CCDS Curation Guidelines

Sections:

1. Definitions:

Conservation:

We define conservation by observing sequence similarity for orthologous loci at the level of the genome sequence between two or more species with an emphasis (for curating human and mouse) on conservation observed in the genomes assemblies for human, chimp, macaque, mouse, rat, dog, and cow. Additionally, genome conservation may also be observed within a species for paralogous loci. Agreement with other independently curated datasets such as Swiss-Prot protein records may also be taken into consideration. Genome conservation may be observed using existing public tools, such as the UCSC Vertebrate conservation track, or in similar in-house tools provided to support curator staff.

a. Strong conservation: genome sequence is conserved in at least 2 species that are evolutionarily distant (e.g., different taxonomic orders). Strong conservation support (or experimental data) is needed when considering a large N-terminal extension (>100aa).
b. Weak conservation: genome sequence is conserved in closely related species but not conserved in more distantly related species (e.g., such as within primates, within rodents, or within mouse strains)
c. ***Note***: Variation at the protein termini is valid and expected and can be lineage-specific. Small differences in N-terminal length between, for instance, human and mouse, are expected. Large differences may be valid but should be supported by available transcript, publication, and conservation data.

Kozak signal strength:

- GCC[A/G]CCaugG[not U] == optimal
- [A/G]NNaugG[not U] == strong; ‘A’ at -3 is stronger than ‘G’
- Anything else = ‘weak’

Some known modulators of initiation sites (general, not specific):

- Secondary structure:
    - The folding of an mRNA can affect ribosomal scanning rates and the likelihood of translation initiation. For example, a hairpin secondary structure downstream of a non-AUG initiation site, or downstream of an AUG with a weak Kozak signal, may facilitate pausing and increase the likelihood of initiation from that site (PMID: 2236042).
- uORF:
    - The translation of upstream ORFs (uORFs) can negatively affect translation initiation at the downstream primary ORF (pORF), but uORFs do not appear to be entirely inhibitory to pORF translation (PMIDs: 19372376 and 23624144). It is generally thought that long uORFs and uORFs in a stronger Kozak sequence context are more inhibitory, whereas short uORFs (e.g., <35 aa) may be more amenable to translation reinitiation, and uORFs in a weaker Kozak context may be more amenable to leaky scanning. The translation of uORFs and their regulatory effects on downstream initiation events may also depend on gene-, developmental-, or spatially-specific conditions.
- pORF:
    - The primary ORF (pORF) considered to be the functional open reading frame based on homology, publications, and/or sequence content (protein domains, etc).
- This list is not intended to be comprehensive.
    - Other factors, such as RNA-binding proteins or post-translational modifications of translation initiation factors, do come into play and may vary in a gene-specific, developmental, or spatial manner.

<u>Note</u>: A strong Kozak signal is less likely to be permissive to leaky scanning that would enable the use of a downstream start codon. Nevertheless, ribosome profiling evidence shows that initiation events can occur downstream of strong Kozak signals (PMID: 22927429). The strength of the downstream pORF Kozak signal is much less important (see PMIDs: 12459250 and 16213112).

<u>Note</u>: More support is required to annotate at an internal AUG site than at the first AUG site, when there are alternate possible start sites available on one transcript. This is because according to the scanning translation model, even if the first AUG is not in an optimal context then translation will generally still occur from that location at some frequency because the ribosome will pause at the AUG. However, because the context is sub-optimal, the ribosome may not always pause long enough for initiation to *<u>consistently</u>* occur at the first AUG site, and a thus the ribosome may continue scanning, detect a downstream AUG, and initiate translation from that alternate site. If there are several AUGs in a weak context then initiation may actually occur at all sites. Once the ribosome encounters an AUG in a strong context it is less likely to continue scanning.

Nonsense-mediated mRNA decay (NMD) candidate:
A transcript is considered an NMD candidate if the stop codon is >50-55 nt upstream of the last splice junction. The CCDS collaboration uses the threshold of >=50 nt. See PMIDs: 15040442 and 23435113 for more information about NMD.

Readthrough transcript:
A transcript that contains exon structure from two or more distinct and adjacent genes on the same strand, i.e., transcription “reads through” the normal termination signals from one gene and into another.

Inferred CCDS representation:
A CCDS representation for which the transcript exon combination lacks full-length support from transcripts deposited in public International Nucleotide Sequence Database Collaboration (INSDC) databases.

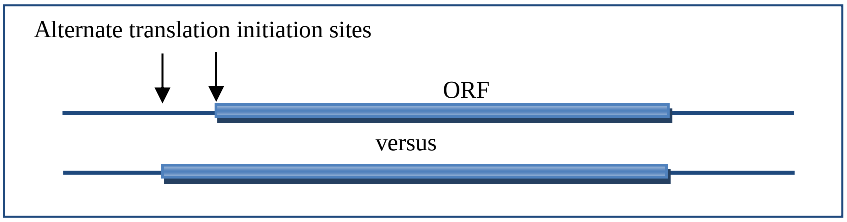
CCDS Attributes:
A tag attached to select CCDS IDs to indicate a particular characteristic of the representation. Attributes can be found in a designated ‘Attributes’ section of individual CCDS report pages, or in the ‘CCDS_attributes.[YearMonthDay/current].txt’ files in the CCDS FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/).

Public Note:
A public statement that is associated with select CCDS IDs. Such notes are found in a designated ‘Public Note’ section of individual CCDS report pages. These may indicate the reason why a CCDS representation was updated or withdrawn, or they may describe supporting evidence for a particular representation (e.g,. data from publications or inferences from orthologous data), or they may further explain an associated attribute.

2. Curation Guidelines

2.A. Extension of the ORF – assessing alternate in-frame translation initiation sites

The following start codon selection guidelines are used for a transcript that contains multiple possible in-frame start codons:



Default Rule: Always annotate the CDS starting from the upstream AUG ***unless*** one of the following exceptions applies.

Note: The expectation is that frequently none of the exceptions will apply; therefore, we will often annotate an upstream AUG with a weak Kozak signal and weak conservation.

Attribute: Downstream AUG exceptions represented in the CCDS data set are tracked with the 'CDS uses downstream AUG' attribute, which can be found in the 'Attributes' section of relevant CCDS reports, or in the 'CCDS_attributes.[YearMonthDay/current].txt' files in the CCDS FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). In some cases, an explanatory Public Note may also accompany the attribute in the CCDS report page.

1. Strong experimental evidence shows that the downstream start codon is used. Experimental evidence may include:

- Protein N-terminal sequencing.
- N-terminal-specific antibody support.
- Translation assays that include both start codons. *Caution:* If the upstream start codon has not been tested in the experimental assay, then evidence showing that the downstream AUG can be used does not necessarily mean that the upstream start cannot be used.
- Evidence of 5' UTR secondary structure/protein interactions that would preclude the use of the upstream AUG.

- Evidence that the shorter protein originating from the downstream AUG is the primary functional product (short protein is more abundant and function has been ascribed to the short protein).
- Ribosome profiling evidence, particularly when there is a credibly high score and when there is no support for use of the upstream in-frame start codon.
- Evidence for cleavage of the N-terminal methionine derived from the downstream AUG, coupled with N-terminal acetylation of the next residue.

Note: Consideration should be given for alternatively spliced transcript variants when applying experimental support for downstream AUG use, e.g., variants that lack the upstream AUG could explain the experimental support for downstream AUG use.

2. The downstream AUG has a strong history of use and is considered to be the community standard. A 'strong history' may be indicated by several (5) high quality publications (or 10+ relevant publications) indicating consistent definition of the protein, especially the N-terminus or publications that cite point mutations at specific locations relative to a CDS standard. **Please contact a scientific expert** to determine if there is experimental support for either AUG in question, to determine if the community is aware of the upstream AUG, and to make a final decision regarding annotation.

Note: We expect that the upstream Kozak signal is not 'strong' and that genome conservation for the upstream AUG may also be weak.
Note: *There may be cases where new data must override historical use because historical use was based on incomplete knowledge*. A low number of publications perhaps based on an initial clone that is currently considered to be 5' incomplete or that represents 5' indels or that prevent the use of the upstream start codon, should not be sufficient evidence to annotate the downstream AUG.

3. The upstream AUG and in-frame extension is not conserved in any other species (does not fulfill our weak conservation definition):

- AND functional information is available or can be inferred for the shorter protein or a similar isoform, either publication support or convincing domain structure that would be indicative of a function (e.g., always annotate the longest ORF for genes of unknown function), OR if no functional or domain information is available, the protein is strongly conserved and all other species use the downstream start codon
- AND some combination of:
    - the N-terminal extension does not add a signal or transit peptide
    - the N-terminal extension does extend a signal peptide to an unreasonable length (roughly, >40 aa)***
    - the N-terminal extension is very large resulting in a protein length significantly different from either homologs or paralogs where the homolog/paralog is itself considered well supported and of full length and the difference is not due to alternate

splicing. ***Note***: if the N-terminal extension under consideration is very large (say, ~>100aa) then require strong conservation support or experimental data.

- the N-terminal extension does not add or complete a domain
- the downstream AUG site has a strong Kozak signal context and/or is more strongly conserved

*** See distribution of eukaryotic signal peptide lengths at http://www.cbs.dtu.dk/services/SignalP-1.1/sp_lengths.html

Note: According to the scanning model, the ribosome does not know *a priori* whether the upstream or downstream AUG is conserved in other species; however, other regulatory factors such as conserved secondary structure could play a role.

4. A signal or transit peptide is predicted for the shorter, but not the longer, N-terminal AND there is clear experimental evidence that leader peptide presence or cleavage is necessary for a functional protein. E.g., there is experimental evidence in the literature indicating that this protein is secreted or targeted to a cellular compartment that requires the leader peptide, and the longer protein lacks the leader peptide. This evidence may be inferred from an ortholog or similar family member.

Note: Consideration should be given for alternatively spliced transcript variants, e.g., there may be variants that can only use the downstream AUG and produce the isoform(s) with a leader peptide.

5. There is a non-canonical in-frame start codon (e.g., CTG, GTG, or ACG) located 5' of the first AUG, and the non-AUG site is:

- Experimentally verified as a translation initiation site (support could be from a homolog)
- Or, use of upstream non-AUG site completes a protein domain
- Or, completes or adds a signal or transit peptide (AND there is experimental support for a targeted location – support could be from a homolog)
- Or, there is very strong genome conservation for the alternate start site

Attribute: Non-AUG initiation codons represented in the CCDS data set are tracked with the 'Non-AUG initiation codon' attribute, which can be found in the 'Attributes' section of relevant CCDS reports, or in the 'CCDS_attributes.[YearMonthDay/current].txt' files in the CCDS FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). In some cases, an explanatory Public Note may also accompany the attribute in the CCDS report page.

Note: Kozak indicates that use of a non-AUG occurs if there is an optimal Kozak signal context and/or the mRNA structure supports pausing of the ribosome over the non-AUG site long enough to allow the inaccurate pairing of the initiating Met-tRNA and the non-AUG codon. There may be leaky scanning and initiation at an alternate downstream AUG site in addition to

initiation at the upstream non-AUG site. Therefore, we should have good support for a decision to annotate the CDS initiating at a non-AUG site.
Note: A hairpin secondary structure downstream of a weak Kozak site may facilitate ribosome pausing and thus increase the likelihood of initiation from that site (PMID: 2236042).

6. Any case that does not cleanly fit into the above guidelines:
    - Final annotation decision is made following discussion among the CCDS collaborators. This discussion may include consultation with other scientists.

Start Codon Selection Example Cases:

For annotating the CDS starting from the first AUG:

a) There is a strong Kozak signal for the first AUG and it is an extension of the primary ORF (regardless of the Kozak signal for the downstream AUG, and regardless of genome conservation).
b) There is a weak Kozak signal but there is strong genome conservation for the first AUG and the extension doesn't conflict with experimental information about translation initiation or localization. In this context, strong conservation at the level of genome sequence is observed between two or more species; the species do not need to be widely diverged (e.g. primate-specific N-terminal differences are valid).
c) There is a weak Kozak signal and there is weak or no conservation for the first AUG, but the extension improves the protein with regard to adding or completing a domain or signal/transit peptide.
d) There is no functional information, whether direct or indirect (domains), for the protein function.

For annotating the CDS starting from an internal AUG:
a) There is a weak Kozak signal and no conservation for the first AUG, and very strong conservation and a strong Kozak signal at a downstream AUG. There is significant genome conservation observed among species with evolutionary distance and there is consistency in the location of the downstream AUG site and N-terminus region of the protein.
b) There is a very strong historical use; the protein as defined from the internal AUG is considered the community reference standard. ***Note***: if you think this is a case where newer data indicates historical use is faulty then it may be useful to consult with an expert on the gene/protein to confirm the N-terminus representation. The community standard N-terminus should be supported by available public data. In other words, there should be a compelling reason to not annotate from the upstream AUG especially when there is conservation support or a good Kozak signal. In one real case, the community expert pointed out that the upstream AUG site being considered

was invalid because the transcript representation was in error; promoter studies determined that the predominant transcript start occurred after the first AUG site that was in question. The transcript representation had been extended further 5’ of the known promoter based on weak transcript support that the scientific expert did not consider valid.

c) ***Note***: if the ‘internal’ AUG site in question is also available as the first AUG site on a different transcript due to use of an alternate promoter, or alternate splicing, then (given sufficient support) both transcripts and both N-terminal protein options can be annotated. Naturally, all transcripts have to themselves meet quality and abundance criteria to be considered for representing as annotated alternate transcripts.

Cases where a leader peptide can be predicted for both N-termini are less clear and may require further discussion, with consideration for the signal peptide length.

2.B. Nonsense-mediated mRNA decay (NMD) exceptions

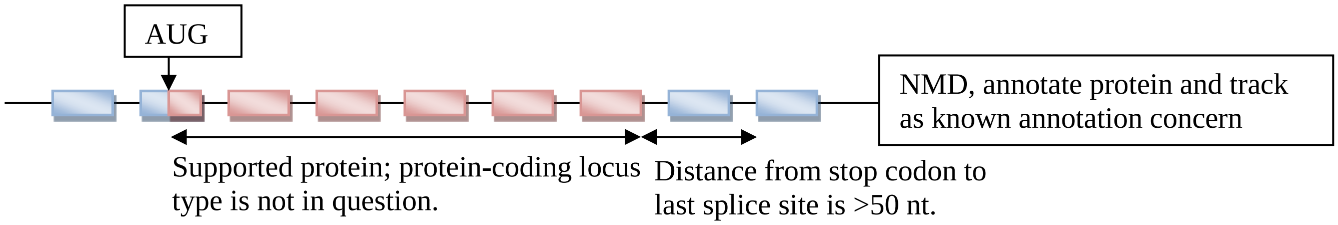
Nonsense-mediated mRNA decay (NMD) is a eukaryotic surveillance pathway that destroys abnormal transcripts encoding a truncated protein due to the presence of a premature termination codon (PMIDs: 12502788, 15040442 and 23435113). The CCDS collaboration NMD guidelines are based on the exon junction complex model (PMIDs: 15040442 and 23435113), whereby a transcript is assumed to be an NMD candidate if the stop codon is located >50 nts upstream of the last exon-exon junction.

The products of NMD transcripts are generally not represented in the CCDS dataset unless the exceptions outlined in Cases 1 and 2 below apply.

Attribute: NMD exceptions with CCDS representation are tracked with the ‘Nonsense-mediated decay (NMD) candidate’ attribute, which can be found in the ‘Attributes’ section of relevant CCDS reports, or in the ‘CCDS_attributes.[YearMonthDay/current].txt’ files in the CCDS FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). In some cases, an explanatory Public Note may also accompany the attribute in the CCDS report page.

Case 1:

The gene is protein-coding; there is abundant transcript data; all available transcripts are NMD candidates. E.g., this is a known protein-coding gene and it is considered an error of omission to not represent the protein.

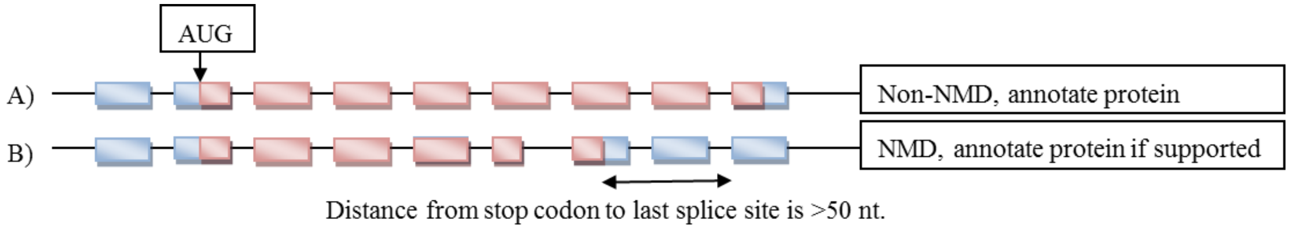

AUG
NMD, annotate protein and track as known annotation concern
Supported protein; protein-coding locus type is not in question.
Distance from stop codon to last splice site is >50 nt.

Decision 1: The protein is represented from an NMD transcript if:

- One or more long transcripts are available
- AND the protein has homology support from two or more species, or has a supporting publication
- AND the gene cannot be considered a pseudogene, e.g., the supported protein could not be derived from non-NMD transcripts from a highly similar paralogous gene

Case 2:

The gene is protein-coding; there is abundant transcript data; some transcripts are NMD candidates; other transcripts are not NMD candidates AND a supported protein can be represented from the non-NMD transcripts. There is also experimental evidence* for the production of an isoform from the NMD transcript.


AUG
A)
B)
Non-NMD, annotate protein
NMD, annotate protein if supported
Distance from stop codon to last splice site is >50 nt.

Decision 2: The protein is represented from a non-NMD transcript. The protein may ALSO be represented from NMD transcripts ONLY IF:

- One or more long transcripts with the different exon combination are available
- AND there is publication support indicating that the NMD transcript is translated, or it is demonstrated that the isoform encoded by the NMD transcript is found *in vivo** and there is no non-NMD transcript variant that can encode it.

Note: Evidence for the endogenous protein is required. Protein support based on transfected cDNA constructs or in vitro translation of mRNAs is not considered sufficient support because pre-processed mRNAs are unlikely to be subjected to NMD based on the exon junction complex model.

2.C. Readthrough loci

Readthrough transcripts arise when transcription initiates in one gene, continues past the normal transcription termination signals for that gene, and terminates within or at the end of a distinct downstream gene on the same strand. The resulting transcripts contain exon structure from both genes.

Some exons of either gene may be skipped and novel exons may be included. Some readthrough transcripts may span more than two genes. Readthrough transcripts may encode a protein derived from coding exons from one or both loci (e.g., may be the same as the downstream locus; may be a fusion protein based on coding sequence from both genes), or they may encode a novel protein product due to CDS frameshifts with respect to one of the genes, or they may be non-coding due to NMD.

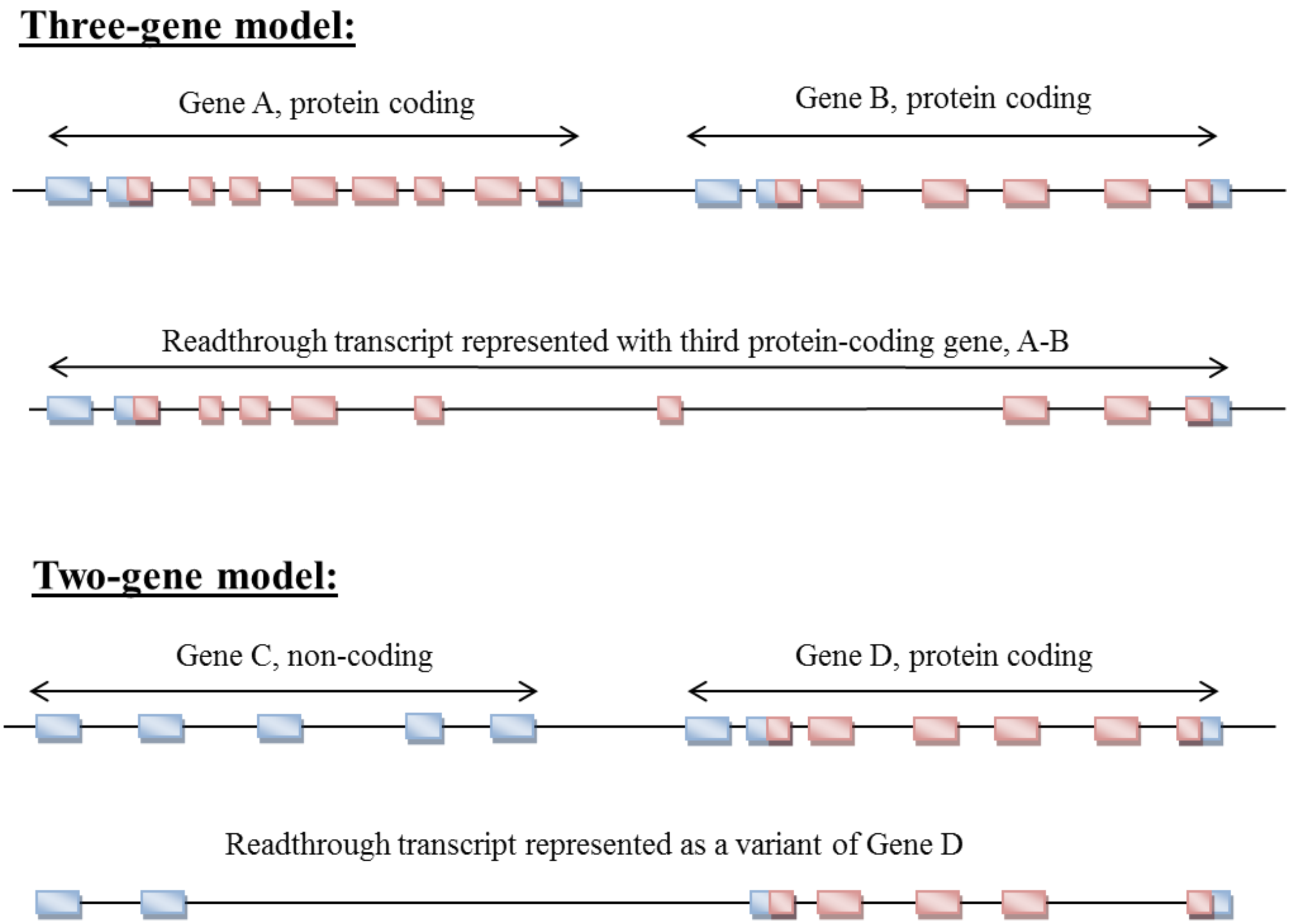
The CCDS collaboration definition of readthrough is very specific in that the individual partner genes must be distinct, and the readthrough transcripts must share >=1 exon (or >=2 splice sites except in the case of a shared terminal exon) with each of the distinct shorter loci. Unlike the broader definition of "conjoined" genes described in Prakash et al. (PMID: 20967262), the CCDS readthrough definition does not include cases where the genes are otherwise considered to be co-transcribed (e.g., human *HOXC4*, *HOXC5* and *HOXC6*) (PMID: 2898768), bicistronic (e.g., human *CERS1* and *GDF1*)(PMID: 2034669), or overlapping each other but not sharing splice sites (e.g., the 3' exon of the mouse *Mon1b* gene overlaps the 5' exon of the *Syce1l* gene), or genes that have nested arrangements relative to each other (e.g., human and mouse protocadherin gene clusters).

Note: Any readthrough transcript that has matching protein annotation from NCBI and Ensembl/Havana is eligible for CCDS representation. However, recent reports in the literature (example PMID:26861889) suggest that readthrough transcripts are pervasive, especially when cells are subject to stress, and may not encode a protein. Furthermore, both NCBI and Ensembl/Havana use two-gene or three-gene models to represent readthrough transcripts (see below), which can result in user confusion about transcript:gene association and/or the third readthrough locus may artificially increase the gene count. Therefore, the CCDS collaboration has decided to represent proteins encoded by readthrough transcripts only if there is strong transcript support from multiple sources or published experimental support for the protein encoded by the readthrough transcript. In future, assignment of CCDS IDs to proteins encoded by readthrough transcripts will be considered by CCDS curators on a case-by-case basis.

In most cases, the collaboration uses a separate locus to represent readthrough transcripts (three-gene model when there are two distinct individual genes, see diagram below). However, depending on the locus type of the individual genes, in some cases a two-gene model is used and the readthrough transcript is treated as a variant of one of the individual genes. The decision to represent a two-gene versus a three-gene model also includes a protein similarity consideration, i.e., a consideration of whether the protein produced from the readthrough transcript is more similar to the protein product of one individual gene versus the other, or if the readthrough product is very different. The following criteria are used for locus type combinations that could produce a protein-coding readthrough transcript:

Treatment	Upstream locus type	Downstream locus type
Three-gene model	Protein-coding	Protein-coding, non-coding, or transcribed pseudogene

Three-gene model	Transcribed pseudogene	Protein-coding
Two-gene model	Transcribed non-coding	Protein-coding


Three-gene model:
Gene A, protein coding
Gene B, protein coding
Readthrough transcript represented with third protein-coding gene, A-B
Two-gene model:
Gene C, non-coding
Gene D, protein coding
Readthrough transcript represented as a variant of Gene D

Evidence required:***

******Note:*** The following requirements are based on current RefSeq guidelines, which are more stringent with respect to transcript completeness and independent support evidence than currently required for Havana readthrough annotation (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf). CCDS representation is possible only if both NCBI and Ensembl/Havana annotate the readthrough protein product, hence the more stringent RefSeq support criteria are necessary for CCDS representation.

- One long transcript that can be used to represent the full length CDS. It must share >=1 exon (or >=2 splice sites except in the case of a shared terminal exon) with each of the distinct shorter loci.
- AND a second line of support for the readthrough event (exon sharing criteria as above); may be a partial transcript and not necessarily the same transcript variant.
- OR published support for the protein originating from the readthrough transcript (in vivo evidence for the endogenous protein).
- OR homology support for the readthrough transcript when the cross-species transcript is itself sufficiently supported to represent the readthrough locus within that species; cross-species support should not be used in the absence of a same species full-length readthrough transcript.

Nomenclature for readthrough CCDS representations:

- The gene symbols used in the CCDS database are propagated from NCBI's Gene database (http://www.ncbi.nlm.nih.gov/gene/), hence the CCDS readthrough nomenclature will be based on the GeneID that the CCDS RefSeq is associated with.
- In most cases, the CCDS ID for the readthrough product will have a readthrough gene symbol (usually a hyphenated amalgamation of the official individual gene symbols, or as determined by the HUGO Gene Nomenclature Committee or Mouse Genome Informatics for human or mouse, respectively).
- In cases where the protein product is the same as one of the individual gene products (e.g., the readthrough transcript shares only UTR structure with one of the genes and has the entire CDS of the other gene), or when a coding readthrough transcript is represented in the two-gene model, the resulting CCDS ID will have the symbol of the relevant individual gene, not a readthrough gene symbol. Non-readthrough transcripts from the individual gene may also be associated with such CCDS IDs.

2.D. Inferred CCDS Representations

The CCDS collaboration aims to represent only high quality and supported protein-coding sequences in the dataset. Ideally, this means that every CCDS representation should include full-length transcript support for the CDS exon combination, in addition to other evidence such as conservation, functional data or predicted domain structure in the protein. In practice, however, many valid transcript variants or very long proteins lack full-length transcript support, as available in public International Nucleotide Sequence Database Collaboration (INSDC) databases.

In order to fulfill the CCDS project's goal to represent as many consistently annotated protein-coding genes as possible, the collaboration allows inferred exon combination representations in the dataset. This typically occurs in two scenarios:

1. When a known protein-coding gene lacks full-length transcript support but a full-length protein can be inferred from partial transcript and/or homology, orthology or publication data, e.g., CCDS44873.1 representing the *KMT2D* gene.

2. When the gene contains cassette exons or multi-exon cassettes that are individually supported (by transcript, conservation or publication data) but where full-length transcript support for a complete complement of exons is lacking. For example, CCDS59435.1 represents the longest possible full-length splice variant of the *TTN* gene with an exon combination that is not supported by any single transcript.

Attribute: CCDS representations based on inferred exon combinations are tracked with the 'Inferred exon combination' attribute, which can be found in the 'Attributes' section of relevant CCDS reports, or in the 'CCDS_attributes.[YearMonthDay/current].txt' files in the CCDS FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). In some cases, an explanatory Public Note may also accompany the attribute in the CCDS report page.

Note: A subset of CCDS representations that have not yet underwent curator review may also be inferred representations. These are annotations created by Ensembl and NCBI automatic processing, i.e., Ensembl annotations that have not been manually annotated by Havana, and NCBI RefSeqs that have 'Provisional,' 'Predicted' or 'Inferred' status (not 'Reviewed' or 'Validated,' which signifies review by a curator). Nonetheless, such non-reviewed CCDS representations may be present in the dataset due to identical CDS annotations from both Ensembl and NCBI.

3. References:

1. Kozak M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 361:13-37. (PMID: 16213112)
2. Kozak M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. Gene. 299(1-2):1-34. (PMID: 12459250)
3. Kozak M. (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc Natl Acad Sci 87(21):8301-5. (PMID: 2236042)
4. Calvo SE, Pagliarini DJ, Mootha VK. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA. 106(18):7507-12. (PMID: 19372376)
5. Somers J, Pöyry T, Willis AE. (2013) A perspective on mammalian upstream open reading frame function. Int J Biochem Cell Biol. 45(8):1690-700 (PMID: 23624144)
6. Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 109(37):E2424-32. (PMID: 22927429)
7. Lewis BP, Green RE, Brenner SE. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA. 100(1):189-92. (PMID: 12502788)
8. Maquat LE. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol. 5(2):89-99. (PMID: 15040442)
9. Schweingruber C, Rufener SC, Zünd D, Yamashita A, Mühlemann O. (2013) Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells. Biochim Biophys Acta. 1829(6-7):612-23. (PMID: 23435113)
10. Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, Nishida Y, Fujikake T, Takeda T, Taylor TD. (2010) Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PLoS One. 5(10):e13284. (PMID: 20967262)
11. Simeone A, Pannese M, Acampora D, D'Esposito M, Boncinelli E. (1988) At least three human homeoboxes on chromosome 12 belong to the same transcription unit. Nucleic Acids Res. 16(12):5379-90. (PMID: 2898768)
12. Lee SJ. (1991) Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. Proc Natl Acad Sci U S A. 88(10):4250-4. (PMID: 2034669)