Proposed extension of the HGVS nomenclature regarding complex changes

Peter E.M. Taschner
Department of Human Genetics
Center for Human and Clinical Genetics
Leiden University Medical Center

As recommended during the Human Variome Project meeting in Paris, this proposal will also be posted at the Human Genome Variation Society website (http://www.hgvs.org) to start the discussion.
Please send comments and suggestions to: P.Taschner@lumc.nl

Current HGVS recommendation:

Complex changes

> *Sequence changes can be very complex, involving several changes at a specific location. The description of such changes using the recommendations given above can become rather complicated and at some point, although literally correct, effectively meaningless. In such cases the recommendation is to submit the sequence that has been determined to GenBank and to use the accession.version number in the description.*

- *c.123_4567conNM_004006.1:c.123_678 describes a gene conversion replacing nucleotides c.123 to c.4567 of the coding DNA sequence of the transcript of interest with nucleotides c.123 to c.678 from a transcript sequence as present in GenBank file NM_004006 (version 1)*
- *c.88+101_oGJB2:c.355-1045del denotes a deletion which ends in the flanking GJB2 gene at position 355-1045 (in the intron between nucleotides 354 and 355) on the reverse strand (the genes are thus located and fused in opposite transcriptional directions, see Discussion)*
- *c.123+54_123+55insAB012345.2:g.76_420 denotes an intronic insertion (between nucleotides c.123+54 and 123+55) of 345 nucleotides (nucleotides 76 to 420 like in GenBank file AB012345 version 2)*

The proposed extension aims to support descriptions of complex changes, which might be considered as "imperfect" duplications, gene conversions, inversions and insertions. According to the current HGVS recommendations, a single nucleotide substitution in the gene conversion and intronic insertion examples above would have to be described using the accession.version number of the sequence submitted to GenBank.

Please note that the purpose of the extended description is:
1) to reduce the necessity to submit new Genbank sequences for every small change observed within a larger change
2) to allow easy comparison between different complex changes.

3) to provide a description format, which can be generated and interpreted by dedicated software tools, such as Mutalyzer (http://www.mutalyzer.nl).

The extended description should not be interpreted as an evolutionary sequence of events, i.e. the molecular mechanism leading to the observed complex variant. It is simply a description to support automatic conversion of any reference sequence into the sequence, which was observed by the submitter.

Symbols used in the extension

- { } = encloses “sub-alleles”, one or several changes within insertions, gene conversions, inversions and duplications.
    - c.76_234inv{80A>C;83G>C} two changes within the inversion
    - c.76_234dup{80A>C;93_94insG} two changes within the duplication.
    - c.76_234dup{inv} a duplication in the opposite orientation, i.e. an inversion of the second duplicon

Position numbering within sub-alleles

Position numbers within sub-alleles are relative to the reference sequence in its original orientation and should not exceed the range of the basic change. If the basic change is an insertion or gene conversion described using an additional reference sequence accession number, the position numbers within the sub-allele refer to that reference sequence in its original orientation. The position numbers within the sub-allele are not affected by inversion of the inserted sequence indicated by the prefix “o”.

Position numbers within sub-alleles follow the numbering scheme of the allele without repetition of the prefix (g., etc.).
Applications:
3) the combination of the nested change format and the composite change format

1 Nested change format

The nested change format supports description of changes within duplications, gene conversions, inversions and insertions.

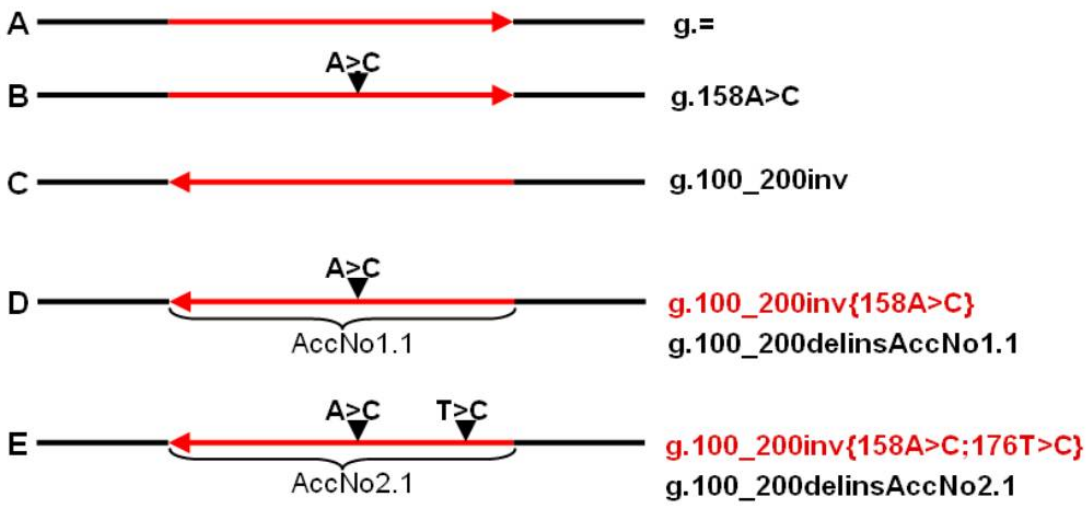

Fig.1. The nested change format reflects the similarity between inversions containing different SNP alleles.
Complex variants of a genomic region (red arrow)(A) with a SNP (B) undergoing perfect (C) or imperfect inversion (D, E) are described using the nested change format (red) and current HGVS nomenclature (black). Current HGVS nomenclature requires the accession numbers AccNo1.1 and AccNo2.1 obtained after submission of the sequences, which might be the result of imperfect inversions.
Please note that the substitutions in D and E are represented relative to the sequence in its original orientation, whereas the inversion will result in the insertion of its reverse complement.

- *D: g.100_200inv{158A>C} denotes inversion of sequence from nucleotide g.100 to nucleotide g.200 with a substitution of nucleotide A at position 158 of the reference sequence in its original orientation by nucleotide C)*
- *E: g.100_200inv{158A>C;176T>C} denotes inversion of sequence from nucleotide g.100 to nucleotide g.200 with a substitution of nucleotide A at position 158 of the reference sequence in its original orientation by nucleotide C and a second substitution of nucleotide T at position 176 of the reference sequence in its original orientation by nucleotide C)*

<u>2 Composite change format:</u>
The composite change format supports concatenation of inserted sequences

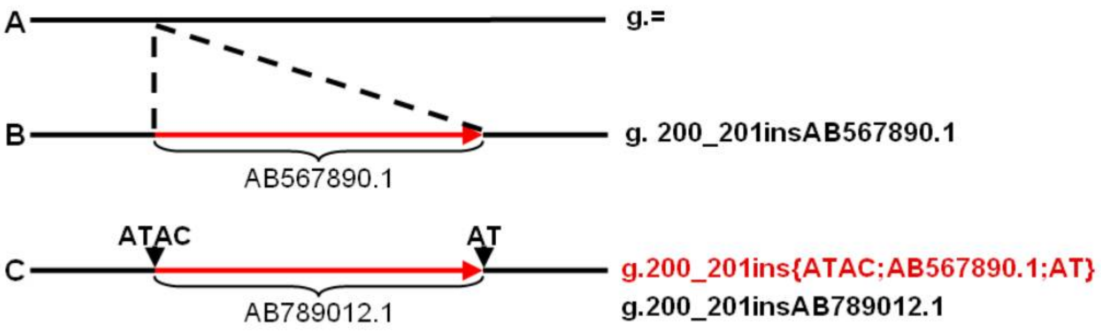

Figure 2. The similarity between insertions is reflected in the composite change format. Complex variants of reference sequence (A) containing insertion of similar sequences (red arrows) without (B) or with additional flanking sequences (C) are described using composite change (red) and current HGVS format (black).

- *C: g.200_201ins{ATAC;AB567890.1;AT} denotes an insertion (between nucleotides g.200 and 201) of nucleotides ATAC, followed by the nucleotides contained in GenBank file AB012345 version 2, followed by nucleotides AT.*

3 The combination of the nested change format and the composite change format

The nested and composite change format can be combined for maximal flexibility to describe changes, which might result from duplication in combination inversions.

Describing changes in duplicon orientation:

Duplications result in a tandem repeat of the same sequence, which can be represented by an arrow (→) for the original or duplicated unit (duplicon). Nesting and concatenation can be used to describe the inversion of one or both duplicons:

dup	(→→)
dup{inv}	(→←)
inv;dup	(←→)
inv;dup{inv}	(←←)

The dup{inv} (→←, head-to-head) and inv;dup (←→, tail-to-tail) formats support descriptions of the current ISCN inverted duplications at molecular level.

- *g. 200_3000dup{inv{258G>C}} denotes an inversion of a copy of the second duplication unit from position 200 until position 3000 of the reference sequence with a G>C substitution at position 258.*

- *g.[221del;200_3000dup{inv{258G>C}}] denotes a single nucleotide deletion at position 221 in the first duplicon and an inversion of a copy of the second unit from position 200 until position 3000 of the reference sequence with a G>C substitution at position 258*

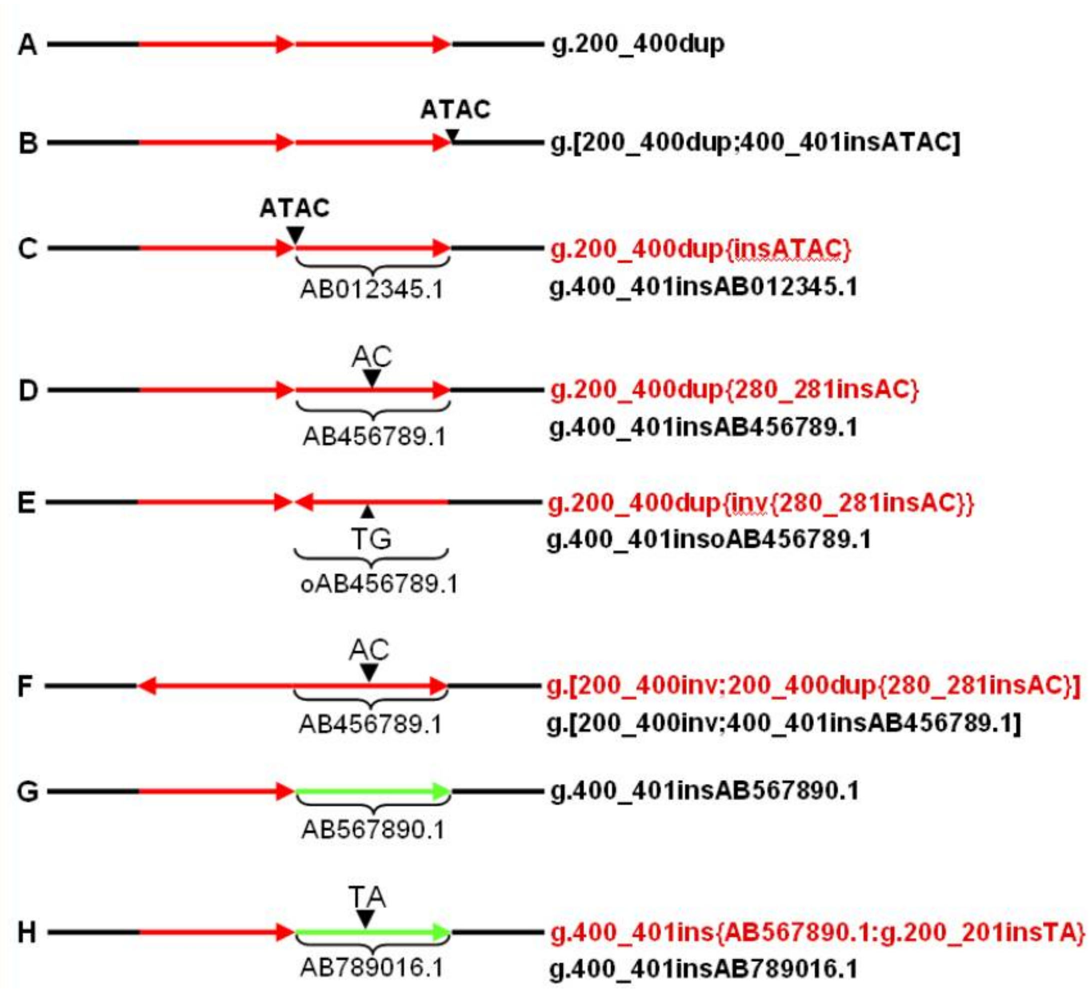

Figure 3. Extended HGVS complex variant descriptions are more informative than conventional ones. A genomic region (red arrow) has undergone perfect (A) or imperfect (B-D) duplication or inversion (E, F). Insertions of unrelated sequences (green in G, H) result in changes of similar size (G, H). In contrast to conventional descriptions (black), the extended descriptions (red) clearly show the similarity between the different variants. These extensions reduce the need to request a new Genbank Accession number (shown below the relevant sequences) for relatively small sequence changes.

- *A: g.200_400dup denotes a duplication of the sequence from nucleotide g.200 to nucleotide g.400.*

- *B: g.[200_400dup;400_401insATAC] denotes a duplication of the sequence from nucleotide g.200 to nucleotide g.400 followed by an insertion of nucleotides ATAC.*
- *C: g.200_400dup{insATAC} denotes an insertion of nucleotides ATAC between position 400 and 401 of the reference sequence followed by a duplication of the sequence from nucleotide g.200 to nucleotide g.400 in its original orientation. Please note that for insertions at the beginning of the second duplicon flanking positions do not preceed the ins variant type designator.*
- *D: g.200_400dup{280_281insAC} denotes a duplication from nucleotide g.200 to nucleotide g.400 of the reference sequence in its original orientation with an insertion of nucleotides AC between positions 280 and 281 in the second duplicon.*
- *E: g.200_400dup{inv{280_281insAC}} denotes an inversion of a copy of the second duplication unit from position 200 until position 400 of the reference sequence with an insertion of nucleotides AC between positions 280 and 281. Please note that the reverse complement of the inserted AC nucleotides is depicted.*
- *F: g.[200_400inv;200_400dup{280_281insAC}] denotes an inversion of the reference sequence from position 200 until position 400 followed by a duplication of the sequence from nucleotide g.200 to nucleotide g.400 in its original orientation with an insertion of nucleotides AC between positions 280 and 281.*
- *G: g.400_401insAB567890.1 denotes an insertion between position 400 and 401 of the reference sequence of the sequence contained in AB567890.1.*
- *H: g.400_401ins{AB567890.1:g.200_201insTA} denotes an insertion between position 400 and 401 of the reference sequence of the sequence in AB567890.1in its original orientation with an insertion of nucleotides TA between positions 200 and 201.*

Examples of nested and composite change descriptions for additional small changes in case of gene conversions and insertions:

- *c.123_4567conNM_004006.1:c.123_678{440G>C} describes a gene conversion replacing nucleotides c.123 to c.4567 of the coding DNA sequence of the transcript of interest with nucleotides c.123 to c.678 from a transcript sequence as present in GenBank file NM_004006 (version 1), but with an additional G>C substitution at position 440 of NM_004006.1.*
- *c.123+54_123+55insAB012345.2:g.76_420{110_115del} denotes an intronic insertion (between nucleotides c.123+54 and 123+55) of 339 nucleotides (nucleotides 76 to 109 and nucleotides 116 to 420 like in GenBank file AB012345 version 2)*
- *c.123+54_123+55ins{AT;AB012345.2:g.76_420;GC} denotes an intronic insertion (between nucleotides c.123+54 and 123+55) of 339 nucleotides (nucleotides AT followed by nucleotides 76 to 420 like in GenBank file AB012345 version 2 followed by nucleotides GC)*

- *c.123+54_123+55ins{AT;AB012345.2:g.76_420{110_115del};GC} denotes an intronic insertion (between nucleotides c.123+54 and 123+55) of 339 nucleotides (nucleotides AT, followed by nucleotides 76 to 109 and nucleotides 116 to 420 like in GenBank file AB012345 version 2, followed by nucleotides GC)*