SPECIAL ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Describing Structural Changes by Extending HGVS Sequence Variation Nomenclature

Peter E.M. Taschner* and Johan T. den Dunnen

Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, Nederland

For the HVP Bioinformatics Special Issue

Received 1 October 2010; accepted revised manuscript 6 December 2010.
Published online 15 March 2011 in Wiley Online Library (www.wiley.com/humanmutation). DOI 10.1002/humu.21427

ABSTRACT: New technologies allow rapid discovery of novel sequence variants among which those involving complex structural rearrangements. The description of such complex variants challenges the existing standard sequence variation nomenclature of the Human Genome Variation Society (HGVS, http://www.hgvs.org/mutnomen), because this mainly focuses on simple variants. Here, we suggest several extensions of the HGVS nomenclature guidelines to facilitate unambiguous description of complex sequence variants at the DNA level. These include: (1) nesting to support description of changes within inversions and duplications, and (2) composite changes to support concatenation of inserted sequences. The advantage of these additions is that inversions and duplications with small differences and more complex variants can be described without reverting to the less informative indel description. In addition, they should provide sufficient flexibility and consistency, thereby limiting alternative interpretations and ambiguous descriptions. The specifications should allow easy implementation in sequence variant nomenclature checkers (e.g., Mutalyzer, http://www.mutalyzer.nl/). We are extending the functionality of Mutalyzer to incorporate the latest version of the HGVS sequence variation nomenclature guidelines. Hum Mutat 32:507–511, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: sequence variation; complex mutation; mutation detection; mutation database; nomenclature; MDI

Introduction

The standard human sequence variation nomenclature (http://www.hgvs.org/mutnomen/) has gradually evolved as the result of continuous additions and changes [Antonarakis et al., 1998; Den Dunnen and Antonarakis, 2000]. The standard nomenclature was predominantly designed and used to describe sequence variants in gene sequence variant databases or in tabular format in publications. The updated standard nomenclature supports the need of the (clinical) genetics community in describing simple variants at different levels, including the noncoding DNA level. Large-scale rearrangements detectable by cytogenetic techniques traditionally belong to the genetic changes covered by the International System for Human Cytogenetics Nomenclature [Shaffer et al., 2009].

The application of array technology by the cytogenetics community and next generation sequencing technology in DNA diagnostics result in the increased detection of complex changes at relatively high resolution. These complex changes include exon duplications with additional variants in the duplicated exon, but also structural variants (SVs) such as large-scale rearrangements (segmental duplications, inversions, translocations, and transpositions) and copy number variations. The molecular nature of the data generated by the new techniques demands extensions of the standard ISCN and HGVS nomenclature systems. We believe that both systems will gradually approach each other and ultimately merge. As a first step, an extension of the HGVS standard nomenclature is proposed to allow accurate and unambiguous description of more complex changes including SVs. The aims of this article are: (1) to investigate the problems encountered when describing more complex variants, (2) to suggest a solution by extending the guidelines to provide sufficient flexibility and consistency for the description of complex sequence variants, (3) to provide a description format, which can be generated and interpreted by dedicated software tools, such as the Mutalyzer sequence variant nomenclature checker [Wildeman et al., 2008].

Current Standard Human Sequence Variation Nomenclature

The standard nomenclature for simple sequence variants, its basic rules and variant type preference have been summarized in Box 1 and Table 1. The standard nomenclature regarding complex sequence variants is recapitulated here to allow assessment of its limitations:

- Complex changes are sequence changes involving two or more changes occurring at the same location. The combination of changes may include substitutions, deletions, duplications, insertions, and inversions, which affect either a single nucleotide or a range of nucleotides, but also larger scale events (translocation, gene conversion, and transposition). When descriptions become too complex, the recommendation is to submit the sequence that has been determined to GenBank and to use the accession and version number in the description. For example, c.123+54_123+55insAB012345.2:g.76_420 denotes an insertion of nucleotides 76 to 420 from GenBank file AB012345 version 2 between nucleotides c.123+54 and 123+55 of the intron.
- Two or more changes in one individual (allele, haplotype, and genotype descriptions) are described by combining the changes, per allele (chromosome) between square brackets ("[]"). When

*Correspondence to: Peter E.M. Taschner, Center for Human and Clinical Genetics S-4-P, Leiden University Medical Center, Albinusdreef 2, P.O. Box 9600, 2300RC Leiden, Nederland. E-mail: P.Taschner@lumc.nl

Contract grant sponsor: European Community's Seventh Framework Program (FP7/2007-2013); Contract grant number: 200754.

changes occur at flanking positions, it is recommended to merge the separate descriptions. For example, the combination of c.76A>C and c.77T>G should not be described as c.[76A>C; 77T>G], but as c.76_77delinsCG.

Box 1. Summary of Current Basic and Extended HGVS Nomenclature Rules at the Genomic DNA Level

Current basic rules in a nutshell[a]
(1) Most 3′ position assigned to be changed. Example: g.5delT (not: g.4delT)
(2) Ranges of a reference sequence involved in deletions, duplications or inversions are indicated by their start and end positions separated by an underscore. Example: g.5_10del
(3) The location of an insertion is indicated by the consecutive positions of the flanking nucleotides separated by an underscore. Example: g.5_6insTA
(4) Single variants in multiple alleles are listed separately between square brackets, which are separated by a semicolon. Example: g.[5delT];[123A>G]
(5) Multiple variants in a single allele are listed between square brackets and separated by a semicolon. Example: g.[1A>T; 7del] (not: g.[7del;1A>T])
(6) Variants in a single allele are ordered from 5′ to 3′
Extended rules
(1) "Suballeles" using nested and composite change formats are preferred to describe changes within the range of duplications, inversions, insertions, and gene conversions
(2) Nucleotide numbers or ranges specifying the position of a variant in a suballele refer to the original reference sequence in its original orientation. They cannot exceed the range of the duplication, inversion, or gene conversion to which it belongs
(3) Variants in a suballele are listed between curly braces and separated by a semicolon
(4) When different levels of nesting are used, variant type hierarchy and order are evaluated from the deepest suballele level upward
(5) Sequences inserted in a suballele can be specified by a stretch of nucleotides and/or their corresponding accession number and version number separated by a semicolon and ordered from 5′ to 3′. Examples: g.5_25dup{11_12insAT}, g.5_2500dup{111_112insAB345678.1} or g.5_2500dup{111_112ins{GC;AB345678.1;AB456789.1;TAC}
(6) Insertions of sequences between 5′ and 3′ copies of a duplicated sequence (i.e., before the start of the 3′ copy) have no insertion position numbers in the suballele description. Example: g.5_25dup{insAT}
(7) Multiple variants in a suballele are ordered from 5′ to 3′

[a]See the HGVS guidelines at http://www.hgvs.org/mutnomen/ for the complete set of abbreviations and definitions.

Limitations of the Current Nomenclature

The description of complex variants using "delins" and a GenBank accession number is convenient in case of large changes where the inserted sequence is a perfect copy. However, duplications, gene conversions, and inversions often have additional small changes and can be regarded as "imperfect." According to the standard HGVS nomenclature, these "imperfect" rearrangements have to be described as deletion–insertions. This makes the descriptions long and complex. In addition, the similarity between the imperfect copies is not easily recognized (Fig. 1).

Furthermore, the standard HGVS nomenclature defines duplication as "a sequence change where a copy of one or more nucleotides is inserted directly 3′-flanking of the original copy." This means that the original sequence (5′ copy) is in the same orientation (i.e., head to tail) as the inserted copy (3′ copy). Therefore, the HGVS definition applies only to tandem duplications. The ISCN nomenclature describes these as direct duplications to distinguish them from inverted duplications, in which the order of the bands with respect to the centromere has changed. In cytogenetic terminology, either one of the copies involved may have been inserted in an orientation opposite to the original sequence.

Extensions and Additional Suggestions

To remove the limitations mentioned above, we suggest the following extensions (see Box 1 for a summary of the extended rules):

Nested Changes

We propose the description format using nesting to describe complex variants. This extension could be included in the "Symbols" section of the standard HGVS nomenclature as:

- Curly braces { and } are used to indicate "suballeles," containing one or several changes within insertions, gene conversions, inversions and duplications. The suballele is used

Table 1. Simple and Complex Variant Type Preference in HGVS Nomenclature

Top level	Basic level[a]	Criteria
Simple Variants of Reference Sequence: 5′-ATGTTAC-3′		
Substitution g.4T>C	Deletion–insertion (not: g.4delTinsC)	1-bp replacement
Duplication g.3_5dup	Insertion (not: g.3_5insGTT)	Inserted sequence (3′ copy) = immediately preceding sequence (5′prime; copy)
Inversion g.3_6inv	Deletion–insertion (not: g.3_4delGTTAinsTAAC)	Inserted sequence = reverse complement of deleted sequence
Complex variants		
Gene conversion g.100_2000conAB345678.1	Deletion–insertion (not: g.100_2000delinsAB345678.1)	Deleted sequence flanks homologous to flanks of inserted sequence, which is from another location in the genome
Duplication (3′ copy inversion) g.30_400dup{inv}	Insertion (not: g.30_400insoAB678901.1)[b]	Inserted sequence (3′ copy) = reverse complement of preceding sequence (5′ copy)
Duplication (with other change) g.30_400dup{125A>C}	Insertion (not: g.30_400insAB789012.1)	Inserted sequence (3′ copy) = preceding sequence (5′ copy) containing the suballele change
Insertion (with other change) g.100_200ins{ATAC;{AB567890.1:g.24_25insGC};AT}	Insertion	Sequence similar to AB567890.1 inserted at similar positions in the reference sequence. Concatenation used to describe additional flanking sequences in 5′ to 3′ order
Inversion (with other change) g.30_400inv{125A>C}	Deletion-insertion (not: g.30_400delinsAB456789.1)	Inverted sequence = reverse complement of deleted sequence containing the suballele change. Please note: inv{dup} not allowed for duplication in tail-to-head orientation; use: g.[30_400inv;30_400dup{inv}]
Gene conversion (with other change) g.100_2000conAB345678.1{125A>C}	Deletion-insertion (not:g.100_2000delinsAB234567.1	Deleted sequence flanks homologous to flanks of inserted sequence, which is from another location in the genome and contains the suballele change

[a]If the criteria are met, the basic level description shown below the variant type is replaced by that below the top level description.
[b]Current HGVS nomenclature uses the prefix "o" before the accession number to indicate that the reverse complement of the specified sequence is inserted.

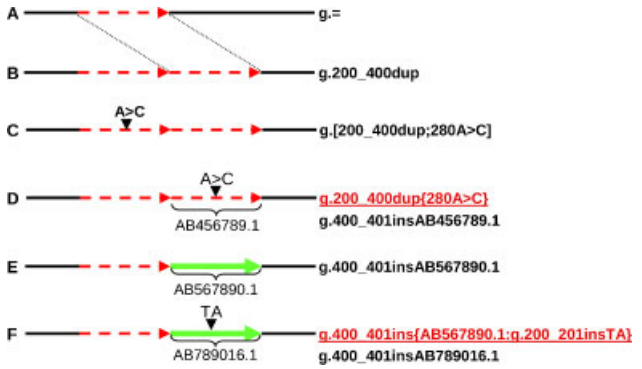

Figure 1. Nested change format reflects similarity between duplications containing different SNP alleles. A genomic region (dashed arrow) (Allele A) has undergone duplication (Allele B) in combination with substitution (Alleles C and D). The imperfect duplication (allele D) is described using the nested change format (underlined) and according to current HGVS nomenclature (not underlined). The presence of the same substitution in different copies is easily recognized (Alleles C and D). The nested change format helps to recognize the difference between the variants with 3′ copy changes and those with insertion of an unrelated sequence (Alleles E and F). Current HGVS nomenclature requires submission of the sequences involved and the use of the corresponding accession numbers to describe the complex variants (Alleles D–F). [Color figures can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

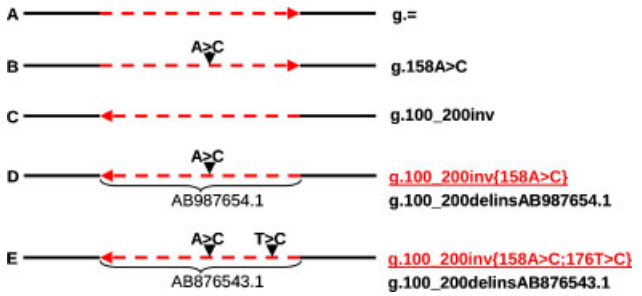

Figure 2. Nested change format reflects similarity between inversions containing different SNP alleles. In a genomic region (dashed arrow) (Allele A) containing a SNP (Allele B) and an inversion (Allele C), two complex variants (imperfect inversion alleles D and E) are described using the nested change format (underlined) and according to current HGVS nomenclature (not underlined). The presence of the same substitution in B, D, and E is easily recognized. Current HGVS nomenclature requires submission of the sequences involved and the use of the accession numbers AB987654.1 and AB876543.1 to describe the complex variants. Please note that the substitutions in alleles 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. [Color figures can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

as a suffix to the original change (e.g., format: dup{change}). The description g.200_400dup{280A>C}, for example, indicates that the 3′ copy of the sequence from nucleotide position 200 to nucleotide position 400 also contains an A to C substitution at position 280. The nested changes follow the numbering scheme of the main variant, unless g. or other position number prefixes are included between braces.

To demonstrate the use of the nested change description format, we will apply it to describe several theoretical changes in the copies of a region with a duplication (g.200_400dup) and a substitution (g.280A>C) (Fig. 1, alleles A–D, description underlined). The substitution in the 5′ copy can be described as g.[200_400dup;280A>C]. According to the standard nomenclature, the same substitution in the 3′ copy should be described after submission of the sequence involved to GenBank. Then, the corresponding accession and version number (AB456789.1) are used to describe the change as g.400_401insAB456789.1 (Fig. 1, allele D, description not underlined). For this allele, the nested change format uses the main change type dup followed by the substitution between braces, for example, g.200_400dup{280A>C}. The position of the nested change refers to the original position within the reference sequence, that is, before duplication. Due to the nested change format, "imperfect" 3′ copies are easily distinguished from insertions of nonhomologous sequences in Figure 1. The similarity between alleles E and F can also be described by adding the change within inserted sequence as a "suballele" to the accession numbers of the inserted sequence. In the description g.400_401ins{AB567890.1:g.200_201insTA} for allele F, the accession number of the inserted sequence is contained within the curly braces. As a consequence, the suballele description of inserted sequences can be evaluated according to the standard nomenclature. Other complex variants, such as changes within sequences involved in inversions, gene conversions or other exchanges, can be described by adding these changes as "suballeles" to the accession numbers of the exchanged sequences.

Application of the nested change format to theoretical "imperfect" inversions in a region with a substitution (g.158A>C) and an inversion (g. 100_200inv) is also easy (Fig. 2). Consider the deletion–insertion allele g.100_200delinsAB987654.1 (Fig. 2, allele D, description not underlined), which becomes g.100_200inv{158A>C} in nested change format. The position of the nested change refers to the original position within the reference sequence, that is, before inversion. In addition, the nucleotide specified is also the original one, that is, not its reverse complement in case of inversions. If the inverted region contains an additional substitution (g.176T>C), the standard allele description rules can be used to create "suballeles" containing the variants separated by a semicolon, for instance, g.100_200inv{158A>C;176T>C} (Fig. 2, allele E, description underlined).

The nested change format is also applicable to "imperfect" duplications with changes in the orientation of one or both copies (Fig. 3). For clarity, in this text (but not in the nomenclature) the orientation of the copies relative to the reference sequence is represented by arrows. The → arrow indicates a copy in the original orientation with the sequence specified by the start and end positions; the ← arrow indicates a copy in the opposite orientation with the reverse complement of the sequence specified by the start and end positions. Perfect duplications have copies in the same head-to-tail orientation (→→) without any change (g.200_400dup, Fig. 3, allele A). A small insertion following the 5′ copy can be described as an allele: g.[200_400dup;400_401insATAC] (Fig. 3, allele B). The same insertion between the copies can be described as duplication with nested insertion: g.200_400dup{insATAC} (Fig. 3, allele C). In this case, the nested insertion has no position number to indicate its insertion before the 3′ copy. A small insertion within the 3′ copy also can be described as duplication with nested insertion: g.200_400dup{280_281insAC} (Fig. 3, allele D). Standard HGVS nomenclature can describe an inversion where the 5′ copy is inverted (tail-to-tail orientation: ←→) in an allele format combining inversion and duplication: [inv;dup]. The other copy orientations, head-to-head (→←) and tail-to-head (←←) can be regarded as duplications with nested inversions, which are easily described as dup{inv} and inv;dup{inv}, respectively. Thus, inversion of the 3′ copy of allele D could be described in the format dup{inv} with two levels of nesting: g.200_400dup{inv{280_281insAC}} (Fig. 3, allele E).

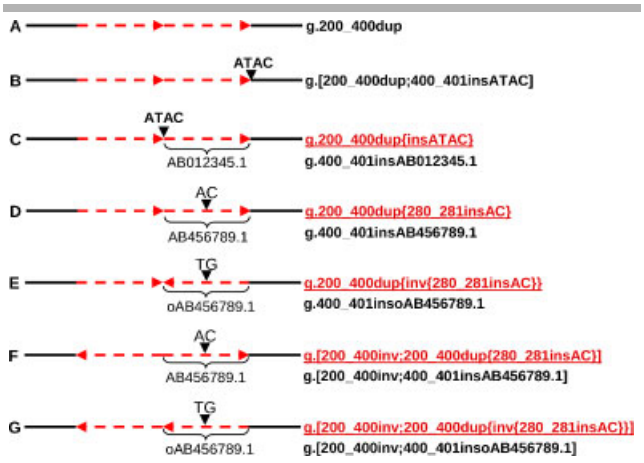

Figure 3. Complex variants involving inversion and duplication are more informative in nested change format. In a genomic region with a duplicated sequence (dashed arrows) (Allele A), variants with insertions near or in the 3′ copy (Alleles B–D), or inversions of one of the copies are observed (Alleles E–G). The nested change format (underlined) clearly shows the difference between the duplications with changes of the 5′ and 3′ copies (Alleles B–G). Current HGVS nomenclature requires submission of the sequences involved and use of the corresponding accession numbers to describe all complex variants, except allele B (not underlined). Current HGVS nomenclature uses the prefix "o" before the accession number to indicate that the reverse complement of the specified sequence is inserted (Alleles E and F). [Color figures can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

The inversion of the 5′ copy of allele D uses the format inv;dup with one level of nesting: g.[200_400inv;200_400dup{280_281insAC}] (Fig. 3, allele F). The inversion of both copies of allele D (or the 5′ copy of allele E) is described in the format inv;dup{inv}with two levels of nesting: g.[200_400inv;200_400dup{inv{280_281insAC}}] (Fig. 3, allele G).

Composite Changes

We propose the composite change description format to describe different insertions of similar sequences. This format uses the "suballele" braces of the nested change format to enclose a description of the inserted sequences, which are separated by a semicolon (;) in their order of appearance from 5′ to 3′ (e.g., format: ins{sequence; accession number.version number; sequence}, g.200_201ins{ATAC; AB567890.1;AT}. The composite change description format allows a combination of accession numbers and IUB nucleotide codes.

Consider a region in which an insertion of the sequence AB567890.1 was found (Fig. 4, alleles A and B). Occasionally, similar insertions are observed, but more nucleotides are inserted than those contained within the AB567890.1 record (Fig. 4, allele C). Allele C is described as g.200_201ins{ATAC;AB567890.1;AT} in composite change format. For optimal flexibility, the composite change format can be combined with nesting to describe the result of mutation events involving sequences of known and unknown origin. Thus, the insertion of a sequence, which differs from AB789012.1 by the additional insertion of nucleotides GC between positions 24 and 25 in AB567890.1, can be described as g.100_200ins{ATAC; {AB567890.1:g.24_25insGC};AT} (Fig. 4, allele D).

Variant Type Preference and the Order of Descriptions

The standard nomenclature uses different variant types to achieve the shortest description of a specific change. The standard

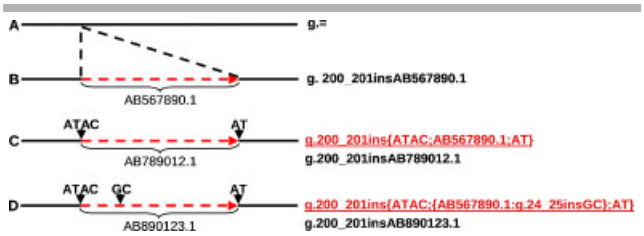

Figure 4. Composite change format reveals the similarity between different insertions at the same location. In genomic region A, insertions of similar sequences (dashed arrow), which only differ by a few flanking nucleotides are observed (Alleles B–D). The composite change format (underlined) captures both the similarity and the difference between the inserted sequences by concatenation of individual nucleotide sequences and accession number AB567890.1. The GC insertion within the AB567890.1 sequence is described using nesting (Allele D). Current HGVS nomenclature requires submission of the sequences involved and use of the corresponding accession numbers (not underlined). [Color figures can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

nomenclature contains a hierarchy to indicate which variant type is preferred for a given change (Table 1). For inserted sequences resulting from a duplication of the immediately preceding sequence (tandem duplication), the variant type duplication is preferred over insertion. Similarly, three other variant types, substitution, inversion, and gene conversion, are preferred above the insertion/deletion (delins) type. New in this extension is that inv;dup{inv} should be used to indicate a duplication in tail-to-head (← ←) orientation.

The variant type preference should be applied within the "suballeles" in the nested or composite change formats introduced here. For the unambiguous reconstruction of the variant sequence, it is important to evaluate the descriptions in the correct order (see Table 2). In general, variants in "suballeles" should be ordered from 5′ to 3′ similar to allele descriptions. In case of descriptions with different levels of nesting, variant type preferences and order have to be applied first on the deepest suballele level (within the most inner braces). For example, working upward in g.200_400dup{280_281insAC}, allows unambiguous recognition of the 3′ copy as the target location of the AC insertion. In addition, effects on the size of the copy are more easily calculated. In composite change formats, descriptions separated by semicolons should be treated independently in the order from left to right. For example, g.100_200ins{ATAC; {AB567890.1:g.24_25insGC};AT} is not equal to g.100_200ins{AT; {AB567890.1:g.24_25insGC};ATAC}, although the different parts of the inserted sequence are the same.

Discussion

The nomenclature extensions proposed here have been available for comments since June 2010 on the HGVS sequence variation nomenclature Website (http://www.hgvs.org/mutnomen), but no suggestions for modifications have been received. The increased flexibility of the nomenclature extensions will allow simple and unambiguous descriptions of closely related complex variants. The new formats support descriptions at the nucleotide level for duplications with nested inversions, which correspond to the large scale rearrangements known by cytogeneticists as inverted duplications [ISCN2009, p. 69]. For whole genome resequencing projects, the extensions might minimize the submission of sequences involved in SVs as separate entities to GenBank or other primary sequence repositories. The extensions will allow comparable descriptions of similar repetitive elements at different locations.

The extended description by itself should not be interpreted as an evolutionary sequence of events, that is, the molecular mechanism leading to the observed complex variant. It should be regarded simply as a description to support automatic conversion of any reference sequence into the sequence observed by the submitter. Similarly, the new formats could be used to describe somatic changes during tumor progression and to follow the evolution of cancer cell genomes. The same reference sequence file would also suffice for other genomic rearrangements, such as gene conversions and translocations leading to gene fusions in leukemia. Furthermore, the use of the same reference file may help to identify variants resulting from retrotransposition of reverse transcribed transcripts in LSDBs or whole genome resequencing data.

The new formats can be expected to remove the limitations of the current version of the standard nomenclature mentioned above. The increased flexibility will, however, add a new layer of complexity to the standard nomenclature. Evaluation of a combination of nested and composite changes will be relatively straightforward and should allow unambiguous reconstruction of the complete complex variant sequence. We have implemented the standard nomenclature in a software tool, the Mutalyzer nomenclature checker [Wildeman et al., 2008]. The nomenclature extensions suggested here are planned to be included in a future version of Mutalyzer and should not be difficult to incorporate in other software with similar functionality.

Acknowledgments

The authors thank Jeroen Laros for stimulating discussions, Emmelien Aten for critically reading the manuscript, Ivo Fokkema for help with the figures, and the participants of the Human Variome Project Paris 2010 meeting for suggestions. Part of the research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 200754, the GEN2PHEN project.

References

Antonarakis SE, The Nomenclature Working Group. 1998. Recommendations for a nomenclature system for human gene mutations. Hum Mutat 11:1–3.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Shaffer LG, Slovak ML, Campbell LJ, editors. 2009. ISCN 2009: an international system for human cytogenetics nomenclature. Basel: Karger.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variation descriptions in locus-specific mutation databases and the literature using the MUTation AnaLYZER (MUTALYZER) mutation nomenclature checker. Hum Mutat 29:6–13.