-justified representation, but neither is guaranteed. Alignment format

* pro: used by VCF, understood by many, supported and validated by vcf specification authors (Global Alliance).
* pro: proactively including the upstream allele would make it easier to reconstruct the VCF file format. NOTE: implementers can choose to optionally store the upstream allele when they store the information to overcome this issue.
* con: the reference allele includes bases that we’re not interested in and potentially cause ambiguity in isolating the precise region that has changed.
* con: deletions occurring at the same ref allele as a substitution will have different reference coordinates
  + g.100A>T, g.100delA (assuming the upstream allele is a G) you would have 100,100,A,T & 99,100,GA,G, respectively.
* con: since the actual location that is impacted must include upstream alleles it conflates the ability to place a type on the region represented by the actual location that is changed. For example, suppose the first 2 bases of an intron are deleted. Ideally the region would be classified, as “splice site”, but our reference coordinate would include both the intronic bases plus one exonic base thus impacting the ability to clearly identify the region type.
* con: special rules/considerations need to be made for regions that occur at the first and/or last base pair for a given reference sequence.

Comments

(chris bison 2/17/15 github issue #22) Note that neither problem is unfixable or unhandleable in the alignment format, but they point, in my mind to a somewhat un-natural way of describing these coordinates: one that sort of requires you to know what the allele is going to be before you define the place where it occurs.

Variant format

* pro: similar (if not identical to) the GVF format, which is also a well adopted standard, however, less so than VCF.
* pro: a very close resemblance to the HGVS genomic numbering scheme. So it would be reasonably easy to go back and forth between genomic HGVS nomenclature and this format. This does not include dupes or repreats only subs, ins, dels.
* con: most HGVS sequence variants are represented using cDNA which would not allow the smooth back and forth translation.
* con: the numbering scheme requires an understanding of the ref allele in order to determine if the position represents an insertion point or a change to the current reference. (see comments)
* con: special rules/considerations need to be made for regions that occur at the first and/or last base pair for a given reference sequence.
* con: does not include the upstream allele which would make it easier to reconstruct the VCF file format. NOTE: implementers can choose to optionally store the upstream allele when they store the information to overcome this issue.
* pro: the most specific region representing the allele is represented allowing for a clear and consistent region classification approach.

Comments

(chris bison 2/18/2015 github issue #22) I think (?) that the GVF format is identical to what we were calling the "Variant Format" in our earlier discussion. The drawback of this method compared to the interval method (or inclusive-exclusive) is that the ref and end are not interchangeable. You really need both to distinguish between talking about a single-base locus and talking about the insertion point just before the base.

Interval format

* con: this is not currently promoted by any variant calling file format standard. Promoting and gaining adoption could be a challenge since it deviates from the current standards.
* pro: the coordinate clearly indicates whether the allele is replacing an existing base or bases vs. inserting at a given position in the sequence. (Unambiguous coordinate specification, independent of ref or alt allele values).
* con: does not include the upstream allele which would make it easier to reconstruct the VCF file format. NOTE: implementers can choose to optionally store the upstream allele when they store the information to overcome this issue.
* pro: simple translation between this format and genomic HGVS nomenclature for substitutions, inserts and deletions.
* pro: much more direct ability to compare coordinates (for the same reference sequence) to determine if they are equivalent, overlapping, intersecting, etc.
* pro: the most specific region representing the allele is represented allowing for a clear and consistent region classification approach.

Comments

This seems to be the best technical approach that aligns with the objectives of the data modeling work groups allele model. Implementers may choose to store the coordinates in any of the formats, but, since the allele model must communicate region types, it would be helpful to have an approach that provides a consistent approach for defining a reference coordinate irrespective of the reference and alternate allele information.