Muddled genetic terms miss and mess the message.

Vihinen, Mauno

Published in:
Trends in Genetics

DOI:
10.1016/j.tig.2015.05.008

2015

Muddled genetic terms miss and mess the message

Mauno Vihinen

Department of Experimental Medical Science, Lund University, BMC D10, SE-22184 Lund, Sweden

A critical aspect of science is the clear communication of complicated matters. However, language is often ambiguous, and the message can get lost in the telling. In particular, genetic terms can have different meanings for different people. Here, I discuss this problem and suggest remedies to clarify the message.

Keywords

genetic terms; systematics; muddled language; clarity of presentation; HGNC; HGVS nomenclature; Variation Ontology; Human Variome Project

Why is it so difficult to write unambiguous genetic texts?

Genetics deals with complicated concepts; thus, it is important to communicate clearly in a way that everyone can understand and interpret in only one way. This ideal is not always achieved, partly because many genetic concepts have somewhat different meanings for different people. Take the term ‘mutation’: a brief search provided various definitions, including a nucleotide change, a genetic change leading to a phenotype different from the parental one, and a permanent change in an organism. Even when considering the context, it is not always possible to be sure about the meaning. Another level of complication emerges from the utilization of nonstandard systems. Examples are the legacy numbering in globin and collagen proteins, which are inconsistent with nomenclature based on gene and protein sequences, causing confusion. Those working in the field may be familiar with such conventions, but nonexperts can easily get lost.

Several projects and organizations aim to unify the systematics in genetics, including gene names, reference sequences, variation nomenclature, variation effects, data exchange standards, and ontologies (Table 1). For example, the Human Variome Project (HVP) has a systematic procedure for the establishment and acceptance of standards and guidelines [1].

Gene and protein names

It is not uncommon when reading a paper to be unsure about which gene or protein is being discussed. Either the whole name is not given or one of the less-utilized alternative names is used. Many genes have numerous names and the same name can refer to completely different entities.

Systematic names generated by the HUGO Gene Nomenclature Committee (HGNC) solve this problem. There are some 39 000 names for human, mainly protein-coding genes but also pseudogenes, noncoding RNAs, phenotypes, and genomic features. In addition, HGNC provides tools for identifying corresponding genes in other organisms.

Mutation versus variation

The most common problematic term in genetic literature is ‘mutation’, which has two widely used meanings: a process generating variation and an outcome of the process. The latter often has negative connotations. Moreover, genetic changes lead to a range of phenotypic effects, which is not captured in the blanket term ‘mutation’. ‘Variation’ is a neutral term that has the same meaning for everybody, making it preferable to mutation. Another muddled concept is ‘polymorphism’, which is used to indicate a benign variation with a certain frequency, typically over 1%, in a population. However, these parameters are often not known, and it is difficult to verify that a variant is harmless in all situations. Thus, variation is also the recommended term for these situations.

There are several recommendations to systematically use the term ‘variation’ for the products of the mutation process. These include the Human Genome Variation Society (HGVS) nomenclature [2], articles [3] including the American College of Medical Genetics guidelines [4], and Variation Ontology (VariO) [5] (Box 1).

Variation in naming systematics

The HGVS nomenclature was developed to enable efficient and accurate reporting, curation, and search of variation data. The HGVS nomenclature allows description of variations at the DNA, RNA, or protein level, and works equally well for short and long alterations. HGVS names contain the variation position and the alteration, providing unambiguous mapping to reference sequence(s). Systematic HGVS names can also be generated with computational tools from sequence information 6 and 7.

HGVS names also offer the advantage that computers can correctly interpret them, but they can be difficult for human readers to comprehend. To address this, VariO translates systematic descriptions of variation type into short English terms for variations, such as DNA substitution or protein truncation [5] (Box 1). The VariOtator tool (http://variationontology.org/VariOtator.php) provides VariO variation-type annotations automatically from the sequence and variation details.

Reference sequences

Given that genetic changes are relative, a reference sequence is needed. However, most articles are published without these details. If a reference sequence is taken from a sequence database, a version number has to be provided to allow for unambiguous tracking. For human data, the preferred reference sequences are Locus Reference Genomic (LRG) sequences [8], which are available for

several genes and proteins. These sequences will never be changed and, thus, no version number is needed. If changes are necessary, a new LRG can be generated.

Utilization of LRGs makes the matching of data more reliable and efficient. In the current situation, for example, locus-specific variation database (LSDB) curators spend lots of time trying to match information in publications to sequence entries [9]. Numbering schemes may differ depending on the field as well as the utilized reference sequences, further confusing matters. LRGs are manually curated permanent reference sequences designed especially for reporting variants with clinical implications.

Missense, nonsense, and nonsynonymous variations

Many scientists use muddled and unclear terms when discussing variations and typically mix the molecular levels. This is especially common in the case of protein variations, probably due to many decades dominated by molecular biology findings and authors' limited knowledge of proteins. The most common misused terms are 'missense' and 'nonsense'. The 'sense' they refer to is the protein-coding triplet code of mRNA and, thus, these terms describe effects on mRNA not on protein. In the case of a 'missense' variation, the wrong amino acid is coded; however, when that change appears in the protein, it is called an amino acid substitution. The original variation is a DNA substitution, which is then passed on to the mRNA and protein by transcription and translation.

By contrast, 'nonsense' turns a coding codon into a stop codon. The outcome of these variations is either a truncated protein or no protein due to degradation of the message by mRNA surveillance mechanisms. The correct terminology for these variants, when discussing the final effect, is either protein truncation or missing protein (for details, see 5 and 10).

'Nonsynonymous' variations describe alterations at the RNA level and, thus, nonsynonymous single nucleotide variation (nsSNV) is a misnomer because SNVs (often erroneously called SNPs) describe DNA changes. Similarly, synonymous also refers to the RNA level and cannot be used to describe DNA changes.

Frameshift

'Frameshift' or 'out-of-frame' are also incorrectly applied to proteins. The shifted frame refers to the mRNA. Frameshift variation occurs when the size of the nucleotide sequence change (deletion, indel, or insertion) in the coding sequence is not divisible by three. This leads to an altered reading frame and a scrambled sequence. There is no frameshift in the protein structure. The correct term for these variations at the protein level is 'amphigoric', according to VariO (Box 1). An amphigoric insertion refers to the case where the sequence from the insertion point onwards is totally different than the reference. Similar to another RNA variation type, 'nonsense variation', frameshift deletions lead to protein truncations or even missing proteins if the deletion is sizable.

Indel

The term ‘indel’ is frequently used to mean insertion, deletion, or both collectively. However, the recommended definition of indel is a variation that comprises both inserted and deleted elements. At the protein level, indels, similar to insertions, can be of two types: sequence retaining or amphigoric, whereas at the nucleotide level, there are no subclasses. Note that amphigoric amino acid deletion does not exist.

Why bother?

One might ask whether these notes are just nitpicking and if readers can still understand the message. Although this is often probably true, it is not guaranteed and may lead to misinterpretations. In a test, experts agreed in only 78% of cases if a sentence was about DNA, mRNA, or protein [11]. This kind of distinction should be crystal clear. This percentage is likely to go down with more complex texts and descriptions of variations. In recent years, computer-aided methods in the domain of natural language processing (NLP) have increasingly been used to analyze and mine texts. Computers do not have the fuzzy processing capabilities of the human brain and cannot correctly interpret ambiguous, messy, or muddled text. Text mining is often the only possibility to screen large bodies of text. Our writing should be clear and concise for readers whether they have brains or processors. Poorly written prose cannot be automatically interpreted. Above all, though, the value of the work that we do as scientists can only be realized if we clearly communicate our message, both within and beyond the scientific community. Adopting common nomenclature for genetic terms is essential to sharing our findings, without which we will not realize the full potential of our research.

References

1 T.D. Smith, M. Vihinen. Standard development at the Human Variome Project Database (2015) http://dx.doi.org/10.1093/database/bav024

2 J.T. den Dunnen, S.E. Antonarakis. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum. Mutat., 15 (2000), pp. 7–12

3 G.R. Cutting. Annotating DNA variants is the next major goal for human genetics. Am. J. Hum. Genet., 94 (2014), pp. 5–10

4 S. Richards, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med., 17 (2015), pp. 405–423

5 M. Vihinen. Variation ontology for annotation of variation effects and mechanisms. Genome Res., 24 (2014), pp. 356–364

6 M. Wildeman, et al. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum. Mutat., 29 (2008), pp. 6–13

7 R.K. Hart, et al. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics, 31 (2015), pp. 268–270

8 R. Dalgleish, et al. Locus reference genomic sequences: an improved basis for describing human DNA variants. Genome Med., 2 (2010), p. 24

9 J. Celli, et al. Curating gene variant databases (LSDBs): toward a universal standard. Hum. Mutat., 33 (2012), pp. 291–297

10 M. Vihinen. Types and effects of protein variations. Hum. Genet., 134 (2015), pp. 405–421

11 V. Hatzivassiloglou, et al. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics, 17 (Suppl. 1) (2001), pp. S97–S106

Table 1. Resources for a unified genetic nomenclature

Organization or project	**Website**
Gene Ontology	http://geneontology.org/
HGNC	http://www.genenames.org/
HGVS	http://www.hgvs.org/
HVP	http://www.humanvariomeproject.org/
Global Alliance for Genomics and Health (GA4GH)	http://ga4gh.org/
LRG	http://www.lrg-sequence.org/
VariO	http://variationontology.org/
Sequence Ontology (SO)	http://www.sequenceontology.org/
VarioML data exchange format	http://www.varioml.org/
The Phenotype and Genotype Object Model (PAGE-OM)	http://www.omg.org/spec/PAGE-OM/

Box 1. Definition of concepts

Systematic definitions for some often-muddled terms as provided by VariO [5].

Amino acid substitution: substitution of an amino acid in protein.

Amphigoric amino acid indel: amino acid indel with sequence completely changed after the indel position; caused by an indel of a number of RNA nucleotides that is not divisible by three, leading to a frameshift of the mRNA reading frame.

Amphigoric amino acid insertion: insertion of one or more amino acids to a protein with the sequence completely changed after the insertion position; caused by insertion of a number of RNA nucleotides that is not divisible by three, leading to a frameshift of the mRNA reading frame.

DNA indel: complex DNA variation comprising both nucleotide insertion(s) and deletion(s).

Mutation: any process generating variation.

Missense variation: nucleotide change in the mRNA triplet codon encodes another amino acid.

Missing protein: variation preventing protein translation because of, for example, initiation codon variation or mRNA surveillance mechanism.

Missing RNA: variation preventing transcription, because of, for example, impaired transcription machinery or variation in the transcription regulation or start site.

Nonsense variation: nucleotide change in the mRNA codon triplet creates a terminator codon.

Nonsynonymous variation: mRNA variation leading to amino acid change in translated protein. The variant alters the codon so that it encodes another amino acid.

Out-of-frame deletion: deletion of nucleotide(s) causing alteration of mRNA coding sequence frame.

Out-of-frame indel: RNA indel containing a number of nucleotides that is not divisible by three and, therefore, causing a change in the mRNA reading frame.

Out-of-frame insertion: insertion of nucleotide(s) causing alteration of the mRNA reading frame.

Protein truncation: shortening of the protein sequence from the terminus.

Sequence retaining amino acid indel: amino acid indel without affecting the sequence after the indel.

Sequence retaining amino acid insertion: insertion of one or more amino acids to the protein without affecting the sequence after insertion.

Synonymous variation: mRNA variation not affecting the sequence of the translated protein.

Variant: a genetic character, organism, or individual with a difference to the reference state.

Variation: alteration in DNA, RNA, or protein.