havana

human and vertebrate analysis and annotation

annotation guidelines

29 October 2009

page left intentionally blank *

* that is of course a paradoxical statement: the act of printing the text contradicts it, negates its truth. It is basically a pseudomenon, the equivalent of a liar paradox.

this is a blank page *

* see remark on previous “blank” page

havana
human and vertebrate analysis and annotation

annotation guidelines

Gene Classification

Currently Havana genes are subdivided into the following locus categories. Only "Known genes" are set directly from the "Known" tag in the Locus. Other types are set directly through transcript types that are attached to the locus. Yet others are typed indirectly from underlying transcript types (shown in *italics* below).

Known gene
is identical to species native cDNA or protein sequences identified by a GeneID or approved gene name/symbol in, depending on model organism:
Human: Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)
Human: HGNC (http://www.genenames.org/)
Mouse: Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)
Mouse: MGI (http://www.informatics.jax.org/)
Zebrafish: Zfin (http://zfin.org/cgi-bin/webdriver?MIval=aa-newmrkrselect.apg)
Protein coding as well as non-coding loci can be tagged as Known, but pseudogenes cannot (even if they have approved gene symbols).

Novel coding gene
has a CDS (coding sequence) and is identical, or has homology, to cDNAs or proteins but does not fall in the above category; can be known in the sense that there are mRNA sequences for it in the public databases, but it is not yet represented in Entrez Gene or has not received an official gene name. Can also be novel in that it is not yet represented by an mRNA sequence in the species concerned or there isn't a locus-specific mRNA for this copy of the gene in a gene family or cluster.

Novel transcript
is as above but no ORF (open reading frame) can be unambiguously assigned as a CDS; it can be a genuine non-coding gene or can be a partial gene because of the limits of the evidence it is based on. Contains four or more exons and/or is supported by at least one mRNA or three ESTs.

Putative novel transcript
is identical, or has homology, to spliced ESTs but is devoid of a significant ORF and polyA features; these are short genes or gene fragments with three or fewer exons, supported by one or two ESTs.

Pseudogene

is characterised by a disrupted CDS (frameshifts, in-frame stop codons) compared to parent gene(s). Pseudogenes can be processed or unprocessed, and transcribed or not.

Transposon

Special category for Zebrafish, not for general use. Used for tagging transposons in the Zebrafish genome.

Artefact

Used to tag mistakes in the public databases (Ensembl/SwissProt/trembl): the transcript model is tagged for its translation to be removed. Usually these arise from high-throughput cDNA sequencing projects which submit automatic annotation, sometimes resulting in erroneous CDSs in what may for example be 3′ UTR.

Full name artefact gene

Also used for variants based on cDNAs with artefactual "splice" sites. These manifest themselves as jumps from the middle of one exon to the middle of one further downstream, often skipping exons in between, presumably through recombination. Sometimes the "splice" is actually within the last exon or 3′ UTR. Characteristically the "splice" junction is repeated on the genome, *i.e.* a number of mRNA bases can be aligned equally well to both sides of the junction. These artefacts are annotated as a variant of the locus.

NOTE: Artefacts must be made from species-specific mRNAs only (not from ESTs or other species).

EXCEPTION: if an artefact transcript has both an artefact event and a genuine variation that is not covered by other evidence, annotate the section of the transcript containing the genuine event up to or starting from the artefact event (depending on if the latter is downstream or upstream of the former) as a normal transcript variant. But still build an Artefact typed transcript representing the entire artefactual cDNA.

TEC

"**T**o be **E**xperimentally **C**onfirmed" is used for non-spliced EST clusters or single-exon mRNAs (that are not otherwise confirmed) that have good polyA features. Experimentalists will use 5′ RACE/ PCR to try to confirm and extend the transcript.
Note the following exception to the conventional naming convention:

Full name TEC

Only use this for a locus, not for a variant.

NOTE: **LEC** (**L**ocus for **E**xperimental **C**onfirmation) is a tag to highlight loci for targeted experimental investigation. For example loci with no best-in-genome support or fragmented loci (*i.e.* loci with discontiguous fragments supported by gappy homology).

Locus Annotation Remark:
LEC

Building Transcript Objects

Each transcript is assigned a type. If there is only one transcript, the locus type is directly derived from this. If there are multiple variant transcripts, each with their own type, the locus type is determined by looking at the hierarchy of transcript types (*i.e.* CDS types trump transcript types, known type trumps others, etc.).

NOTE: most of the suggested rules shown here can be set aside in the face of strong homology and cross-species evidence.

Using non-best-in-genome and non-organism-supported evidence to build transcript models

When full-length best-in-genome evidence (*i.e.* locus-specific) is present, do not use non-b-i-g evidence (*i.e.* from paralogs or homologs) to support splice variants, extensions of locus-specific evidence based variants or polyA features.

Non-organism-supported evidence (*i.e.* from orthologous loci or other species in general) can be used to build variants on the condition that homology is perfectly co-linear and all splice sites are canonical. Do not build retained intron or NMD variants based on non-organism evidence or use it to extend variants based on locus-specific evidence.

If there is either only partial or no b-i-g locus-specific evidence, transcript models can be built using evidence from either other loci (non-b-i-g evidence) or other species (non-o-s evidence) and should preferably be full-length. See ***Table 1*** and ***Table*** 2 below for guidance.

IMPORTANT

NOTE: non-b-i-g and non-o-s evidence should never be used to support polyA features.

NOTE: it is especially important for loci that appear in clusters of very similar genes to make sure that supporting evidence is locus-specific and not from another locus in the cluster or from another species.

NOTE: where non-b-i-g and/or non-o-s evidence is indicating a number of different potential splice variants in the absence of locus-specific evidence, choose only one representative variant. Where possible the best match, the longest, with most exons, greatest coverage and longest CDS.

Table 1: when and how to use non-best-in-genome evidence

Coding	▪ Make sure locus is coding and not a pseudogene
Transcript	▪ Only annotate splice variants based on non-b-i-g mRNAs; DO NOT use non-b-i-g ESTs
Putative transcript	▪ Only annotate when non-b-i-g ESTs splice perfectly and support a structure >= 3 exons long

Transcript Annotation Remark:
non-best-in-genome evidence

Table 2: when and how to use non-organism-supported evidence

Coding	▪ Make sure locus is coding and not a pseudogene ▪ DO NOT use non-o-s evidence to extend UTRs of coding splice variants

Transcript Annotation Remark:
non-organism_supported

Genomic sequence errors
If a genome sequence error is suspected, check whether it is a known validated SNP/DIP (see also **Polymorphic Pseudogenes** on page 18) (use Ensembl, UCSC). If it isn't, mail the designated Havana team member that deals with these issues with the genomic clone accession number, the twenty or so bases of sequence flanking the error (in case of indels and substitutions) with the affected base(s) marked, the cDNA coordinate(s) of the error on a disagreeing cDNA (with accession umber), the gene symbol of the affected gene, the details of the error, the amino-acid change(s) if any, the number and nature of sequences disagreeing with the genomic sequence (*e.g.* 14 human ESTs, 3 human, 1 chimp and 1 cow cDNA) and the accession numbers of at least a representative sample of these cDNAs and ESTs. Build a transcript as a Transcript type if the error has a detrimental effect on the CDS. If the error is a simple indel or substitution in UTR or non-fatal substitution in the CDS, make transcript coding as normal. Either way build the transcript as if the error wasn't there if possible. Add a visible remark only if the CDS is affected:

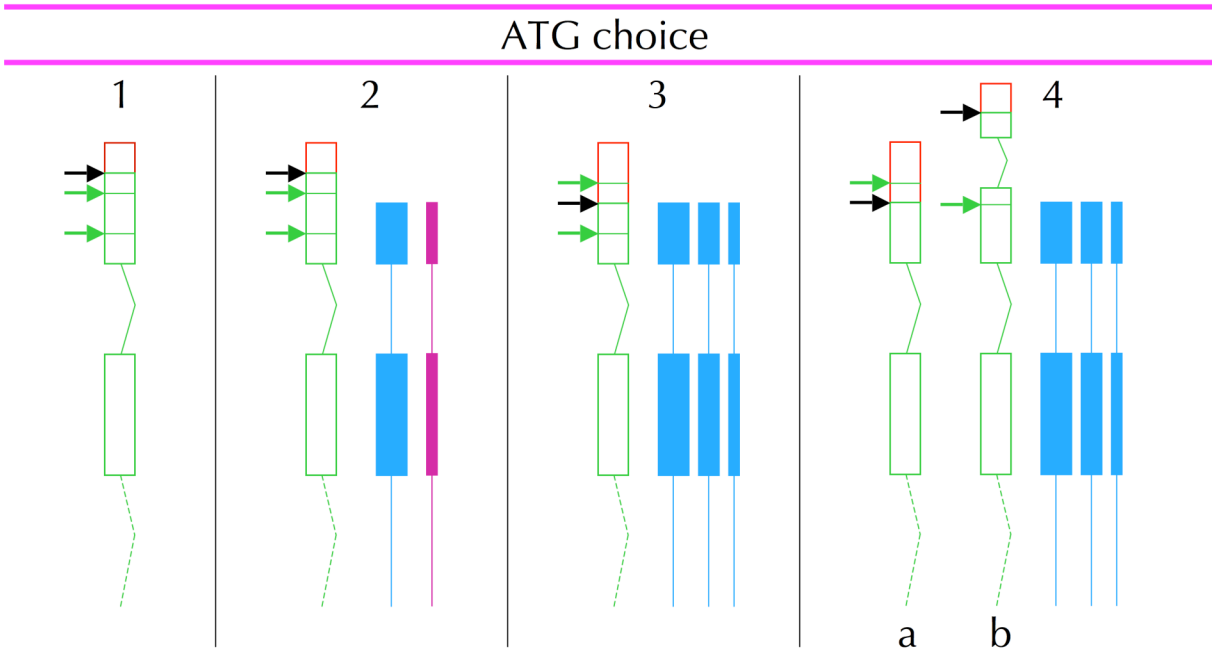
Transcript Visible Remark:
suspected genomic sequence error affecting CDS in exon <exon number>

WARNING: Genoscope mRNAs are modified to correspond to genomic sequence so should not be trusted in deciding whether a potential sequencing error could be a polymorphism.

Defining the coding region

As we only annotate one CDS per variant we have to take several factors into account when assigning an ATG in an attempt to annotate the CDS most likely to represent the function of the variant. The scanning model of initiation proposed for eukaryotes suggests that some degree of translation will initiate from the first ATG the ribosome encounters, however, the level of transcription from an ATG is highly dependent on its context and may range from negligible to 100%. The longest ORF may also not encode the main functional protein product of a variant. Where strong evidence that a downstream ATG starts the functional protein *e.g.* conservation (making the assumption that sequences are conserved because they have a conserved function) or published evidence for structure or activity of the shorter protein, the downstream ATG should be used.

Figures below show the practical application of these guidelines (***Figure 1***, ***Figure 3***, ***Figure 4***). In ***Figure 1***, Locus 1 has no protein support so the most upstream ATG should be used. Locus 2 has same-species SwissProt protein support, cross-species Trembl support or inconclusive conservation in UCSC browser; again the most upstream ATG should be used. Locus 3 has good cross-species support for a downstream ATG, *i.e.* SwissProt protein from ≥1 other species using the same ATG or strong conservation of the downstream ATG in the UCSC browser. If either or both of these is true and there is no strong conservation of the upstream ATG in UCSC browser, then the downstream ATG should be used and the upstream_ATG- annotation remark should be added. These rules should be applied specifically to each splice variant where multiple coding variants are present. Locus 4 has two alternative splice variants **a** and **b**. Variant **a** has good conservation evidence for a downstream ATG and as such the downstream ATG should be annotated (with an upstream_ATG- annotation remark) variant **b** has no conservation or functional evidence and so the most upstream ATG should be used. In all cases, published functional or structural evidence supersedes ATG order and conservation evidence in assigning an initiating ATG. Taking into account the strength of the Kozak sequence (***Figure 2***) also helps deciding on the best start ATG. A strong Kozak sequence suggests that the ATG is likely to initiate translation. A weak one will do some of the time but the ribosome may scan past it and initiate at a downstream ATG.



1 - No support: use most upstream ATG
2 - Same-species/weak cross-species support: use most upstream ATG
3 - Good cross-species support (SwissProt or UCSC browser conservation tracks): use conserved ATG and add upstream_ATG tag
4 - Good cross-species support and alternative variant with weak or no support: for main variant (a) use conserved ATG and add upstream_ATG tag. For alternative variant (b) use most upstream ATG.

Figure 1: alternative ATGs

As mentioned above, when an ATG further downstream is used tag as follows:

Transcript Annotation Remark: upstream_ATG–<distance in aa upstream>
upstream_ATG-10

(if there are multiple upstream in-frame start codons just note the most distant)

NOTE: this tag is used only on the main (reference) variant. Splice variants that have unique upstream ATGs (owing to a novel 5′ exon) will use that ATG and are typed Putative_CDS (***Figure 9***).

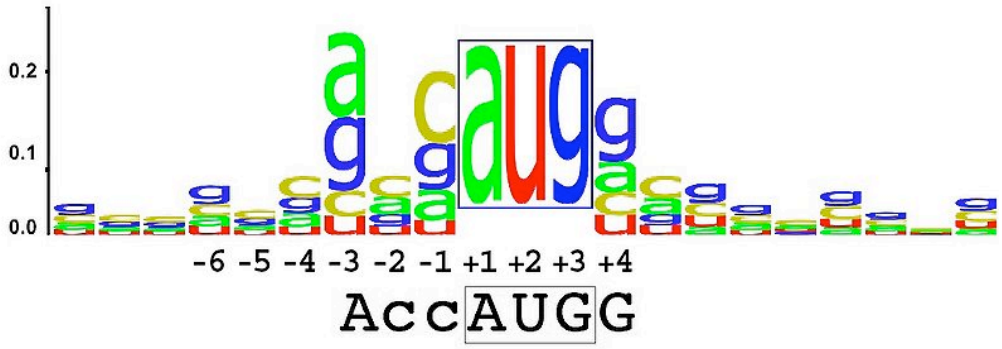


Figure 2: Kozak sequence LogoGraph

NOTE: in the Kozak sequence, the most critical positions are **-3** and **+4**:

A at **-3** = ***strong***
G at **-3** plus **G** at **+4** = ***strong***
Anything else = ***weak***

Any splice variant that produces a translation >35aa (including the stop) based on a reference ATG, where the stop codon is >50bp from a downstream splice site, should be labelled NMD (***Figure 7***). If the upstream translation is <35aa, translation may be re-initiated from an internal downstream ATG or a unique downstream ATG; such a CDS can be used if it passes the normal CDS criteria providing it shares at least some translation in the same frame as a reference (*i.e.* a locus cannot have a coding variant with a translation that has nothing in common with any other translaton).

The stop codon must be in the last exon or no further than 50bp from the end of the penultimate exon, as otherwise it is likely to be a target for NMD (unless experimental evidence or publications indicate otherwise) (***Figure 8***).

Transcripts can have non-ATG starts, which should be annotated just like ATG, provided the validity is supported by publication and/or conservation. Add in annotation remark:

Transcript Visible Remark: non-ATG start codon

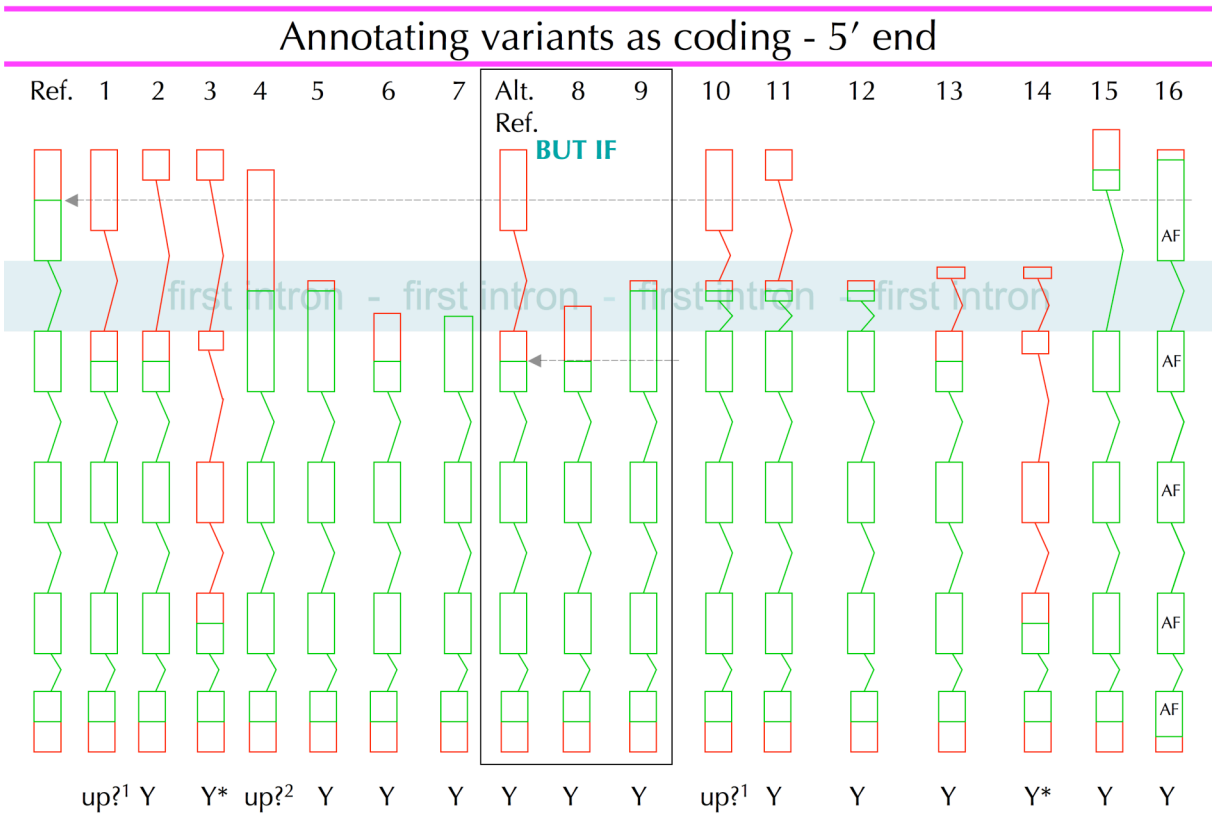


Figure 3: annotating variants as coding - 5' end

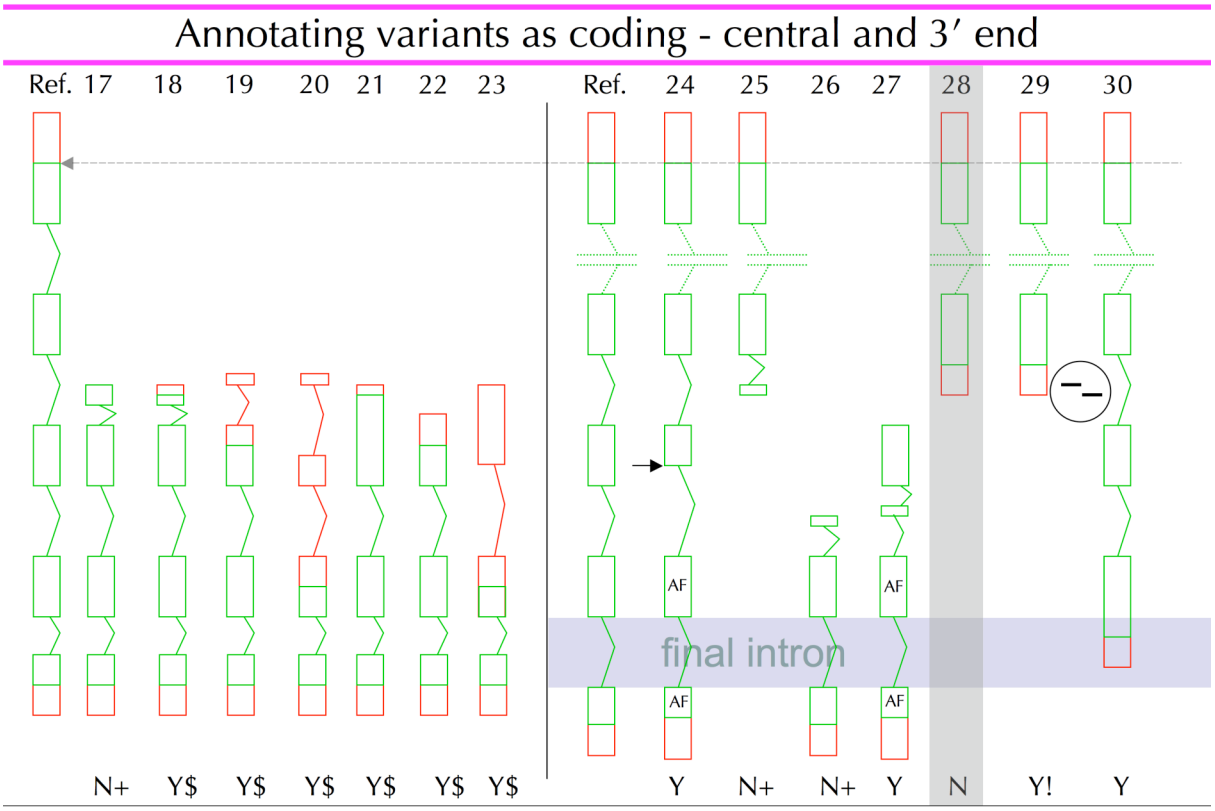


Figure 4: annotating variants as coding - central and 3' end

Using unsupported SwissProt evidence

Some SwissProt evidence for variants is not full-length or not at all supported by transcripts. In these cases check whether there is any literature support and follow *Figure 5* to decide on the use of this evidence.

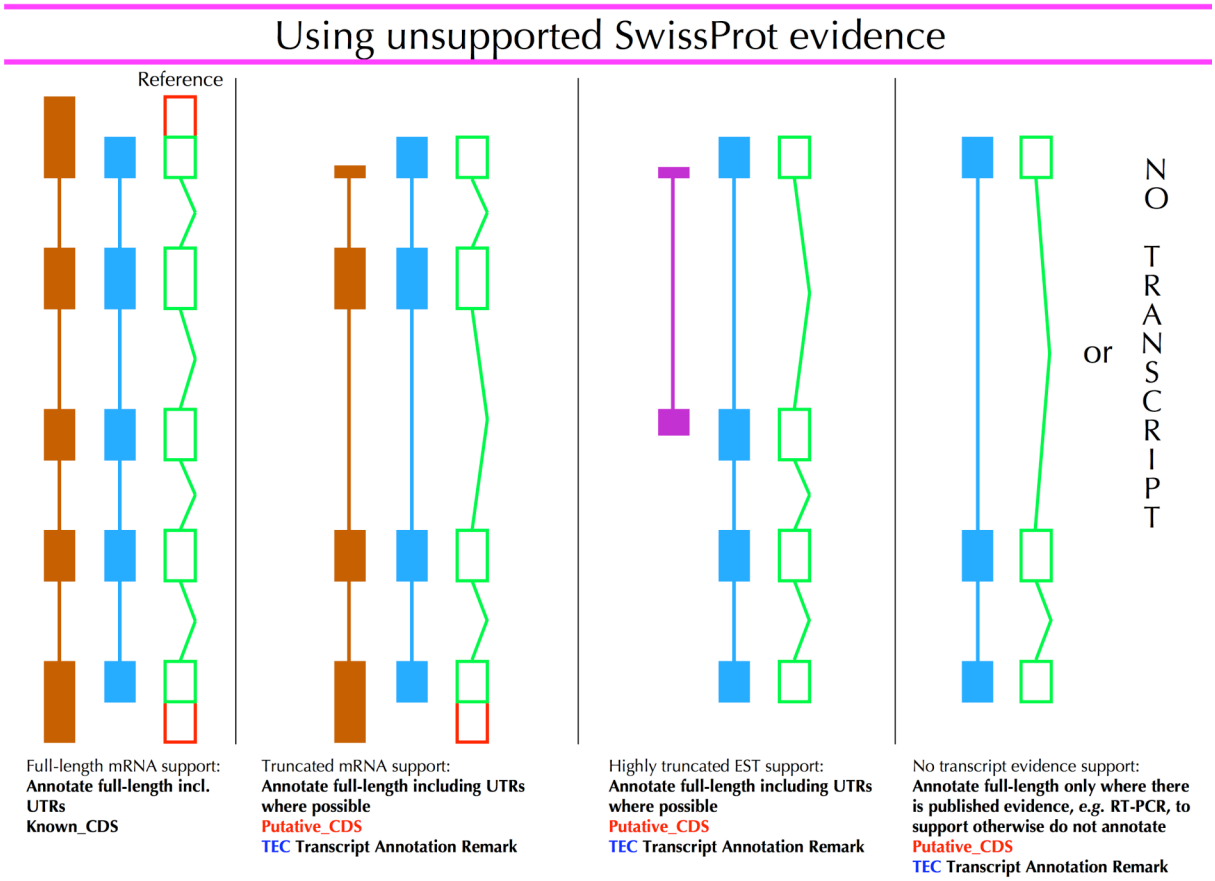


Figure 5: using unsupported SwissProt evidence

NOTE: some SwissProt evidence may be translations from cDNAs that are part of the 3′ UTR or that we annotate as retained intron transcripts. If that is the case ignore SwissProt evidence and contact SwissProt at hsf-curators@sanger.ac.uk to request removal of or the addition of a note to that entry.

Defining first and last coding introns

Novel exons lying within first and last coding introns are treated differently from novel exons in internal introns for a number of reasons. Protein structures are more tolerant of changes at their N- and C-termini so we are less likely to annotate CDSs incapable of folding if we include coding spice variants with novelty at the termini. A novel exon in the first intron may well be utilizing an alterative promoter, which are more likely clustered at the 5′ end of genes (see ***Figure 3***). A novel exon in the final intron is unlikely to be subject to NMD even if it lacks the polyA features to confirm its end.

When a novel internal exon is confirmed by EST/mRNA support, a CpG island (and circumstantially by CAGE or DiTag evidence), this creates a novel first intron where normal first intron annotation rules apply. Similarly, where a novel final exon is confirmed by polyA features, a novel final intron is created where normal final exon annotation rules apply. See ***Figure 6***.

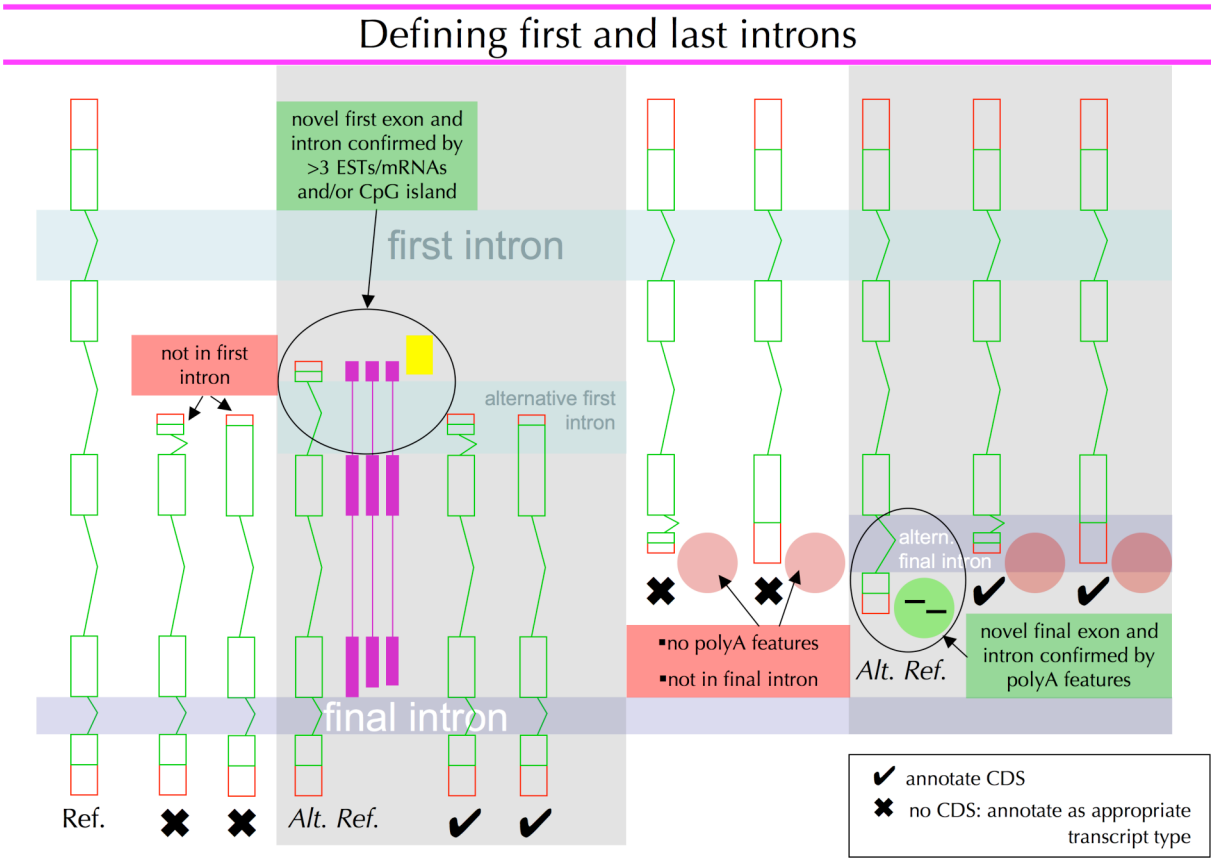
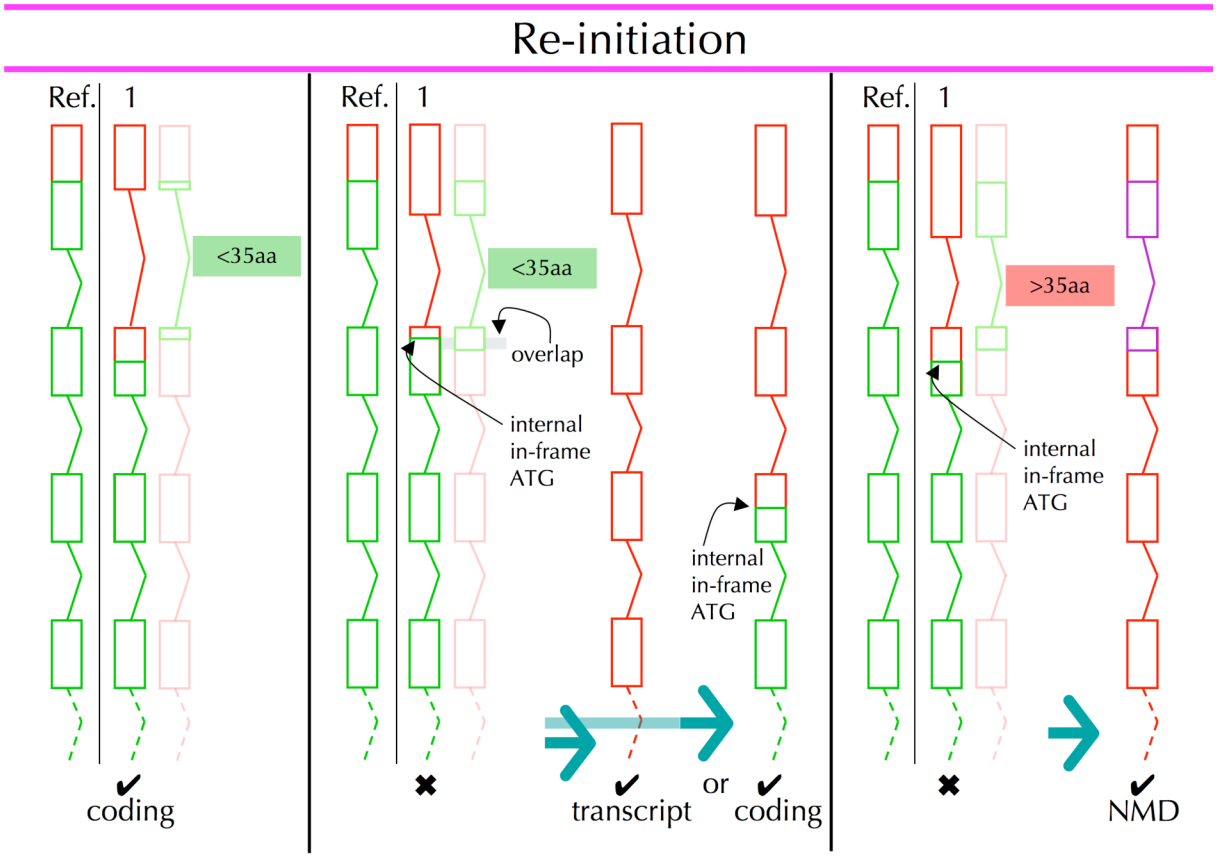


Figure 6: defining first and last introns

Re-initiation

Re-initiation is dependent on the length of the uORF, if the uORF >35aa then re-initiation will not occur and the variant should be annotated as NMD with a CDS starting from the ATG shared with the main variant (*Figure 7*). If the uORF < 35aa re-initiation will occur at the next ATG downstream of the stop codon. The ribosome cannot reverse to use ATGs even slightly upstream of the stop codon. If the next ATG is in frame with other coding variants at the locus annotate a CDS (most likely a putative_CDS). If the next ATG is out of frame and would lead to NMD annotate the variant as a transcript (as we do not have enough confidence that the ATG could initiate translation to annotate as NMD). The distance between the stop codon of the uORF and the ATG used is immaterial (it has been reported that the longer the distance the more efficient the re-initiation). To add uORFs we currently only use ATGs shared with other coding variants as these give a reasonable indication that the ATG is functional. uORFs initiating at ATGs upstream of shared ATGs should not be annotated.

NOTE: these rules do not apply to the main reference variant



Re-initiation is dependent on the length of the uORF: >35aa effectively prevents use of downstream ATG; distance to the downstream ATG (*i.e.* length of the "new" 5′ UTR) does not affect translation.

Figure 7: re-initiation

Classifying coding transcripts

The coding regions are classified as one of the following four categories depending on the evidence available. This applies to every coding transcript individually.

Known_CDS: 100% Identical to RefSeq NP or Swiss-Prot entry. Remember to check var_splic entries from SwissProt in Blixem.

Novel_CDS: shares >60% length with known CDS from RefSeq or Swiss-Prot or has cross-species/family support or domain evidence.

Putative_CDS: shares <60% length with known CDS from RefSeq or Swiss-Prot, or has an alternative first or last coding exon. Can be applied to a variant transcript as well as the sole transcript for a locus that has no variants.

Nonsense_mediated_decay: if there are one or more splice junctions >50bp downstream of the end of the CDS (using the appropriate reference CDS) the transcript is tagged as NMD (see ***Figure 8***). If the stop codon is <50bp from a splice site but there is another splice site further downstream (>50bp from stop), the variant is still NMD. If the variant does not cover the full reference CDS, annotate as NMD if NMD is unavoidable (*i.e.* no matter what the exon structure of the missing portion is, the transcript will be subject to NMD).

EXCEPTION: If a transcript looks like it is subject to NMD but publications, experiments, or conservation support the CDS then a coding transcript should be made and the following tag added:

Transcript Annotation Remark:
NMD_exception
[PMID <id>, <publication reference>]

NMD_exception
PMID 12345678, Wilming et al. (2007) Nature 501

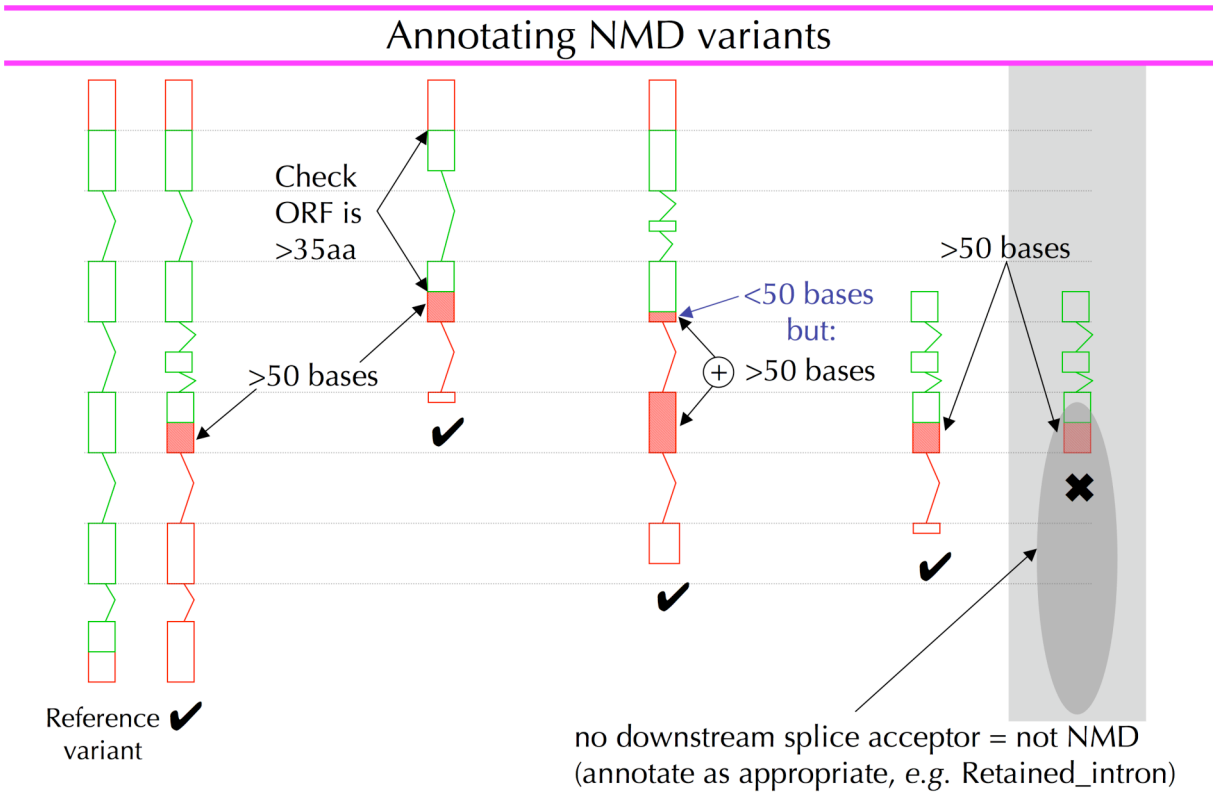


Figure 8: annotating NMD variants

Generally a transcript is considered a splice variant (and not a separate gene) when it shares at least one exon (or part thereof) with another variant. If the overlapping exons in the two transcript models have CDSs in different frames they should be annotated as separate loci.

If a variant has a novel first or last internal exon relative to a reference transcript and no polyA features, conservation, SwissProt, domains or paralog homology to support the putative CDS then annotate as a transcript. Reasoning: the transcript could be incomplete, the full-length version may be sharing more exons upstream/downstream with the reference, which would likely induce NMD.

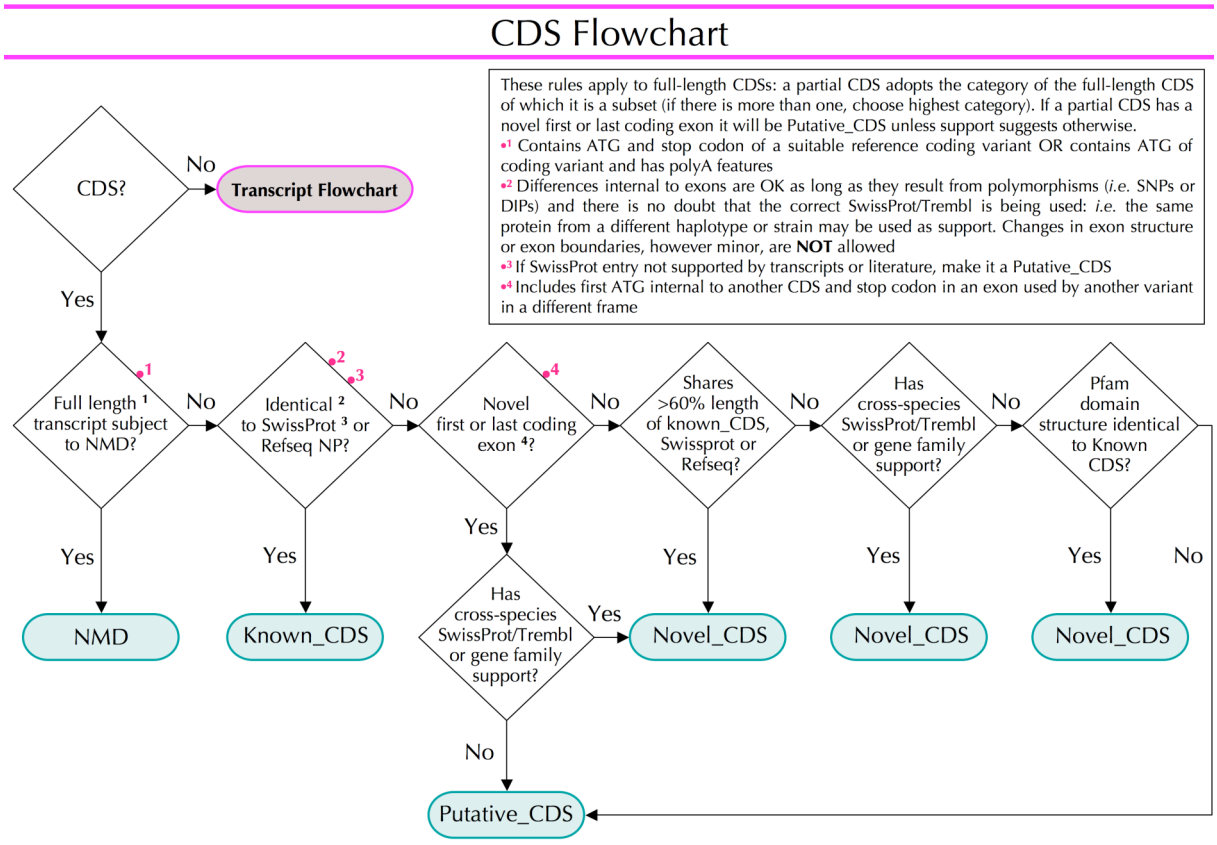


Figure 9: CDS decision graph

Two/three exon solitary gene objects should never have a CDS annotated unless it is a known gene or there is evidence (homology, domains, conservation). But see below.

Orphan proteins

Many independent transcripts (*i.e.* not part of another coding or non-coding locus) based on splicing mRNAs or ESTs contain at least one possible ORF. These transcripts might have biological function at the RNA level (*i.e.* lincRNAs) but there is a possibility that some of these ORFs encode functional proteins. Such ORFs are generally short, poorly conserved (not conserved beyond primates for human or beyond rat for mouse), lack paralogs and contain no functional domains (*e.g.* pfam).

A CDS should be annotated where the orphan protein is >50aa in length and there are no other possible CDSs/ORFs that would interfere with the translation of the proposed orphan CDS. The CDS may be contained within the transcript or open-ended at one or both ends. See ***Figure 10***. An annotated orphan protein may be tagged as Known_, Novel_ or Putative_CDS depending on the supporting evidence (SwissProt, RefSeq).

Add the following remark:

Locus Annotation Remark:
orphan

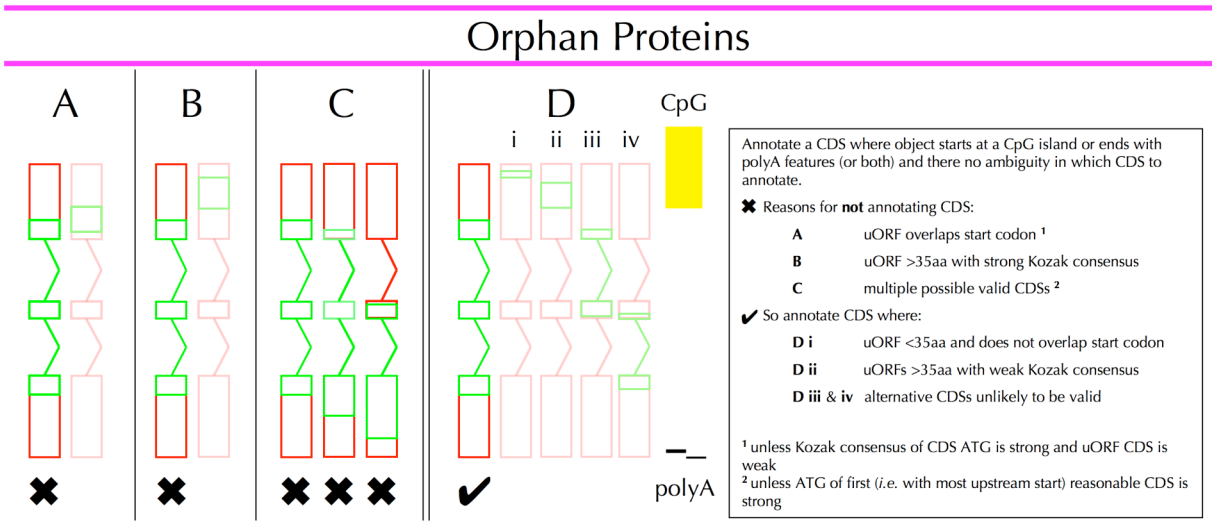


Figure 10: orphan proteins

Selenocysteine proteins

Nonsense codon TGA can encode selenocysteine in certain proteins by using tRNAs with a UCA anticodon carrying selenocysteine. The following comments should be added to each selenocysteine transcript and locus, but only when the presence of selenocysteine is known from the SwissProt entry:

Transcript Annotation Remark:
selenocysteine

Locus Visible Remark:
selenoprotein

Defining untranslated regions and polyA features

5′ UTRs are extended as far upstream as species specific spliced ESTs and cDNAs allow. For variants that share an identical CDS but have alternative 5′ UTRs use following remark:

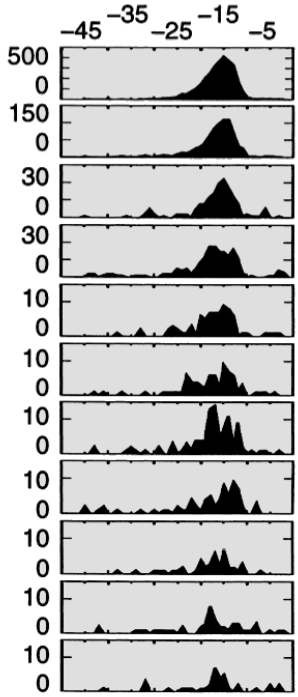
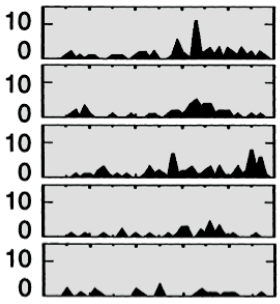
Transcript Visible Remark:
alternative_5′_UTR

3′ UTRs are extended to the furthest downstream genomically encoded nucleotide (*i.e.* before start of the polyA tail) (*Figure 11*). Annotate polyA signals (according to *Table 3*) up to approximately 50 bp upstream of the polyA site. Gaps between spliced evidence and the cluster of polyA containing 3′ ESTs typically seen at the 3′ end are allowed. When multiple polyA signals are associated with the same site annotate the most common signals (AATAAA or ATTAAA) over rare ones. Multiple discrete sets of polyA features (*i.e.* polyA site with corresponding pA signal) are annotated, but the gene is

stretched to the downstream-most set (unless a specific splice variant is associated with a specific polyA feature set). Often there is polyA_site "wobble" where the exact position varies by a few nucleotides. In this case annotate at minimum the downstream-most.

NOTE: polyA signals are never annotated in isolation, only combined with polyA sites. On the other hand, polyA sites can be annotated in absence of a polyA signal.

Table 3: variation in polyA signals and their frequency in humans (Beaudoing et al. 2000)

Hexamer	Observed (expected)[a]	% sites	*P*[b]	Position average ± SD	Location[c]
Most Significant Hexamers in 3′ Fragments: Clustered Hexamers					
AAUAAA	3286 (317)	58.2	0	−16 ± 4.7	
AUUAAA	843 (112)	14.9	0	−17 ± 5.3	
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-18}	−17 ± 5.9	
GAUAAA	72 (21)	1.3	2×10^{-18}	−18 ± 6.9	
AAUAUA	96 (33)	1.7	2×10^{-19}	−18 ± 6.9	
AAUACA	70 (16)	1.2	5×10^{-23}	−18 ± 8.7	
AAUAGA	43 (14)	0.7	1×10^{-9}	−18 ± 6.3	
AAAAAG	49 (11)	0.8	5×10^{-17}	−18 ± 8.9	
ACUAAA	36 (11)	0.6	1×10^{-08}	−17 ± 8.1	
Most Significant Hexamers in 3′ Fragments: Scattered Hexamers					
AAGAAA	62 (10)	1.1	9×10^{-28}	−19 ± 11	
AAUGAA	49 (10)	0.8	4×10^{-18}	−20 ± 10	
UUUAAA	69 (20)	1.2	3×10^{-18}	−17 ± 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	

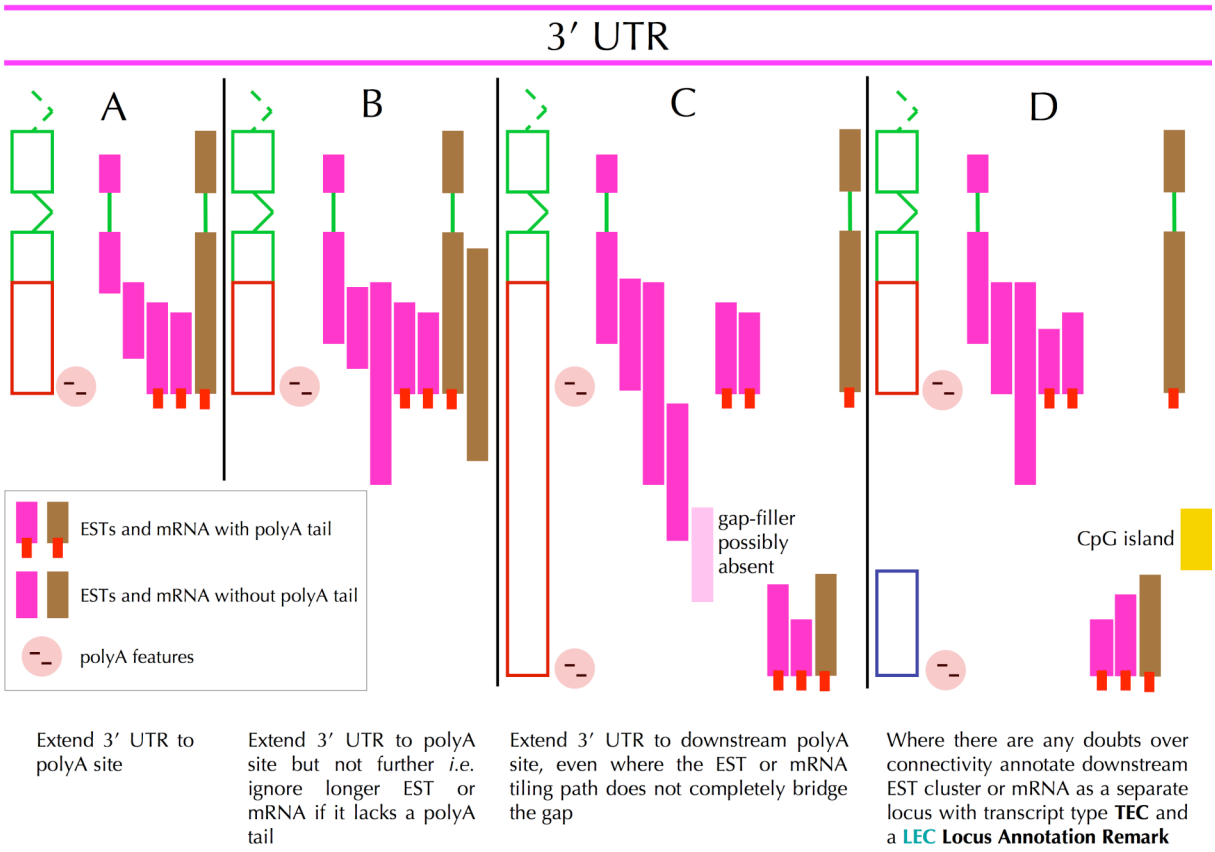


Figure 11: 3' UTR annotation

Transcripts without CDS

Non-coding transcripts (*i.e.* no CDS is annotated because it would not fulfil CDS criteria mentioned earlier) are labelled with one of the following tags (*Figure 12*).

NOTE: a 2-exon EST that overlaps with UTR or any exon but doesn't share splices becomes a variant of the locus.

Transcript: if the transcript does not fit any of the sub-categories below it's tagged as an un-typed transcript.

Non_coding: for known non-coding RNA from literature or experiments, *e.g.* Leu2. Include references in Transcript Annotation Remark.

Retained_intron: the transcript has retained intronic sequence compared to a reference variant and there is no believable evidence such as alternative ATG or polyA features or strong cross-species stop codon conservation that this is genuine. Any variant with a retained intron should be tagged as Retained_intron, unless the entire retained intron is open and in-frame with the flanking coding exons. Where the first or last "coding" intron (relative to a suitable reference) is retained consult *Figure 6*.

EXCEPTION: if another variation upstream of retained intron induces NMD, tag it NMD.
EXCEPTION: if the retained intron is the last intron and gives rise to a novel stop, tag it Putative_CDS.
EXCEPTION: if the retained intron is in a UTR intron and thus doesn't affect CDS, annotate as coding.
EXCEPTION: if the retained intron is only supported by other species evidence, don't annotate (unless annotation depends on evidence from closely related species, *e.g.* human transcripts in gorilla).

Antisense: only used for a known (from literature or experiments) antisense mRNA, *e.g.* TSIX. Just overlapping or nested transcripts on opposite strands is not sufficient. Include references in Transcript Annotation Remark.

Putative: 2-3 exon transcript supported by only 1-2 ESTs.

IG_gene: only for Immunoglobulin gene building blocks.

Transcripts that are not fully supported (whole length) by species-specific evidence are labelled as non-organism supported as follows:

Transcript Annotation Remark:
non-organism_supported

Transcript Flowchart

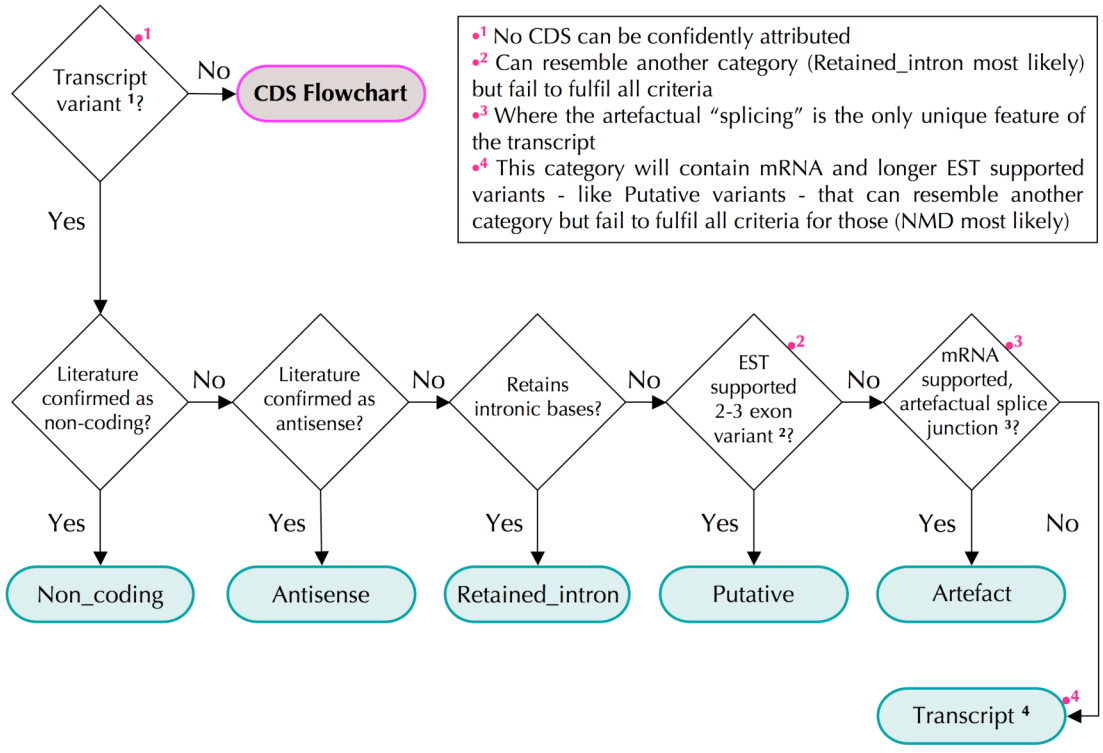


Figure 12: transcript decision graph

Pseudogenes

We divide pseudogenes into five categories with processed and unprocessed pseudogenes each further divided into two sub-categories: transcribed and untranscribed. Most pseudogenes are clear, with homology to existing proteins but containing a disrupted CDS (frameshifts, in-frame stop codons) and having one or more active parent genes. We only annotate the extent of the protein match and they are built wherever there is a recognizable non-spurious match. Unless the whole parent protein matches, use Dotter to check for a more complete alignment.

NOTE: Pseudogenes with a known locus symbol are not tagged "Known", but the approved symbol and description is used.

Processed_pseudogene

Because they are made from reverse-transcribed processed mRNA transposed into the genome, processed pseudogenes don't have the exon structure of the parent gene anymore and are therefore single exon. However, this single exon may be interrupted by repeat sequences (LINEs and SINEs) or even other processed pseudogenes inserted into it, giving the appearance of splicing. Such insertions should not be included in the annotated pseudogene. Processed pseudogenes often have a recognizable remnant of the polyA tail integrated into the genome. Add the corresponding "Pseudo-polyA signal" to indicate incorporation of the tail (only use most common signals AATAAA or ATTAAA). Sometimes processed pseudogenes have an intact CDS similar or even identical to their unprocessed parent, in which case it is still made into a pseudogene unless there is locus-specific transcription evidence.

Unprocessed_pseudogene

Unprocessed pseudogenes still have their exon structure because they are produced as a result of gene or genomic duplication. As a consequence they often appear in a cluster with their active parent genes (*e.g.* histones, olfactory receptors). They may actually be single exon, if their parents are single exon or have a single exon CDS. If the parent is a single exon gene (e.g. olfactory receptors) and the pseudogene has a slightly 5′ or 3′ truncated CDS (compared to other family members), check for missing or truncated domains to determine pseudogene status. These instances always occur in clusters and the pseudogenes are unprocessed because they arose from genomic duplication, not retrotransposition.

Transcribed_processed_pseudogene & **Transcribed_unprocessed_pseudogene**

Sometimes protein homologies unequivocally point to a locus being a pseudogene, but overlapping locus-specific transcription evidence indicates transcription. In that case annotate a pseudogene object (as first variant) and a transcript object under the same locus, the former tagged with the appropriate Transcribed_....._pseudogene tag and the latter with Transcript.

Polymorphic_pseudogene

If owing to a deleterious SNP/DIP the locus being annotated is a pseudogene, but it is known that in other individuals/haplotypes/strains the gene is translated, the gene is labelled Polymorphic_pseudogene. Only used if a known polymorphism (look in Ensembl/UCSC) or if there is transcriptional support for both versions of the locus (*i.e.*

cDNAs/ESTs that contains the SNP/DIP and ones that disagree with the genomic sequence at the SNP/DIP position and have an intact CDS).

WARNING: Genoscope mRNAs are modified to correspond to genomic sequence so should not be trusted in deciding whether the locus is polymorphic or not.

Unitary_pseudogene

A pseudogene for which the ortholog is a coding gene in another reference species. It doesn't have a parent in that it hasn't arisen from recent duplication: it was generated from a deleterious mutation in a previously functional coding gene. These are generally unprocessed pseudogenes and they can actually have more than one "orthologous parent". For example certain gene families (*e.g.* Mup's, Vnr's) have expanded in rodents and at the syntenic position in human the sole representation of the gene family is one or more pseudogenes.

NOTE: requires in-depth conservation analysis or strong published evidence that this is a species-wide pseudogenization event and not a polymorphism. Check that it is not a known confirmed SNP.

IG_pseudogene

Special category for pseudogene versions of Immunoglobulin gene building blocks.

Supporting evidence

In case of splice variation, the main variant gets variant specific evidence plus non-specific evidence (*i.e.* evidence that supports multiple variants). For remaining variants use only variant specific evidence for each transcript and do not re-use the non-specific evidence for multiple transcripts at the same locus. If there is a lot of evidence, select only locus/species specific evidence and only add ESTs if they extend 5'UTR (must splice) or support 3' UTR or polyA features (don't need to splice).

Also check var_splic annotation in SwissProt (entries are visible in Blixem).
Where appropriate add the ids of the supporting evidence to the variants (note the RefSeq protein id for cases that lack SwissProt):

Transcript Visible Remark:
<IDs of variant-specific evidence>

1234567H02Rik
FLJ12345, KIAA1234
DKFZp123E4567Q8, MGC:12345, NP_123456

Multipart genes

Within a contiguous region

If homologies are too weak or incomplete to resolve large gaps in homology (suggesting missing exons), the gene is annotated as a set of separate objects, numbered preferably in consecutive order, with the same gene (locus) name. A note in the objects should point to the fact that these fragments belong to one gene. Be sure that the fragmented homologies are in the correct order and not duplicated (i.e. the same homologies pop up on more than one place on the genome, indicating a gene duplication or multiplication). If the gene spans more than one clone, the most 3' fragment's locus name will be used as the locus name for all fragments, but each fragment will have its own unique transcript name.

For human and mouse add LEC (**L**ocus for **E**xperimental **C**onfirmation) to the locus remark to flag the locus for possible future experimental completion.

Transcript Visible Remark:

gene fragments <this transcript name> and <other transcript name> [and <other transcript name>] are part of the same gene; the exact exon structure linking the fragments is yet to be determined.

> gene fragments RP23-123H10.3-001 and RP23-123H.10.4-001 and RP23-11B11.1-001 are part of the same gene; the exact exon structure linking the fragments is yet to be determined

Locus Annotation Remark:

fragmented_locus

LEC

Spanning a gap

If homologies are fine but you can't make a complete transcript because one or more exons are missing owing to a gap in the assembly or a mis-assembly, use the following:

Transcript Visible Remark:

gene fragments <this transcript name> and <other transcript name> [and <other transcript name>] are part of the same gene; an assembly gap between them contains one or more exons.

> gene fragments RP24-11A2.9-001 and RP23-123H10.3-001 and RP23-99D8.1-001 are part of the same gene; an assembly gap between them contains one or more exons

Locus Annotation Remark:

fragmented_locus

Transcripts that span a gap but are complete (*i.e.* the gap does not contain exons) are annotated as one-piece transcripts across the gap(s) without any of the above remarks.

Variants

We use the term variants to describe different alternative splicing events at the same locus. The minimum requirement for two objects to be classed as variants of each other is that they share at least one exon or part thereof. In general, variants are only annotated to the extent of their supporting evidence (EST, mRNA). This is because there is a (small) chance that a variant has more than one alternative event, possibly outside the homology.

Any partial CDS (*i.e.* start not found and/or end not found) that follows the reference CDS needs to be annotated, however small: even if it is just one amino-acid. This mostly applies to UTR variants.

When encountering a NAGNAG variant, an in-frame type of variation where at the acceptor site some variants splice after the first AG and others after the second AG (see ***X1*** below), it is advisable to add a Transcript Annotation Remark to that effect in all the affected variants. This makes it easier when the gene is revisited later because the affected variants look identical unless zoomed in very far.

3′(aceptor)		3′(acceptor)
intron-nnn**CAG**|**CAG**nnn-exon	&	intron-nnn**CAGCAG**|nnn-exon
intron-nnn**CAG**|**TAG**nnn-exon	&	intron-nnn**CAGTAG**|nnn-exon
intron-nnn**TAG**|**TAG**nnn-exon	&	intron-nnn**TAGTAG**|nnn-exon
intron-nnn**TAG**|**CAG**nnn-exon	&	intron-nnn**TAGCAG**|nnn-exon

***X1**: examples of the most common NAGNAG variations; any other combination is permissible (hence NAGNAG), for example AAGTAG etc., but combinations other than the ones shown above are not seen very often.* | denotes intron-exon boundary (splice junction)

Canonical splice sites

Check that splicing follows consensus splice sites. The LogoGraph below (***Figure 13***) shows the frequency of occurrence of different bases at key positions. But shown first in ***X2*** is a list of most commonly observed splice sites and uncommon ones not visible in the figure (G|GC and |AT-AC|).

NOTE: 5′ signals shown below can occur in any combination with the 3′ signals, except the AT-AC pair, which only occur as a pair.

	5′(donor)	3′(acceptor)	
more common ^	exon-nnn**G**|**GT**nnn-intron		
|	exon-nnn**A**|**GT**nnn-intron		
|	exon-nnn**T**|**GT**nnn-intron		
|	exon-nnn**C**|**GT**nnn-intron		
less common |	exon-nnn**G**|**GC**nnn-intron		
		intron-nnn**CAG**|nnn-exon	^ more common
		intron-nnn**TAG**|nnn-exon	|
		intron-nnn**AAG**|nnn-exon	|
		intron-nnn**GAG**|nnn-exon	| less common

exon-----|**AT**----intron-----**AC**|----exon (in combination only)

***X2**: examples of the most common splice donors and acceptors and some that are not visible in the LogoGraph* | denotes intron-exon boundary (splice junction)

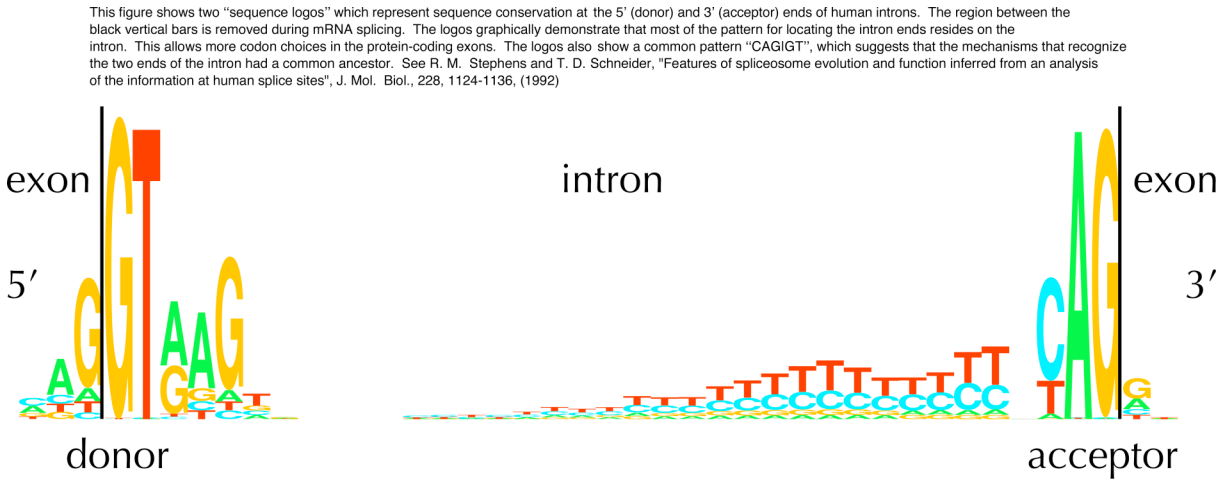


Figure 13: splicing LogoGraph

Non-canonical splice sites

If a splice site doesn't fit the above canonical pattern we are dealing with non-canonical splice sites. Sometimes both donor and acceptor are affected, sometimes only one or the other. If the non-canonical event is conserved in another reference species or it is a published event, this can be used. Add a remark indicating the presence and location of non-canonical splice sites:

Transcript Visible Remark:

non-canonical splice sites between exons <exon #> and <exon #>

non-canonical splice sites between exons 3 and 4

non-canonical splice donor site between exons <exon #> and <exon #>

non-canonical splice donor site between exons 5 and 6

non-canonical acceptor splice sites between exons <exon #> and <exon #>

non-canonical splice acceptor site between exons 1 and 2

TEC

If you annotate a variant with a CDS that breaks protein domain structure or is otherwise very different (truncated) from the reference CDS, add a TEC (**T**ranscript for **E**xperimental **C**onfirmation) transcript annotation remark to flag it for possible future experimental confirmation of expression and investigation of expression pattern:

Transcript Annotation Remark:

TEC

Locus-spanning (readthrough) transcripts and nested genes

Readthrough
A few loci in mouse and human have approved separate locus names for the readthrough transcripts, for example Cbx6-Nptxr. The gene description is like "Cbx6-Nptxr readthrough transcript". Use these symbols and descriptions. So in these cases the loci are annotated as three separate loci: upstream, downstream and readthrough.

In cases where there is no approved separate readthrough locus, follow the flowcharts below (***Figure 14***, ***Figure 15***).

In case of overlapping loci, the following Locus Annotation Remark is added to all the loci involved:

Locus Annotation Remark:
overlapping_locus

---

and the following Transcript Annotation Remark to any transcript that overlaps two or more loci:

Transcript Annotation Remark:
readthrough

Readthrough Flowchart 1

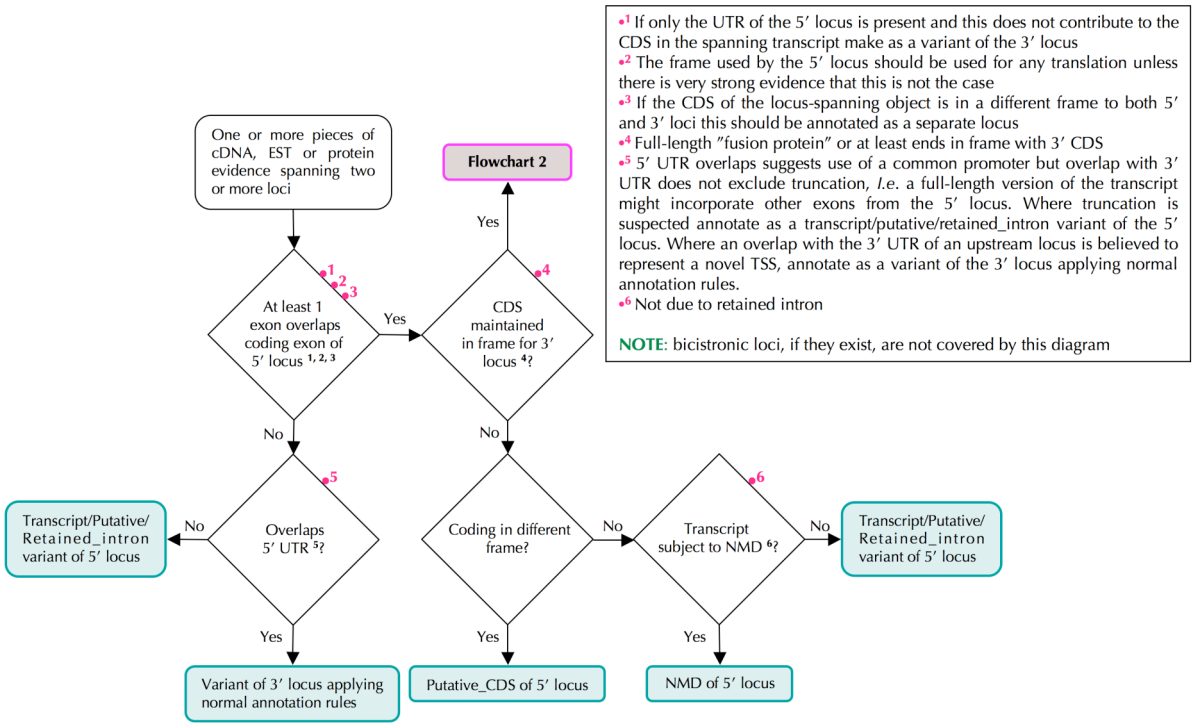


Figure 14: readthrough flowchart 1

Readthrough Flowchart 2

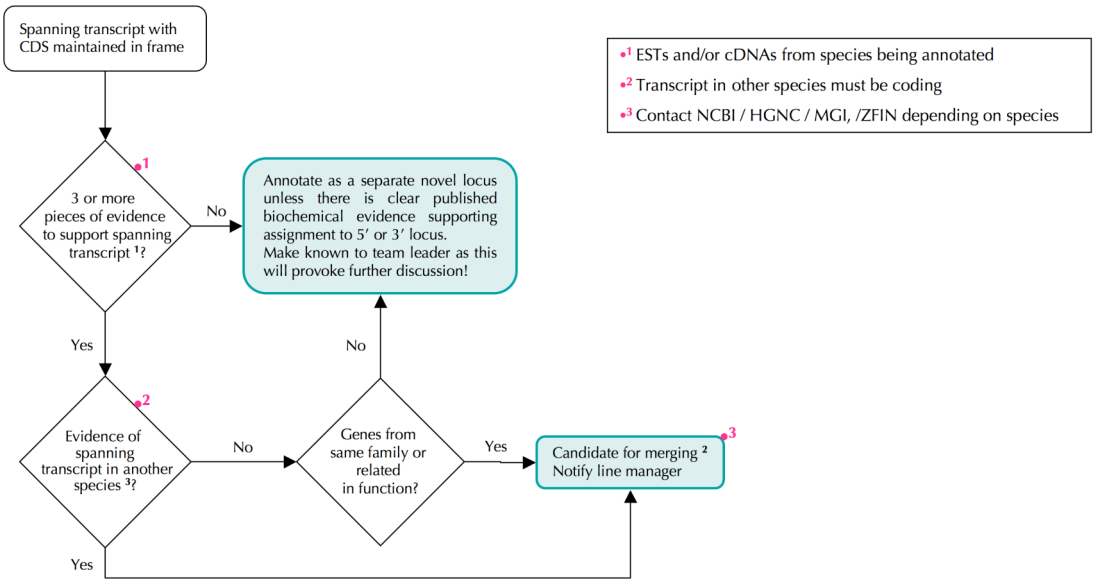


Figure 15: readthrough flowchart 2

Nesting

Transcripts that (partially) reside inside other transcripts on the same strand, whether entirely within an intron, spread over a number of introns, or partially in introns, partially outside the other transcripts, are considered separate loci if they do not overlap on the exon level. See ***Figure 16***. If there is overlap, even if no shared splice junctions, the transcript is actually a variant of the reference locus.

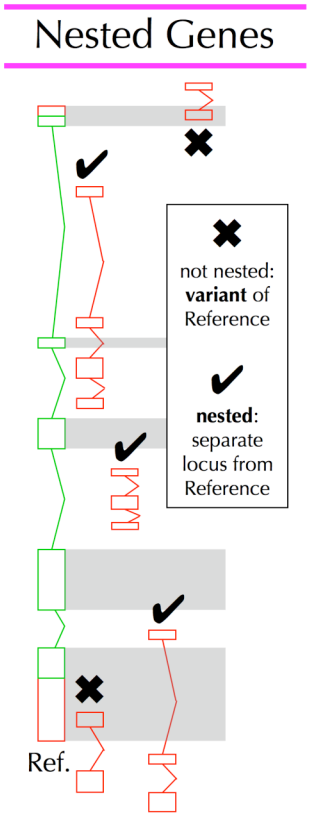


Figure 16: nested genes as separate loci

Naming Genes

Known named genes

The locus of a newly annotated gene that is identical to a known gene is named after the approved symbol for that gene if available in Entrez Gene and the approved gene name is used for the description (Full_name). Interim symbols can be used, but symbols such as accession numbers, Riken, KIAA or FLJ identifiers, genetic marker names, etc. (often found as approved symbols in mouse) are not acceptable as locus names. For human the only irregular symbol we use is the Corf type (*e.g.* C21orf56).

Below the locus **Symbol** is only shown when it needs to be changed. If not shown keep the automatically generated symbol.

Symbol TAP1
Full name transporter 1, ATP-binding cassette, sub-family B
Known ✔

Known anonymous genes

A novel gene that's not really novel but a known gene from mass screening projects, like the Japanese KIAA and FLJ type genes and the German DKFZ ones. Use any helpful information available (pfam domains or families).

Full name novel protein (FLJ10034)
novel protein (KIAA0023)
novel C2H2 type zinc finger protein (KIAA0109)
Known ✔

Extension of known anonymous gene

Sometimes the newly annotated gene extends a known anonymous gene considerably (*i.e.* several more exons), and may even link up two or more separate known gene fragments.

Full name novel zinc finger protein (contains KIAA1234 and KIAA0090)
Known ✔

Known genes with non-approved symbols

Sometimes a human gene has a provisional symbol or what looks like a proper symbol but not HGNC approved. Don't use the symbol, but use the description when it is identical to what's used in an approved mouse gene name or consistently used in a number of other species. Otherwise follow normal rules for naming.

Full name selenoprotein M (SELM) *unapproved symbol SELM, consistent with many species*
novel protein kinase-like protein *unapproved symbol SGK493; Pkdcc in mouse*
Known ✔

Homologous genes

A gene product based on homology to a known protein is named after the best homology if possible, or after the broader family (with description copied from the pfam hit, if available).

Full name novel protein similar to adenine synthetase 3 AS3
novel serine/threonine kinase
novel histone H2a family protein

Occasionally there is good reason to believe the gene is the orthologue of a known gene in another species (*i.e.* very high cross-species homology to that one type of protein from different species), in which case it is acceptable to call it a possible orthologue. This usually applies only to homology between rodent and human genes.

Full name novel protein, possible orthologue of rodent adenylcyclase 5 like Ac5l

Homology to model organism predicted/hypothetical genes

Occasionally the only homology detected is to a series of hypothetical proteins from model organisms, usually from genomic sequencing projects of *C. elegans, D. melanogaster, S. cerevisiae, S. pombe, A. thaliana* and scores of pathogens.

Full name novel protein

Novel genes with non-informative matches

For a gene based on just ESTs or anonymous mRNAs (not from one of the large cDNA sequencing projects).

type "Novel CDS" or "Putative CDS":
Full name novel protein

type "Transcript" or "Putative":
Full name novel transcript

Pseudogenes

Pseudogenes are named after the gene that is obviously the parent or, if that cannot be determined, after the general family. A pseudogene of an anonymous non-informative gene is a "novel pseudogene".

Full name 60S ribosomal protein L17 (RPL17) pseudogene
C2H2 zinc finger protein pseudogene
novel pseudogene

Known pseudogenes take their given description and symbol (but are not tagged "Known"!).

Symbol ASSP9
Full name argininosuccinate synthetase pseudogene 9

DE (Description) Lines

In the DE line of a genomic clone, the genes are generally listed in the order in which they appear. The basic format is "the <locus symbol> gene for <locus full name>". This information is automatically generated when clicking the "Generate" button in the clone editing window. However, the text may need to be edited slightly to conform to the required format. Below are a few points to look out for (highlighted in the example). Genes with un-informative locus full names like "novel protein" or novel transcript" are labelled with the "novel gene" moniker and should be enumerated where necessary and if already enumerated the number digit replaced with the number word. Loci with descriptions that are identical save for the member number or subfamily identifier can be grouped with the full description only used once (see example below). Genes that are not completely on the genomic clone but have their end on it are prefixed with the appropriate qualifier ("the 5′ end", "the 3′ end"). The auto-generated text will only print "part of", irrespective of whether the gene has indeed only internal exons or has an end on the clone. Pseudogenes with official symbols will need editing to the format shown in the example. Where necessary the "a" needs to be replaced with "an" (*i.e.* in front of words starting as pronounced with vowels: A, E, I, O, U, X, or (letter only) F, H, L, M, N, R, S). Finally, if a clone only contains intronic sequences then the automatically generated reference to that gene ("part of ….") needs to be removed. Also any reference to "artefact gene" needs to be removed.

> Contains **the 5′ end** of the HIRA gene for HIR histone cell cycle regulation defective homolog A (S. cerevisiae), **three** novel genes, **an** ATP-binding cassette, subfamily A (ABC1), member 6 (ABCA6) pseudogene, **olfactory receptor, family 1, subfamily R, member 1 pseudogene OR1R1P**, a novel pseudogene, a gene for a novel protein similar to SH3-domain GRB2-like 3 SH3GL3, **the RASGRP1, RASGRP2 and RASGRP3 genes for RAS guanyl releasing protein 3 (calcium and DAG-regulated) 1, 2 and 3** and **the 3′ end** of the gene for a novel phosphoinositide-3-kinase (PIK3) family.

Here are some examples of auto-generated DE lines, with parts to be edited underlined, preceded by a description of the genes they contain.

5′ end of HIRIP3 + novel protein + novel transcript + actin, beta pseudogene 8 ACTBP8

```
Contains a actin, beta pseudogene 8, part of the HIRIP3
gene for HIRA interacting protein 3 and 2 novel genes.
```

beta-2-microglobulin (B2M) pseudogene + intron of HIRIP3

```
Contains part of the HIRIP3 gene for HIRA interacting
protein 3 and a beta-2-microglobulin(B2M) pseudogene.
```

3′ end of HIRIP3 + SHROOM1 + SHROOM2 + SHROOM3

```
Contains the SHROOM1 gene for shroom family member 1,
the SHROOM3 gene for shroom family member 3, part of
the HIRIP3 gene for HIRA interacting protein 3 and the
SHROOM2 gene for shroom family member 2.
```

Reference Tables, Figures and Lists

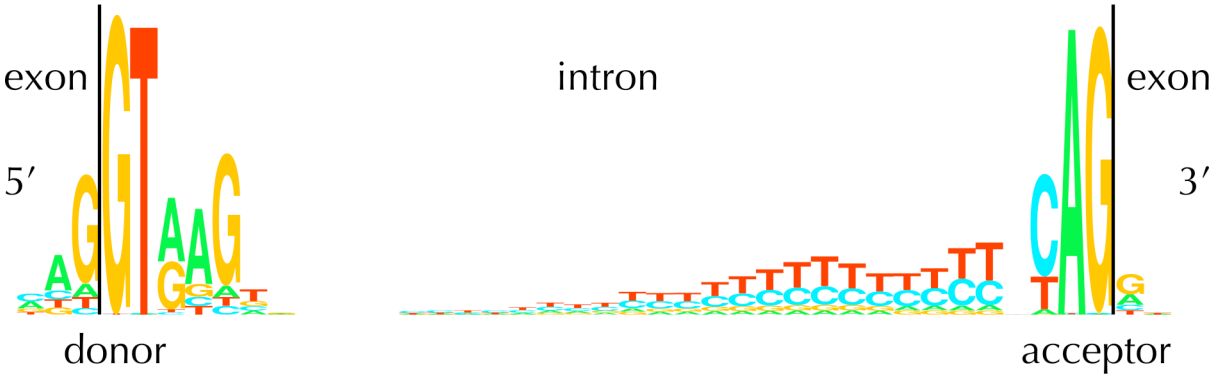
Codon table

	non-polar	polar	basic	acidic	(stop codon)

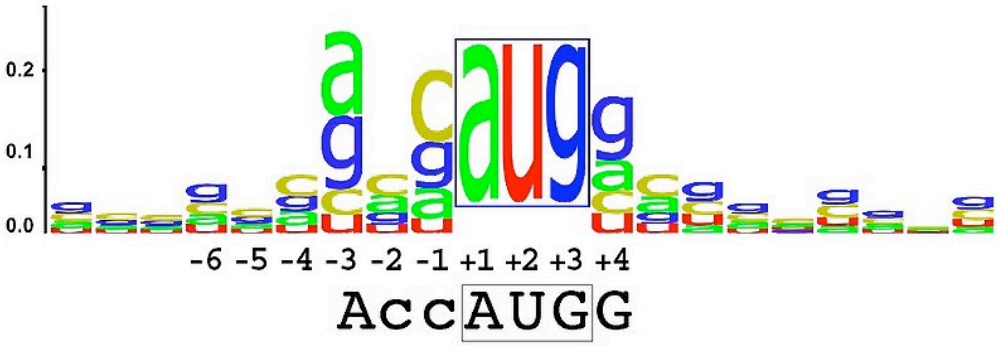
1st base	2nd base: T	C	A	G
T	**TTT** (Phe) Phenylalanine (F)	**TCT** (Ser) Serine (S)	**TAT** (Tyr) Tyrosine (Y)	**TGT** (Cys) Cysteine (C)
	TTC (Phe) Phenylalanine (F)	**TCC** (Ser) Serine (S)	**TAC** (Tyr) Tyrosine (Y)	**TGC** (Cys) Cysteine (C)
	TTA (Leu) Leucine (L)	**TCA** (Ser) Serine (S)	**TAA** Ochre (*Stop*) (*)	**TGA** Opal (*Stop*) (*)
	TTG (Leu) Leucine (L)	**TCG** (Ser) Serine (S)	**TAG** Amber (*Stop*) (*)	**TGG** (Trp) Tryptophan (W)
C	**CTT** (Leu) Leucine (L)	**CCT** (Pro) Proline (P)	**CAT** (His) Histidine (H)	**CGT** (Arg) Arginine (R)
	CTC (Leu) Leucine (L)	**CCC** (Pro) Proline (P)	**CAC** (His) Histidine (H)	**CGC** (Arg) Arginine (R)
	CTA (Leu) Leucine (L)	**CCA** (Pro) Proline (P)	**CAA** (Gln) Glutamine (Q)	**CGA** (Arg) Arginine (R)
	CTG (Leu) Leucine (L)	**CCG** (Pro) Proline (P)	**CAG** (Gln) Glutamine (Q)	**CGG** (Arg) Arginine (R)
A	**ATT** (Ile) Isoleucine (I)	**ACT** (Thr) Threonine (T)	**AAT** (Asn) Asparagine (N)	**AGT** (Ser) Serine (S)
	ATC (Ile) Isoleucine (I)	**ACC** (Thr) Threonine (T)	**AAC** (Asn) Asparagine (N)	**AGC** (Ser) Serine (S)
	ATA (Ile) Isoleucine (I)	**ACA** (Thr) Threonine (T)	**AAA** (Lys) Lysine (K)	**AGA** (Arg) Arginine (R)
	ATG (Met) Methionine (M)	**ACG** (Thr) Threonine (T)	**AAG** (Lys) Lysine (K)	**AGG** (Arg) Arginine (R)
G	**GTT** (Val) Valine (V)	**GCT** (Ala) Alanine (A)	**GAT** (Asp) Aspartic acid (D)	**GGT** (Gly) Glycine (G)
	GTC (Val) Valine (V)	**GCC** (Ala) Alanine (A)	**GAC** (Asp) Aspartic acid (D)	**GGC** (Gly) Glycine (G)
	GTA (Val) Valine (V)	**GCA** (Ala) Alanine (A)	**GAA** (Glu) Glutamic acid (E)	**GGA** (Gly) Glycine (G)
	GTG (Val) Valine (V)	**GCG** (Ala) Alanine (A)	**GAG** (Glu) Glutamic acid (E)	**GGG** (Gly) Glycine (G)

Splicing

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAGIGT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



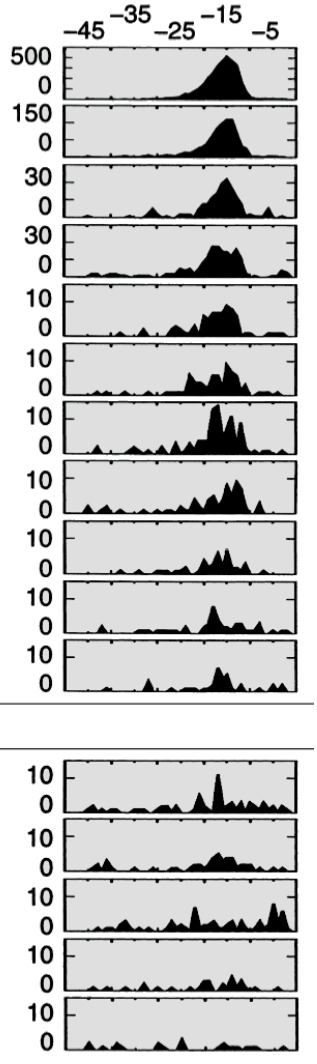
Start codon Kozak sequence



A at **-3** = ***strong***
G at **-3** plus **G** at **+4** = ***strong***
Anything else = ***weak***

PolyA signals

Hexamer	Observed (expected)[a]	% sites	*P*[b]	Position average ± SD	Location[c]
Most Significant Hexamers in 3′ Fragments: Clustered Hexamers					
AAUAAA	3286 (317)	58.2	0	−16 ± 4.7	
AUUAAA	843 (112)	14.9	0	−17 ± 5.3	
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-18}	−17 ± 5.9	
GAUAAA	72 (21)	1.3	2×10^{-18}	−18 ± 6.9	
AAUAUA	96 (33)	1.7	2×10^{-19}	−18 ± 6.9	
AAUACA	70 (16)	1.2	5×10^{-23}	−18 ± 8.7	
AAUAGA	43 (14)	0.7	1×10^{-9}	−18 ± 6.3	
AAAAAG	49 (11)	0.8	5×10^{-17}	−18 ± 8.9	
ACUAAA	36 (11)	0.6	1×10^{-08}	−17 ± 8.1	
Most Significant Hexamers in 3′ Fragments: Scattered Hexamers					
AAGAAA	62 (10)	1.1	9×10^{-28}	−19 ± 11	
AAUGAA	49 (10)	0.8	4×10^{-18}	−20 ± 10	
UUUAAA	69 (20)	1.2	3×10^{-18}	−17 ± 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	



Controlled vocabulary remarks

Transcript Annotation Remark:
upstream_ATG–<distance in aa upstream>

upstream_ATG-10

Transcript Annotation Remark:
NMD_exception
[PMID <id>, publication reference]

NMD_exception
PMID 12345678, Wilming et al. (2007) Nature 447

Transcript Annotation Remark:
selenocysteine

Locus Visible Remark:
selenoprotein

Transcript Visible Remark:
alternative_5′_UTR

Transcript Annotation Remark:
non-organism_supported

Transcript Annotation Remark:
non-best-in-genome evidence

Locus Annotation Remark:
fragmented_locus

Transcript Visible Remark:
gene fragments <this transcript name> and <other transcript name> [and <other transcript name>] are part of the same gene; the exact exon structure linking the fragments is yet to be determined.

gene fragments RP23-123H10.3-001 and RP23-123H.10.4-001 and RP23-11B11.1-001 are part of the same gene; the exact exon structure linking the fragments is yet to be determined.

gene fragments <this transcript name> and <other transcript name> [and <other transcript name>] are part of the same gene; an assembly gap between them contains one or more exons.

gene fragments RP24-11A2.9-001 and RP23-123H10.3-001 and RP23-99D8.1-001 are part of the same gene; an assembly gap between them contains one or more exons.

Transcript Annotation Remark:
TEC

Locus Annotation Remark:
LEC

Transcript Visible Remark:
suspected genomic sequence error affecting CDS in exon <exon number>

Transcript Annotation Remark:
readthrough

Locus Annotation Remark:
overlapping_locus

Locus Annotation Remark:
orphan

Transcript Visible Remark:
non-ATG start codon

Transcript Visible Remark:
non-canonical splice sites between exons <exon #> and <exon #>
non-canonical splice donor site between exons <exon #> and <exon #>
non-canonical acceptor splice sites between exons <exon #> and <exon #>