

Guidelines for human gene nomenclature

Standardized gene naming is crucial for effective communication about genes, and as genomics becomes increasingly important in health care, the need for a consistent language to refer to human genes becomes ever more essential. Here, we present the current HUGO Gene Nomenclature Committee (HGNC) guidelines for naming not only protein-coding genes but also RNA genes and pseudogenes, and we outline the changes in approach and ethos that have resulted from the discoveries of the past few decades.

Elspeth A. Bruford, Bryony Braschi, Paul Denny, Tamsin E. M. Jones, Ruth L. Seal and Susan Tweedie

The first guidelines for human gene nomenclature were published in 1979 (ref. ¹), when the Human Gene Nomenclature Committee was originally established and given the authority to approve and implement standardized human gene symbols and names. In 1989, the Nomenclature Committee was placed under the auspices of the newly founded Human Genome Organization (HUGO) and thus became the HUGO Gene Nomenclature Committee (HGNC). Subsequent revisions to the nomenclature guidelines were published in 1987 (ref. ²), 1995 (ref. ³), 1997 (ref. ⁴) and 2002 (ref. ⁵). In the intervening years, the HGNC has published online updates to the guidelines to reflect the major changes and the increase in knowledge and data during this exciting period in human genomics. More than 40,000 human loci have been named by the HGNC to date; approximately half of these are protein-coding genes, and most resources now agree that the human genome contains around 19,000–20,000 protein-coding genes, a range considerably lower than some earlier estimates. Beyond the naming of protein-coding genes, substantial progress has been made in the nomenclature of different classes of RNA genes and pseudogenes. All approved human gene symbols can be found in the online HGNC database⁶ (<https://www.genenames.org/>).

The philosophy of the HGNC used to be that gene nomenclature should evolve with new technology and that symbol changes, if supported by most researchers working on a gene, would be considered if they reflected new functional information. Since the advent of clinical genomics, such changes have much wider impact, and disseminating nomenclature updates to all clinicians, patients, charities and other parties interested in genes is impossible. Therefore, the stability of gene symbols, particularly those associated with disease, is now a key priority for the HGNC. Nevertheless, novel information can be encapsulated in the gene name without changing the gene symbol.

Because human gene symbols are also routinely transferred to homologous vertebrate genes, including in our sister project, the Vertebrate Gene Nomenclature Committee (VGNC), we now avoid references to human-specific traits in nomenclature whenever possible.

We strongly advise researchers to contact us at hgnc@genenames.org whenever they are considering naming a novel gene or renaming an existing gene or group of genes, for all locus types, not just protein-coding genes. We are not always able to approve the gene symbols requested, but we strive to work with researchers to find acceptable alternatives. Requesting an approved symbol ensures that the published symbol will appear in biomedical databases, including our own and others. We further encourage journal editors and reviewers to check that approved nomenclature is being used and to require authors to contact the HGNC about any novel gene symbols before publication. Submitters should bear in mind that the HGNC is committed to making minimal future changes to gene symbols and that we do not take publication precedence into account when approving nomenclature.

Readers should note that the following are guidelines and recommendations (Box 1) but not strict rules. We are aware of numerous exceptional legacy symbols and names that remain approved. The HGNC considers the naming of each and every gene on a case-by-case basis, and deviations from these guidelines may be made given sufficient evidence that the nomenclature will ultimately aid in communication and data retrieval.

Gene naming

For many years, the HGNC has maintained the definition of a gene as “a DNA segment that contributes to phenotype/function. In the absence of demonstrated function, a gene may be characterized by sequence, transcription or homology.” Because there is still no universally agreed-upon alternative, we continue to use this definition.

Box 1 | Summary of the guidelines

1. Each gene is assigned a unique symbol, HGNC ID and descriptive name.
2. Symbols contain only uppercase Latin letters and Arabic numerals.
3. Symbols should not be the same as commonly used abbreviations.
4. Nomenclature should not contain references to any species or ‘G’ for gene.
5. Nomenclature should not be offensive or pejorative.

Ideally, gene symbols are short, memorable and pronounceable, and most gene names are long-form descriptions of the symbol. Names should be brief and specific, and should convey something about the character or function of the gene products but not attempt to describe everything known. Each gene is assigned only one symbol; the HGNC does not routinely name isoforms (that is, alternate transcripts or splice variants). Therefore, there are no separate symbols for protein-coding or non-coding RNA (ncRNA) isoforms of a protein-coding locus, or alternative transcripts from an ncRNA locus (Box 2).

When authors wish to use their own isoform notation, we advise stating clearly that this notation denotes an isoform of a particular gene and then quoting the HGNC symbol for that gene.

In exceptional circumstances, in response to community demand, separate symbols have been approved for gene segments in complex loci (for example, the *UGT1* locus, the clustered protocadherins at 5q31, and the immunoglobulin and T-cell-receptor families). Putative bicistronic loci may be assigned separate symbols to represent the distinct gene products (for example, *PYURF*, ‘PIGY upstream reading frame’, is encoded by the same transcript as *PIGY*, ‘phosphatidylinositol glycan anchor biosynthesis class Y’).

Box 2 | Cases for which HGNC does not provide official nomenclature

Sequence-variant nomenclature. This is the responsibility of the Human Genome Variation Society (HGVS)¹⁹, which provides recommendations for defining variations in DNA, RNA and protein sequences. HGVS endorses the use of HGNC gene symbols within their notation.

Products of gene translocations or fusions. We are not aware of official naming guidelines for these. SYMBOL1-SYMBOL2 is widely used, but we use this format for readthrough transcripts (as described in the main text), and hence we specifically recommend not using this format for translocations or fusions. We recommend the format SYMBOL1/SYMBOL2, which has been used in some publications (for example, *BCR/ABL1*).

Protein nomenclature. We have no authority over naming proteins, but we

coordinate closely with specialist groups that name specific subsets of proteins, such as the Enzyme Commission. The recently devised International Protein Nomenclature Guidelines (https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/) were written with the involvement of the HGNC. In agreement with these guidelines, we recommend that “protein and gene symbols should use the same abbreviation.” We further advise referring to proteins by using non-italic gene symbols, to distinguish them from genes.

Nomenclature for regulatory genomic elements. These elements include promoters, enhancers and transcription-factor-binding sites. We also do not provide nomenclature for transposable-element insertions in the human genome. Protein-coding and lncRNA genes that fit the criteria outlined in

Mayer et al.²⁰ may be named as endogenous retrovirus (ERV)-derived genes, but ERV insertions will not be named.

Nomenclature for human loci associated with clinical phenotypes and complex traits. Although HGNC historically named these loci, this activity has been taken over by Online Mendelian Inheritance in Man (OMIM)²¹. All HGNC entries with the locus type ‘phenotype only’ now have the status ‘entry withdrawn’. Note that some uncharacterized genes shown to be causative for a specific phenotype have adopted the phenotype symbol and name. If these phenotypic symbols have become entrenched in the literature, we aim to update the corresponding gene names to reflect an aspect of the normal function of the gene and its products (for example, *TSC1*, ‘tuberous sclerosis 1’ is now *TSC1*, ‘TSC complex subunit 1’).

Table 1 summarizes key factors considered when assigning gene nomenclature. Additionally, Supplementary Table 1 lists characters recommended for specific usage in gene symbols, Supplementary Table 2 highlights specific conventions used in gene names, and Supplementary Tables 3 and 4 provide Greek-to-Latin alphabet conversions and single-letter amino acid symbols, respectively.

Gene naming by biotype

Protein-coding genes. We aim to name protein-coding genes on the basis of a key normal function of the gene product. Many protein-coding genes of known function are named in collaboration with internationally recognized bodies composed of experts in a specific field. Where possible, related genes are named by using a common root symbol to enable grouping, typically on the basis of sequence homology, shared function or membership in protein complexes.

Gene group members should be designated by Arabic numerals placed immediately after the root symbol (for example, *KLF1*, *KLF2* and *KLF3*). More rarely, single-letter suffixes may be used (for example, *LDHA*, *LDHB* and *LDHC*). Some large gene families may include a variety of number and letter combinations to indicate subgroupings (for example, the cytochrome P450 superfamily members *CYP1A1*, *CYP21A2* and *CYP51A1*).

For genes involved in specific immune processes, or for those encoding an

Table 1 | Key factors in assigning gene nomenclature

Gene symbols	Gene names
Must be unique within a given genome	Should be brief and specific
Must not be offensive or pejorative (ideally in any language)	
Must not use superscripts or subscripts, or punctuation ^a	Should minimize punctuation; commas, hyphens and parentheses may be included for clarity ^b
Must contain only uppercase ^c Latin letters and Arabic numerals	Must be written in American English
Must start with a letter	Must start with a lowercase letter (unless starting with an eponym or capitalized abbreviation)
Should not include ‘G’ for gene, ‘H’ for human, Roman numerals or Greek letters	Should not include the words ‘gene’ or ‘human’
Should not spell proper names or common words or match commonly used abbreviations	Should start with the same letter as the symbol (to facilitate alphabetical listing and grouping)
Should avoid duplicating symbols in other species (unless orthologous)	Should not reference any species, taxa, tissue specificity, molecular weight, chromosomal location, human-specific features and phenotypes, or familial terms
Should avoid using letters or combinations of letters that have specific meanings when used in gene symbols (Supplementary Table 1)	Descriptive modifiers usually follow the main part of the name, to enable the use of a common root symbol for a gene group (for example, <i>ACADM</i> for ‘acyl-CoA dehydrogenase medium chain’ and <i>ACADS</i> for ‘acyl-CoA dehydrogenase short chain’).

^aSupplementary Table 2 shows punctuation exceptions for gene symbols. ^bPunctuation exceptions are made for enzyme names. ^cSole exception of C#orfs.

enzyme, receptor or ion channel, we consult with specialist nomenclature groups (Supplementary Note). For other major gene groups, we consult a panel of advisors when naming new members and discussing proposed nomenclature updates. A list of our specialist advisors is provided on our

website⁶, and we welcome suggestions of new experts for specific gene groups.

In the absence of functional data, protein-coding genes may be named in the following ways. (1) Naming may be based on recognized structural domains and motifs encoded by the gene (for example,

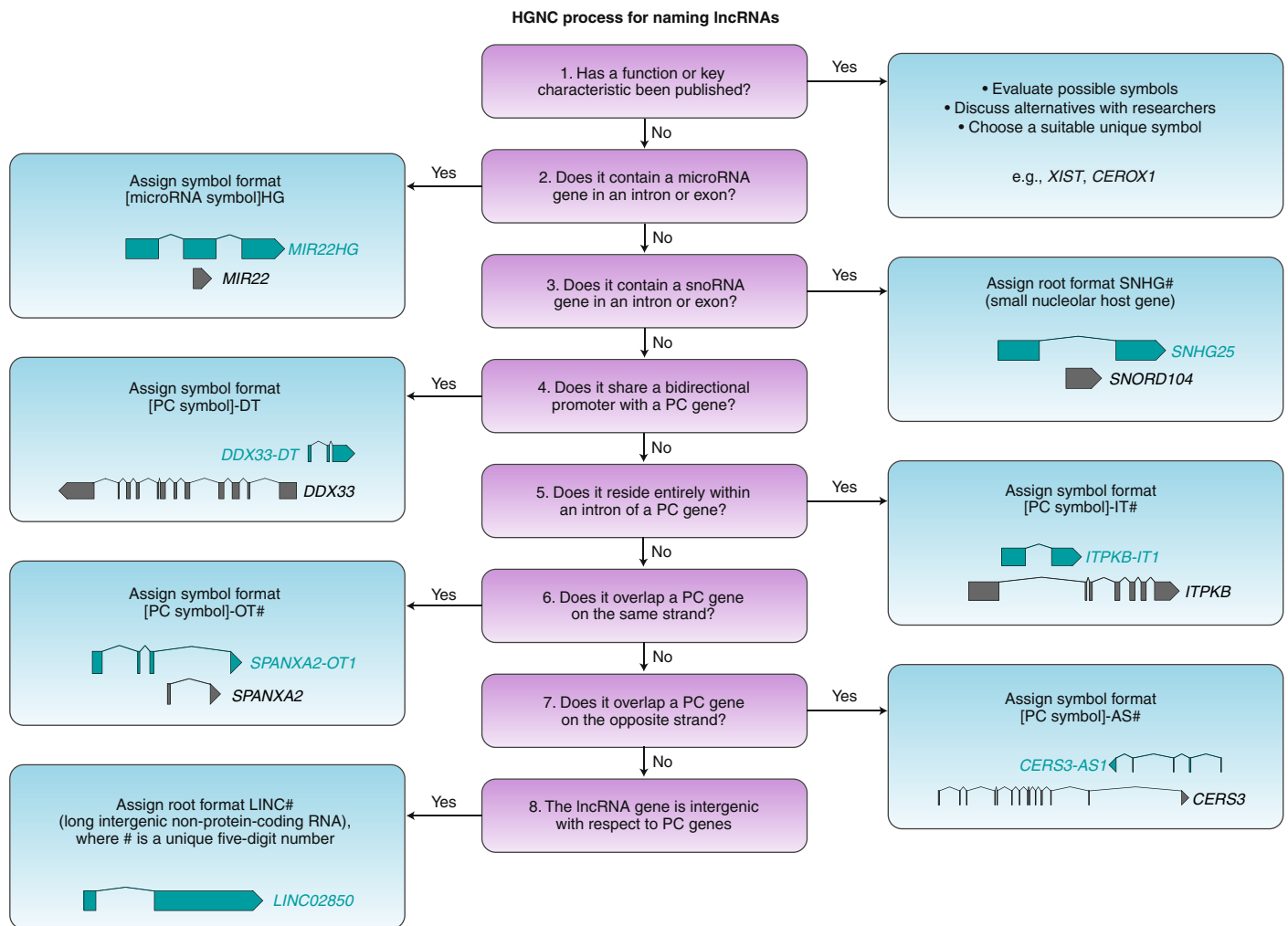


Fig. 1 | The HGNC's systematic process for naming lncRNA genes. In the absence of suitable published information, lncRNA genes are named on the basis of genomic context. HG, host gene; PC, protein coding; DT, divergent transcript (used for lncRNA genes that share a promoter with a PC gene); IT, intronic transcript; OT, overlapping transcript; AS, antisense RNA; LINC, long intergenic non-protein-coding RNA.

ABHD1, 'abhydrolase domain containing 1'; *HEATR1*, 'HEAT repeat containing 1'). Because these features can provide insight into the character of the gene product, this type of symbol is commonly retained even after the normal function of the gene product has been elucidated, although further information may be added to the gene name. (2) Naming may be based on homologous genes within the human genome. When naming is based on characterized homologs, genes of unknown function are given the next symbol within a designated series but with a different gene-name format (for example, *CASTOR3* is 'CASTOR family member 3' rather than 'cytosolic arginine sensor for mTORC1 subunit 3'). The placeholder root symbol FAM ('family with sequence similarity') is used when no information is available for any of the homologous genes.

Each homologous family has a unique FAM number (for example, FAM3), and each family member is distinguished by a letter or letter and number (for example, *FAM3A* or *FAM3C2P*). Of note, this root can be applied to both protein-coding and non-coding gene families. (3) Naming may be based on homologous genes from another species. If there is a one-to-one ortholog, the same or equivalent symbol will be approved (for example, human *CDC45*, 'cell division cycle 45', based on *S. cerevisiae CDC45*). A unique number or letter suffix is added if there is more than one human homolog (for example, *UNC45A* and *UNC45B* are co-orthologs of *Caenorhabditis elegans unc-45*). Gene names are updated to be appropriate for vertebrates (for example, 'unc-45 myosin chaperone' instead of 'UNCoordinated 45'). (4) Naming may be based only on the presence of an

open reading frame. Genes of unknown function that fit none of the above criteria are designated by the chromosome of origin, with the letters 'orf' for open reading frame (in lowercase to prevent confusion between the letter 'O' and the numeral '0', which may be part of the chromosome number) and a number in a series (for example, *C3orf18*, 'chromosome 3 open reading frame 18'). When the coding potential of the locus is in doubt, we include the word 'putative' in the name (for example, 'chromosome 18 putative open reading frame 15').

Historically, genes of unknown function identified by the Human cDNA project at the Kazusa DNA Research Institute⁷ have been named by using the KIAA# identifiers assigned by this project.

Pseudogenes. We define a pseudogene as a sequence that is incapable of producing a

functional protein product but has a high level of homology to a functional gene. In general, we name only pseudogenes that retain homology to a substantial proportion of the functional ancestral gene.

Most pseudogenes are processed and named after a specific parent gene (for example, *DPP3P1*, ‘DPP3 pseudogene 1’). Such pseudogene numbering is usually species specific, and hence orthology cannot be inferred from identical pseudogene symbols in different species.

Pseudogenes retaining most of the coding sequence as compared with other family members, which are usually unprocessed, are named as new family members with a ‘P’ suffix (for example, *CBWD4P*, ‘COBW domain containing 4, pseudogene’). This naming format is also used for genes that are pseudogenized relative to their functional ortholog in another species (for example, *ADAM24P*, ‘ADAM metallopeptidase domain 24, pseudogene’ is the pseudogenized ortholog of mouse *Adam24*). In rare cases, such pseudogenes do not include the ‘P’ if the symbol is well established (for example, *UOX*, ‘urate oxidase (pseudogene)’).

A small number of genes are currently pseudogenized in the reference genome but are known to have coding alleles segregating in the population. Such loci are given the locus type ‘protein-coding’ and are indicated by ‘(gene/pseudogene)’ included at the end of the gene name (for example, *CASP12*, ‘caspase 12 (gene/pseudogene)’).

Non-coding RNA genes. We name ncRNA genes according to their RNA type, as described in our recent review⁸. For small RNAs for which an expert resource exists, we follow the existing naming schema (for example, miRBase⁹ for microRNAs and the genomic tRNA database (GtRNAdb)¹⁰ for transfer RNAs). Other ncRNA classes, such as small nuclear RNAs, are named in collaboration with specialist advisors.

Long non-coding RNAs (lncRNAs), whenever possible, are named on the basis of a key function or characteristic of the encoded RNA. When functional information is not available, a systematic nomenclature is applied (Fig. 1).

Readthrough transcripts. Readthrough transcripts are normally produced from adjacent loci and include coding and/or non-coding parts of two (or more) genes. The HGNC names only readthrough transcripts that are consistently annotated by both the RefSeq annotators at the National Center for Biotechnology Information (NCBI)¹¹ and the GENCODE annotators at Ensembl¹². These transcripts have the

Box 3 | Scenarios that may merit a symbol change

- **Adoption of a more appropriate or commonly used alias.** For example, *RNASEN* was updated to *DROSHA* (drosha ribonuclease III) because of overwhelming community usage.
- **Domain- or motif-based nomenclature.** For example, *TMEM206* (transmembrane protein 206) is now *PACCI* (proton activated chloride channel 1).
- **Phenotype- or disease-based nomenclature.** For example, *CASC4* (cancer susceptibility candidate 4) was renamed *GOLM2* (golgi membrane protein 2), removing reference to the phenotype and making it consistent with its paralog *GOLM1*.
- **Location-based nomenclature.** For example, *TWISTNB* (TWIST neighbor) is now *POLR1F* (RNA polymerase I subunit F).
- **Pejorative symbols.** For example, *DOPEY1* was renamed to *DOP1A* (DOP1 leucine zipper like protein A).
- **Misleading or incorrect nomenclature.** For example, *OTX3* was initially erroneously named as an OTX family member and has been renamed *DMBX1* (diencephalon/mesencephalon homeobox 1).
- **Symbols that affect data handling and retrieval.** For example, all symbols that autoconverted to dates in Microsoft Excel have been changed (for example, *SEPT1* is now *SEPTIN1*; *MARCH1* is now *MARCHF1*); tRNA synthetase symbols that were also common words have been changed (for example, *WARS* is now *WARS1*; *CARS* is now *CARS1*).

locus type ‘readthrough transcript’ and are symbolized by using the two (or more) symbols from the parent genes, separated by a hyphen (for example, *INS-IGF2*) with ‘readthrough’ appended (for example, ‘INS-IGF2 readthrough’). The name may include additional information about the potential coding status of the transcript, such as ‘(NMD candidate)’.

Gene segments. For specific complex loci, the HGNC assigns symbols to individual gene segments, solely on the basis of community request. Examples include the immunoglobulins and T-cell receptors, the *UGT1* locus and clustered protocadherins.

Genomic regions. The HGNC previously named genomic regions referenced in the literature (for example, *XIC*, ‘X chromosome inactivation center’), and gene clusters were assigned symbols suffixed with the ‘@’ character (for example, *HOXA@*, ‘homeobox A cluster’). We no longer routinely provide symbols for genomic regions, but some symbols, such as those for fragile sites, have been retained when they have been used in publications, and this information would otherwise be lost.

Genes found within subsets of the population

Historically, the HGNC has approved symbols for only genes in the human reference genome. Rare exceptions have been made when requested by particular communities (for example, structural variants within the HLA and KIR gene

families, both of which have dedicated nomenclature committees). Future naming of structural variants will be restricted to those on alternate loci that have been incorporated into the human reference genome by the Genome Reference Consortium (GRC; <https://www.ncbi.nlm.nih.gov/grc>). The underscore character is reserved for genes annotated on alternate reference loci (for example, *GTF2H2C_2* is a second copy of *GTF2H2C* on a 5q13.2 alternate reference locus; *APOBEC3A_B* is a deletion hybrid on a 22q13 alternate reference locus that includes exons from both the *APOBEC3A* and *APOBEC3B* parent genes).

Status

All HGNC gene records have a status: the majority are ‘approved’, but when new evidence shows that a previously named gene is no longer considered to be real, the entry’s status changes to ‘entry withdrawn’. Whenever possible, we avoid reusing symbols from ‘entry withdrawn’ records, because doing so can cause considerable confusion.

Naming across vertebrates

We recommend that orthologous genes across vertebrate species (and, when appropriate, non-vertebrate species) have the same gene symbol.

The Vertebrate Gene Nomenclature Committee

The VGNC (<https://vertebrate.genenames.org/>) is an extension of the HGNC responsible for assigning standardized names to genes in vertebrate

species that currently lack a nomenclature committee. The VGNC coordinates with the five established existing vertebrate nomenclature committees—the Mouse Genomic Nomenclature Committee (MGNC)¹³, Rat Genome Nomenclature Committee (RGNC; <https://rgd.mcw.edu/nomen/nomen.shtml>), Chicken Gene Nomenclature Consortium (CGNC)¹⁴, *Xenopus* Nomenclature Committee (XNC)¹⁵ and Zebrafish Nomenclature Committee (ZNC)¹⁶—to ensure that vertebrate genes are named in line with their human homologs.

Orthologs of human C#orf# genes are assigned the human gene symbol, with the other species chromosome number as a prefix and an ‘H’ denoting human. Therefore, because the ortholog of human *C1orf100* is on cow chromosome 16, the cow gene symbol is *C16H1orf100*, with the corresponding gene name ‘chromosome 16 C1orf100 homolog’.

Gene families with a complex evolutionary history should ideally be named with the help of an expert in the field, as has already been implemented for the olfactory receptor¹⁷ and cytochrome P450 gene families.

Species designation. To distinguish the species of origin for homologous genes with the same gene symbol, we recommend citing the NCBI taxonomy ID¹⁸ as well as either the current name or the GenBank common name (for example, taxonomy ID 9598 and either *Pan troglodytes* or chimpanzee).

Nomenclature updates

Although we are committed to minimizing symbol changes, some updates will still be appropriate. All requests for change are considered on a case-by-case basis and often involve community consultation. We anticipate that most future changes will fall into one of the following categories.

Symbol updates for placeholders.



FAMs, C#orfs and KIAAs are regarded as placeholder symbols and updated with structure- and/or function-based designations whenever possible. However,

when specific placeholder symbols have become entrenched in the literature, we may make exceptions and retain the placeholder, while updating the gene name (for example, *FAM20B* has been retained with the updated gene name ‘FAM20B glycosaminoglycan xylosylkinase’).

Replacing underused and problematic nomenclature. We may consider updating symbols that have rarely or never been published, are not suitable for transfer to other vertebrates, and/or have been widely used but could cause substantial problems (examples in Box 3).

Gene-symbol usage. The HGNC endorses the use of italics to denote genes, alleles and RNAs, to distinguish them from proteins.

We advise authors to quote the approved gene symbol at least once in the abstract of any publication. Every gene with an approved symbol also has a unique HGNC ID in the format HGNC:number (for example, gene symbol *BRAF*, HGNC ID HGNC:1097). Although we aim to minimize symbol changes, some updates are inevitable, and sometimes an approved symbol can be used to denote a different gene in the literature; therefore, we advise quoting the HGNC ID for each gene to avoid ambiguity. HGNC IDs are associated with the gene sequence and do not change unless the gene structure undergoes extreme alteration (such as being merged with another locus or split into multiple loci). This association ensures effective and reliable tracking of data regardless of any nomenclature changes. □

Elsbeth A. Bruford^{1,2} , Bryony Braschi¹, Paul Denny¹, Tamsin E. M. Jones¹, Ruth L. Seal^{1,2} and Susan Tweedie¹ 

¹HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. ²Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge, UK.

✉e-mail: elsbeth@genenames.org

Published online: 3 August 2020

<https://doi.org/10.1038/s41588-020-0669-3>

References

- Shows, T. B. et al. *Cytogenet. Cell Genet.* **25**, 96–116 (1979).
- Shows, T. B. et al. *Cytogenet. Cell Genet.* **46**, 11–28 (1987).
- McAlpine, P. *Trends Genet.* (Mar), 39–42 (1995).
- White, J. A. et al. *Genomics* **45**, 468–471 (1997).
- Wain, H. M. et al. *Genomics* **79**, 464–470 (2002).
- Braschi, B. et al. *Nucleic Acids Res.* **47**, D786–D792 (2019).
- Nagase, T., Koga, H. & Ohara, O. *Brief. Funct. Genomics Proteom.* **5**, 4–7 (2006).
- Seal, R. L. et al. *EMBO J.* **39**, e103777 (2020).
- Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. *Nucleic Acids Res.* **47**, D155–D162 (2019).
- Chan, P. P. & Lowe, T. M. *Nucleic Acids Res.* **44**, D184–D189 (2016).
- O’Leary, N. A. et al. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- Frankish, A. et al. *Nucleic Acids Res.* **47**, D766–D773 (2019).
- Maltais, L. J., Blake, J. A., Eppig, J. T. & Davison, M. T. *Genomics* **45**, 471–476 (1997).
- Burt, D. W. et al. *BMC Genomics* **10** (Suppl. 2), S5 (2009).
- James-Zorn, C. et al. *Genesis* **53**, 486–497 (2015).
- Ruzicka, L. et al. *Nucleic Acids Res.* **47**, D867–D873 (2019).
- Olender, T., Jones, T. E. M., Bruford, E. & Lancet, D. *BMC Evol. Biol.* **20**, 42 (2020).
- Federhen, S. *Nucleic Acids Res.* **40**, D136–D143 (2012).
- den Dunnen, J. T. *Methods Mol. Biol.* **1492**, 243–251 (2017).
- Mayer, J., Blomberg, J. & Seal, R. L. *Mob. DNA* **2**, 7 (2011).
- Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).

Acknowledgements

We thank all current and former members of the HGNC team, particularly the late professor Sue Povey, who was HGNC’s principal investigator from 1996 to 2007, and our specialist advisors and advisory-board members past and present. The HGNC relies heavily on the expertise and feedback of researchers, and we are grateful for all input that we receive. The HGNC is currently funded by the National Human Genome Research Institute (NHGRI) grant U24HG003345 (to E.A.B.) and Wellcome Trust grant 208349/Z/17/Z (to E.A.B.).

Author contributions

E.A.B. directed and obtained funding for the project. E.A.B., R.L.S. and S.T. wrote the original draft. E.A.B., R.L.S., S.T., B.B. and T.E.M.J. revised the manuscript. T.E.M.J. designed Fig. 1. All authors contributed to, and commented on, the manuscript before submission and contributed to the development of the current nomenclature guidelines.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-020-0669-3>.