havana

human and vertebrate analysis and annotation

annotation guidelines

Written by and with contributions from

Laurens Wilming (lw2@sanger.ac.uk)
Adam Frankish
Jane Loveland
Jonathan Mudge
Charles Steward
Jennifer Harrow
HAVANA team

V.20

10 April 2012

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.

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. These categories are part of the gene biotype, which is determined by the hierarchy of transcript biotypes. See appendix for the list of gene and transcript biotypes.

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

Transposon

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

Artifact

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 that submit automatic annotation, sometimes resulting in erroneous CDSs.

Full name artifact gene

Also used for variant transcript models based on cDNAs we consider artifactual, typically because it has a non-canonical splice junction where it "splices" from the middle of one exon, skips one or more exons and "splices" into the middle of another exon.

TEC

"**T**o be **E**xperimentally **C**onfirmed" is used for single-exon mRNAs without polyA features and/or one or two ESTs with polyA (see Figure 11). 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: "**confirm experimentally**" is an attribute 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 Attribute: confirm experimentally

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.

Canonical splice sites

Check that splicing follows consensus splice sites. The LogoGraph below (***Figure 1***) shows the frequency of occurrence of different bases at key positions. ***Figure 2A*** Shows different splice sites with an indication of relative frequency, including uncommon sites not visible in the ***Figure 1***, like G|GC and |AT-AC|.

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

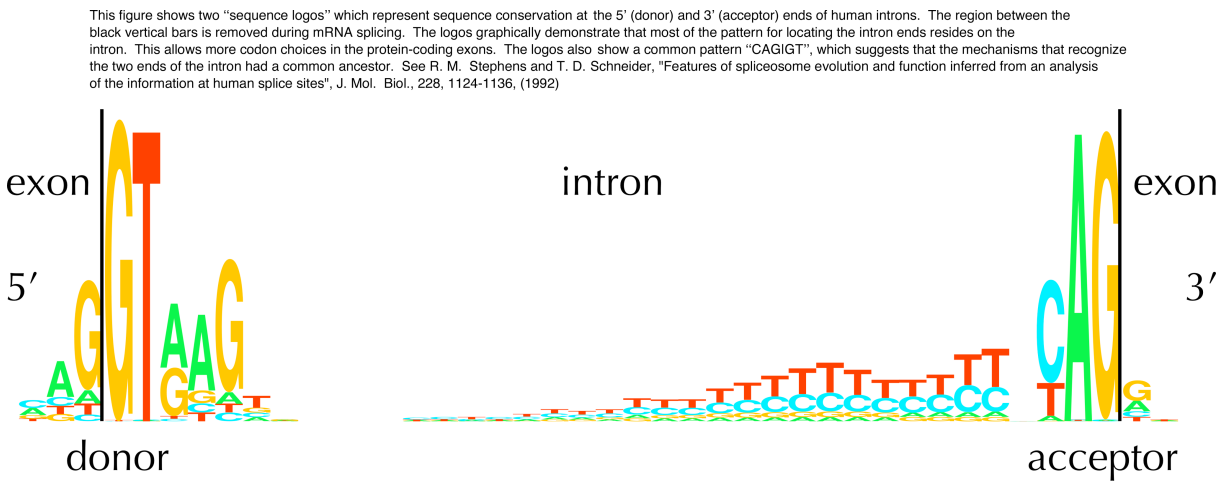


Figure 1: splicing LogoGraph

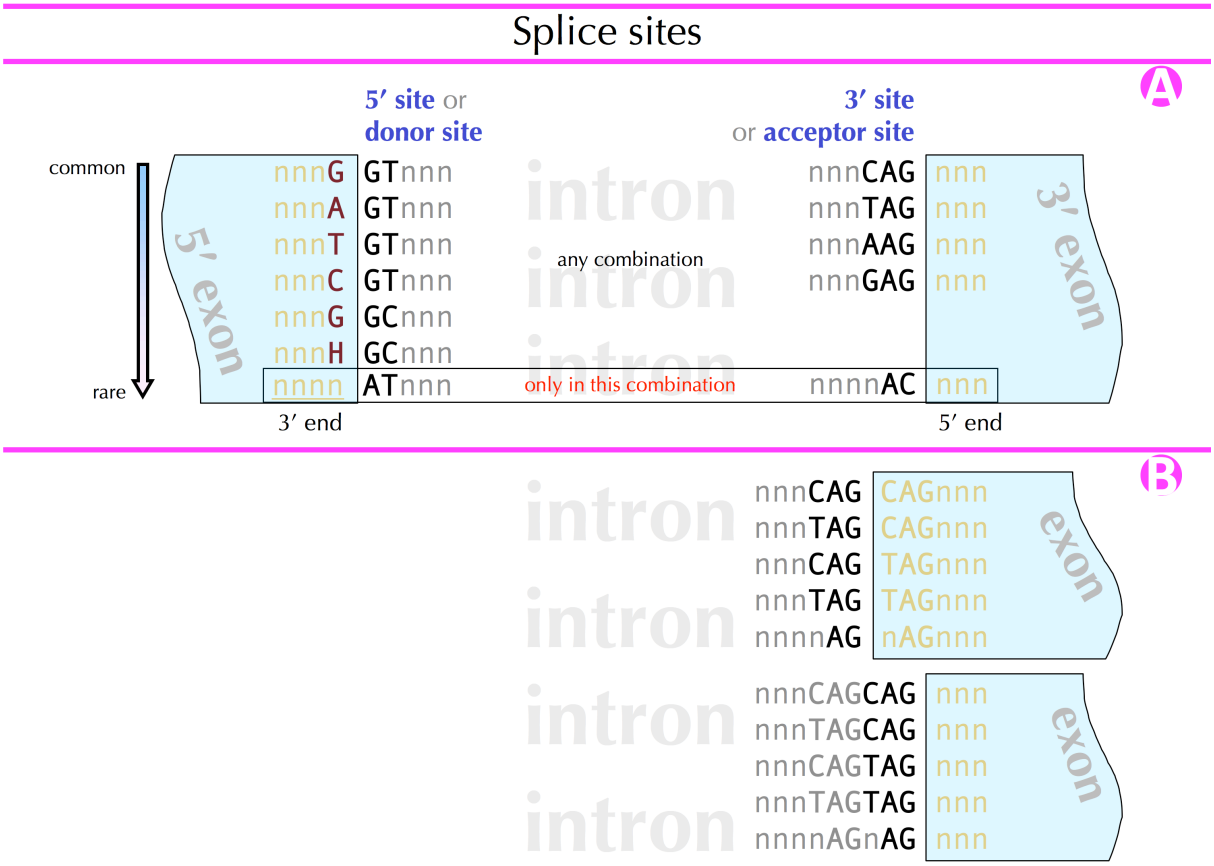


Figure 2: A - splice donor and acceptor sites ordered by frequency; B - NAGNAG splice acceptor sites
A - (H is A, T or C); B - explicitly shows most common ones in addition to nAGnAG

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 ***Figure 2B***), add the following *Transcript Attribute* (found under the "*Splice*" submenu) in all the affected variants:

Transcript Attribute: NAGNAG splice site

Non-canonical splice sites

If a splice site doesn't fit any of the above canonical sites it can be described as non-canonical splice sites. Donor or acceptor sites, or both, may be affected. A non-canonical site may be used in a transcript model if it is supported or explained by any of the following:

- conservation in other species
- genomic sequencing error
- SNP
- U12 intron (*i.e.* AT-AC splice sites)
- mRNA editing
- published support (add PMID as visible remark)

When non-canonical splice sites are used, add the *Transcript Attribute* (under the "*Splice*" submenu) that best describes the reason for the non-canonical nature:

Transcript Attribute:
non canonical U12
non canonical conserved
non canonical genome sequence error
non canonical other
non canonical polymorphism

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

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

mRNA, EST or proteins homology evidence from orthologous loci from other species (*i.e.* non-organism-supported (non-o-s)) can be used to build variants on the condition that homology is perfectly co-linear and all normal splicing rules are upheld, *i.e.* splice sites are canonical or if not the rules for non-canonical splicing are obeyed. You can use non-organism evidence to build NMD variants or to extend variants based on locus-specific evidence, providing the exon structure they share is identical.

As a rule-of-thumb all non-organism mRNAs from orthologous loci should be thoroughly checked (using dotter if required) while non-organism ESTs should only be checked with dotter if there is a good chance they contribute a novel splice feature.

If there is only partial or no b-i-g locus-specific evidence, transcript models can be built using evidence from either paralogous loci (non-b-i-g evidence) or other species (non-o-s evidence).

IMPORTANT
NOTE: do not build non-o-s based retained intron variants unless they have another variation as well, in which case build a partial variant that excludes the retained intron.
NOTE: non-b-i-g and non-o-s evidence should not be used to support polyA features.
NOTE: for loci that appear in clusters of very similar family members only use locus-specific supporting evidence.
NOTE: where locus-specific evidence is not present and the non-b-i-g and/or non-o-s evidence indicates a number of different potential splice variants, choose only one representative variant. Where possible this would be the best match and/or the longest and/or with most exons and/or greatest coverage and/or longest CDS.

Transcript Attribute: not best-in-genome evidence
Transcript Attribute: not organism-supported

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 the following *Transcript Attribute* in all variants except one "reference" variant:

Transcript Attribute: alternative 5′ UTR

Alternative 5′ UTR variants inherit their biotype from the reference, even if the CDS is incomplete. This applies to alternative splicing in the 3′ UTR as well (assuming it does not induce NMD of course). Use the following *Transcript Attribute*:

Transcript Attribute: alternative 3′ UTR

3′ UTRs are extended to the furthest downstream genomically encoded nucleotide (*i.e.* before the start of the polyA tail) (***Figure 4***). Annotate polyA signals (see ***Table 1***) up to 50 bp upstream of the polyA site. In the presence of a polyA signal, 2bp of unaligned As (forward strand) or Ts (reverse strand) in matching evidence is sufficient for a polyA site. Gaps in the tiling path between spliced evidence and the cluster of polyA containing 3′ ESTs typically seen at the 3′ end are allowed if smaller than 200 bp (***Figure 4***). Multiple discrete polyA features (*i.e.* polyA site with corresponding pA signal) are annotated, but the gene is stretched to the downstream-most set, unless the tiling path has gaps larger than 200 bp or a specific splice variant is associated with a specific polyA feature set. See panels D-G in ***Figure 4***. Where multiple polyA signals are associated with the same site, annotate the most common signals (AATAAA or ATTAAA) only. Often there is polyA site "wobble" where the exact position varies by a few nucleotides, in which case annotate at minimum the downstream-most site.

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.
NOTE: for technical reasons polyA sites are annotated as a pair of coordinates, namely the penultimate and the last genomically encoded base (***Figure 3***).

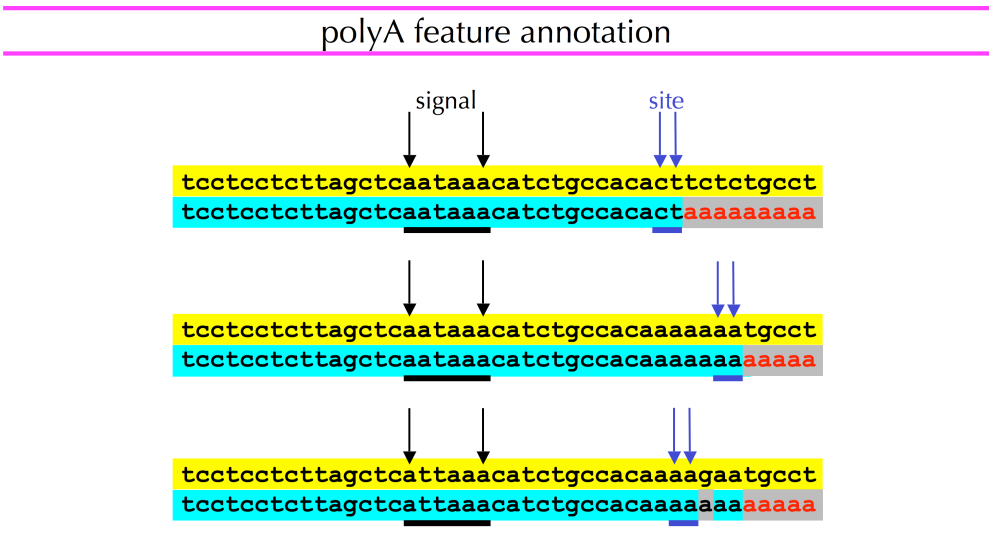
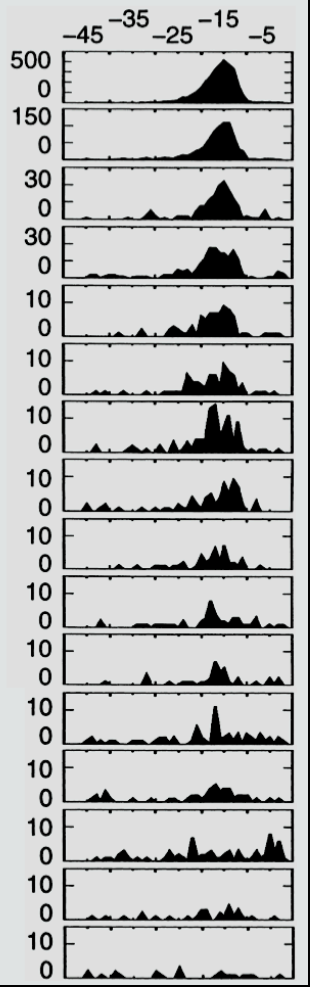


Figure 3: coordinate pairs used for annotation of polyA signals and sites

Table 1: 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]
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	
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	

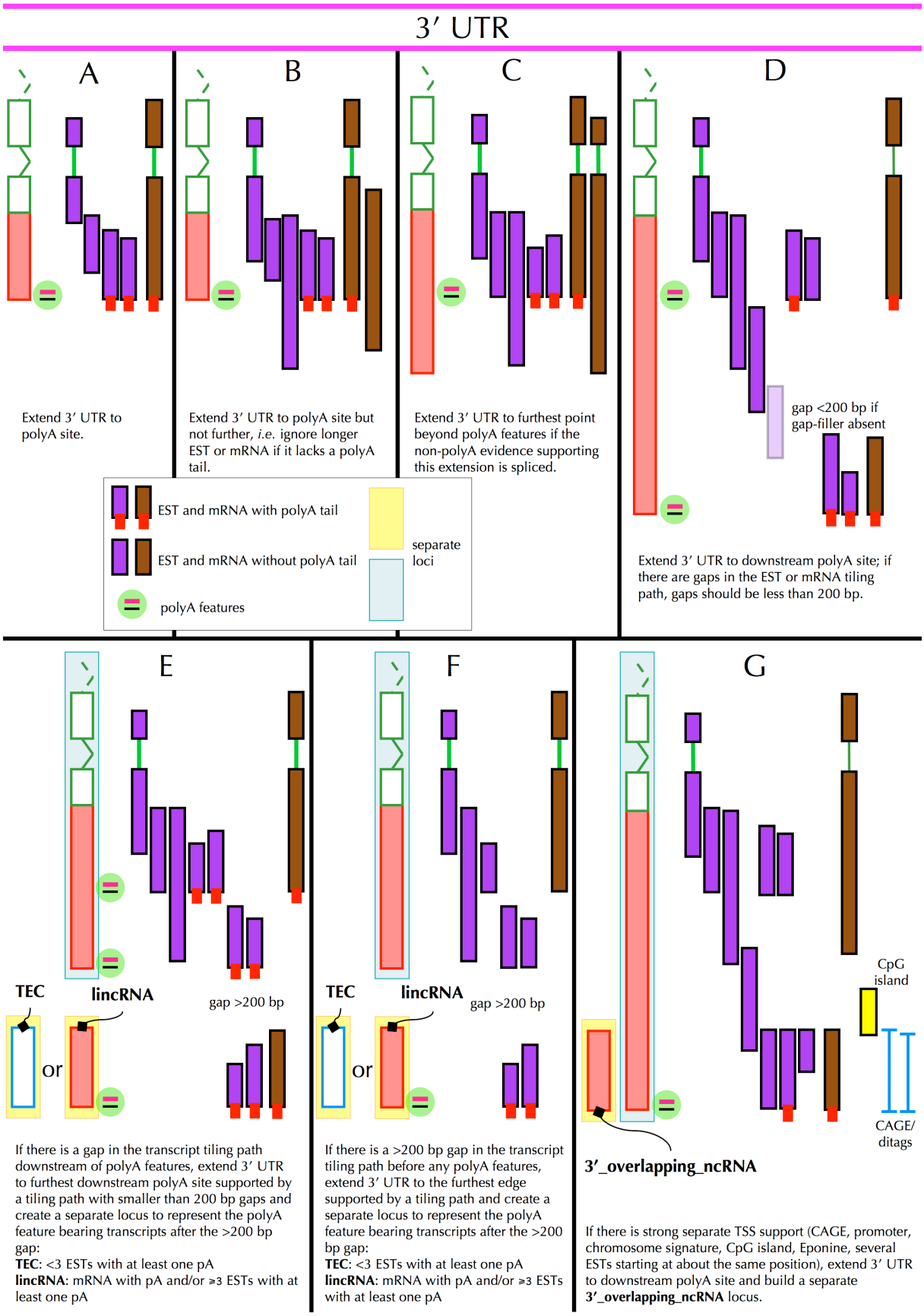


Figure 4: annotation of 3′ UTRs in the context of different types of evidence

Genomic sequence errors

If a genome sequence error is suspected, check whether it is a known validated SNP/DIP (see also **Polymorphic Pseudogene** on page 21 and **WARNING**s below), using Ensembl or UCSC browsers. If it isn't, mail the designated Havana team member that deals with these issues with the following:

- genomic clone accession number
- the cDNA coordinate(s) of the error on, and accession number of, a disagreeing cDNA
- the details of the error
- the SNP id if it is an un-validated SNP
- the gene symbol of the affected gene
- 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)
- 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 a UTR, or a 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 attribute: sequence error

WARNING: do not use Genoscope mRNAs to make any decisions regarding sequencing errors or polymorphisms. Genoscope mRNAs sequences are modified to match the genomic sequence.

WARNING: genome sequence errors often have SNP IDs, with a polymorphism being wrongly called on differences between the reference and other completely sequenced genomes. A SNP call should not necessarily be trusted unless there is good allele frequency data to support it (Figure 5).

Reliable SNPs in dbSNP

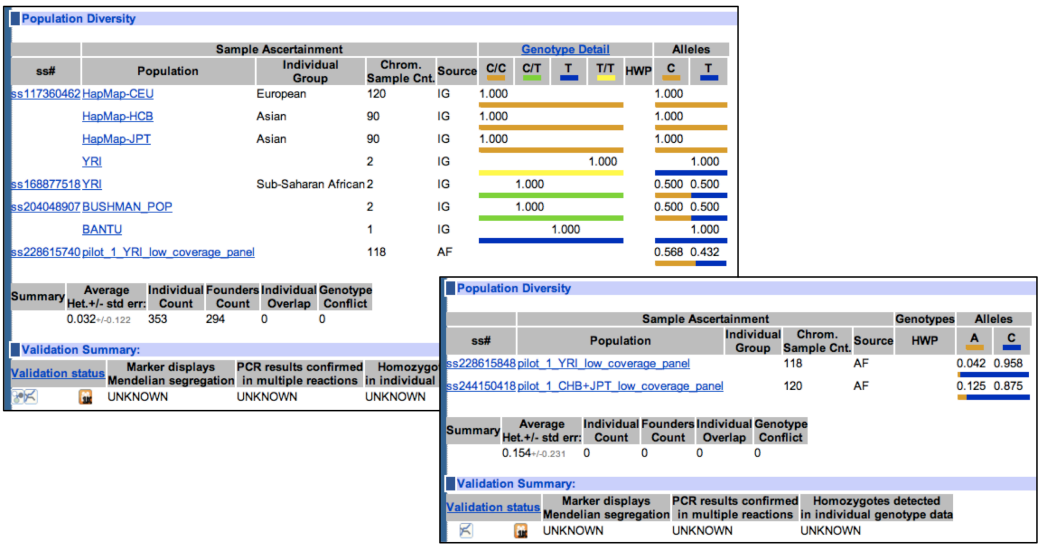


Figure 5: examples of frequency data of real SNPs in dbSNP

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. The default position is the annotation of the most upstream ATG.

Figures below show the practical application of these guidelines (***Figure 6***, Figure 13, ***Figure 14***). In ***Figure 6***, 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 *Transcript Attribute* 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 choice of **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 attribute). Variant **b** has no conservation or functional evidence, 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 7***) 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.

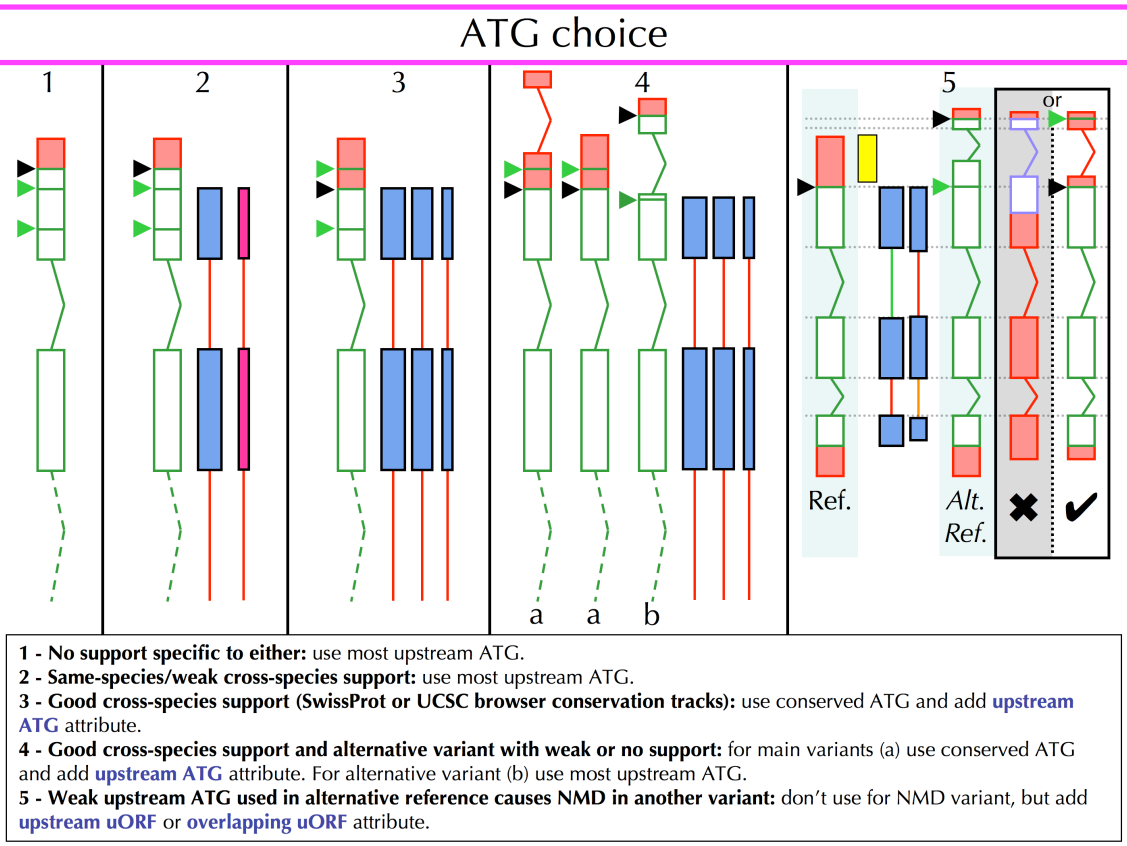


Figure 6: translation start site annotation in the case of alternative ATGs

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

Transcript Attribute: upstream ATG

Conversely, when an upstream ATG is used where a downstream ATG seems more evolutionary conserved, tag as follows:

Transcript Attribute: downstream ATG

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 8***).

NOTE: where 1) an alternative low confidence upstream ATG is used for a coding variant (that still contains the conserved high confidence canonical downstream ATG) and 2) a further variant contains both ATGs but the use of the upstream ATG would result in NMD, refrain from using the upstream ATG for the NMD variant and use the canonical ATG instead (example 5 in ***Figure 6***). Do add the appropriate uORF attribute:

Transcript Attribute: upstream uORF
Transcript Attribute: overlapping uORF

and add the downstream ATG attribute to the alternative reference variant:

Transcript Attribute: downstream ATG

If the upstream ATG is well supported (strong Kozak, conservation, TSS features, etc.), or if both ATGs are equally weak/(un)supported, do use this ATG for NMD variants.

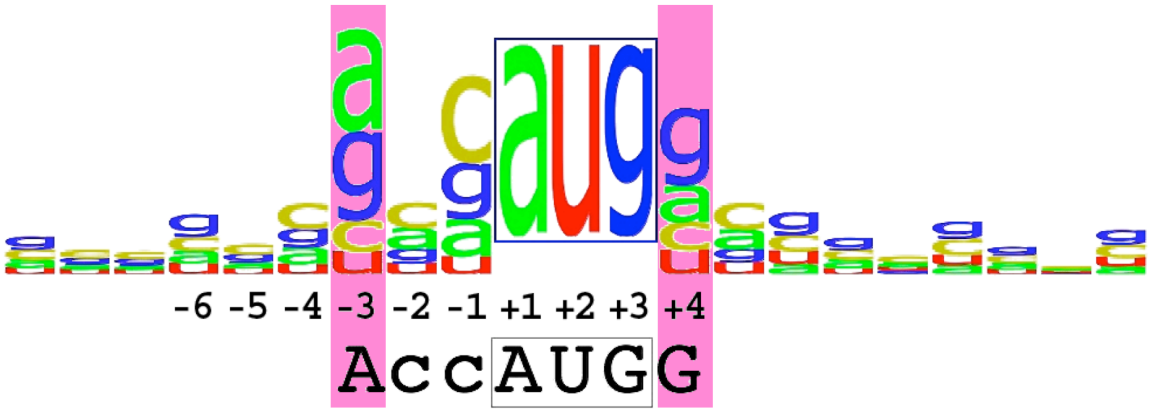


Figure 7: Kozak sequence LogoGraph with important bases highlighted

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***

- Translation from a reference ATG ≥35aa (including the stop) with the stop codon >50bp from a downstream splice site?
  - Make transcript type NMD (***Figure 19***).
- Upstream translation <35aa?
  - Translation may be re-initiated from a downstream in-frame internal or unique ATG. Use such an ATG only if the resulting translation shares at least some peptide sequence with a reference translation.
- Stop codons 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 19***).
- CDSs can have non-ATG starts, which should be annotated just like ATG, provided the validity is supported by publication or conservation. Add following transcript attribute:

Transcript Attribute: non-ATG start codon

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_seq 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: there are one or more splice junctions >50bp downstream of the end of the CDS (using the start of an appropriate reference CDS) (see *Figure 19*).

Non_stop_decay: there is no stop codon before the polyA site is reached (see Figure 22).

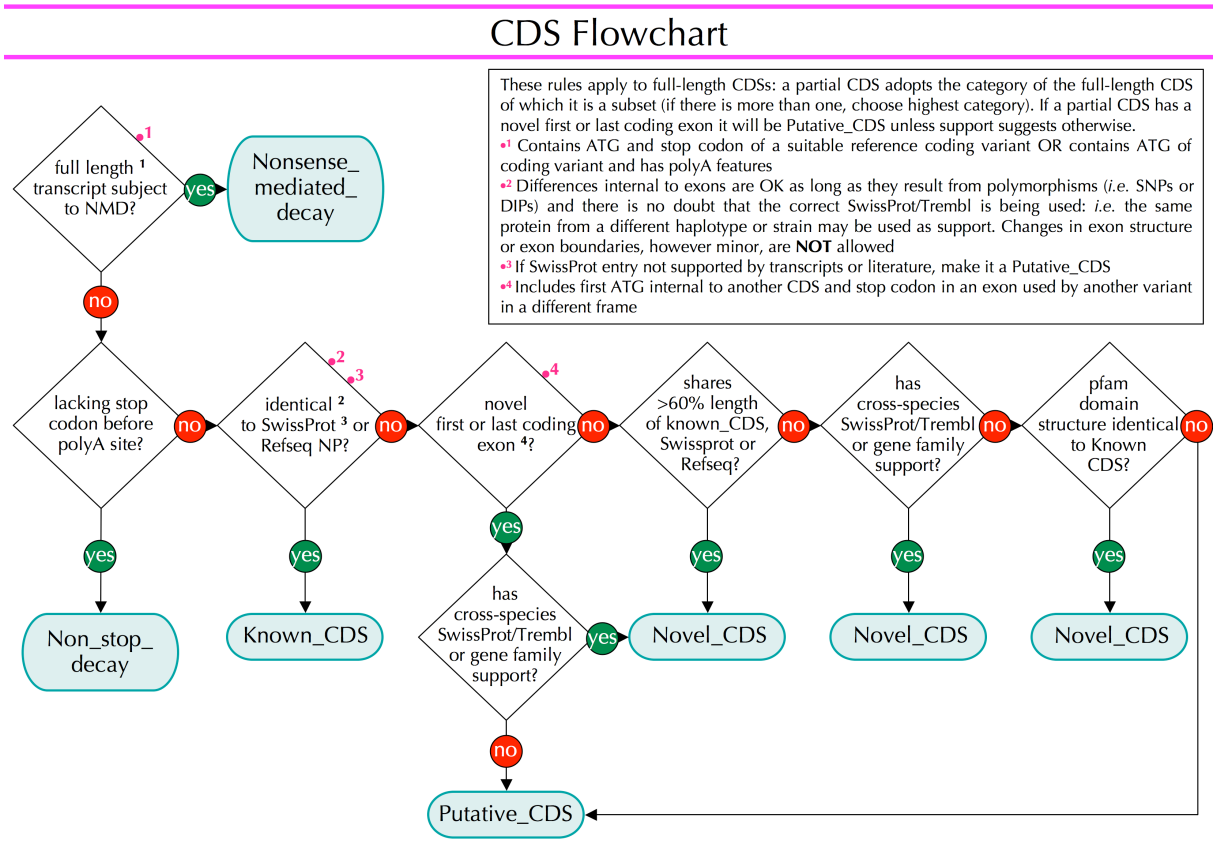


Figure 8: CDS decision graph

Two- or three-exon solitary gene objects will generally not have a CDS annotated unless it is a known gene or there is supporting evidence in the form of homology, domains or 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 like Pfam domains.

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 9***. An annotated orphan protein may be tagged as Known_, Novel_ or Putative_CDS depending on the supporting evidence (SwissProt, RefSeq).

Add the following:

Locus Attribute: orphan

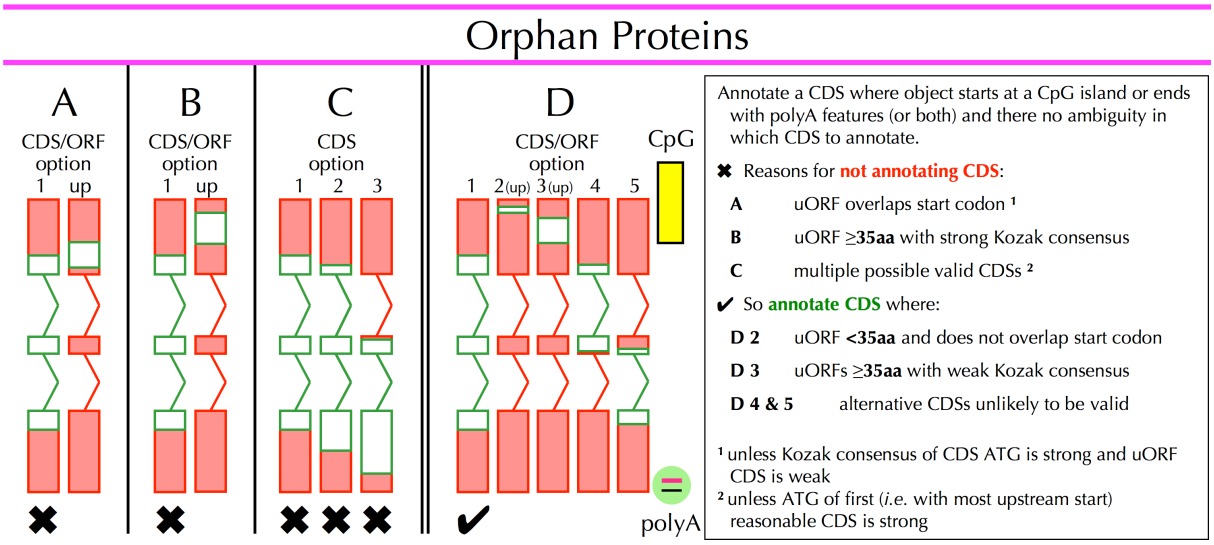


Figure 9: when to annotate orphan proteins

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 Attribute: selenocysteine
Locus Visible Remark: selenoprotein

---

ncRNA hosts

Make sure the ensembl_ncRNA and das_WashU_PASA_human_mRNAs tracks are switched on to see miRNAs, snoRNAs, piRNAs etc. The PASA track doesn't distinguish them specifically but you can recognise them because they will show up as a multitude of roughly same sized small single-exon models in introns or UTRs of the host model.
Any locus, whether coding or not, that is a host for small non-coding RNAs will need the following Locus Attribute and Visible Remark:

Locus Attribute: **ncRNA host**

Locus Visible Remark: **<name or type of hosted small ncRNA>**

MIR26A2 host
snoRNA host

Non-coding loci

Loci where none of their variants have a CDS are annotated with one of the following ncRNA biotypes:

lincRNA[0]: long intergenic non-coding RNA locus. Requires lack of coding potential and is often not conserved between species. If not supported by spliced cDNAs or three or more ESTs, an anchored 5′ end (CpG island, chromatin signature, ditags) or 3′ end (polyA features) is required. Single-exon cDNAs can be lincRNA and have a TEC attribute. Use Ensembl predicted lincRNAs only as a guide.

Antisense[#]: transcripts overlapping the genomic extent of one or more coding loci on the opposite strand. Also for published instances of antisense transcripts regulating a coding gene. As this is a locus level biotype, variants that are not physically antisense, are still labelled antisense by virtue of being a variant of a transcript that is antisense.

Sense_intronic: transcripts that are in introns of coding genes and do not overlap any exons. Add **Locus Attribute:** **overlapping locus** to all relevant loci.

Sense_overlapping: transcripts that contain a coding gene in their intron on the same strand. Add **Locus Attribute:** **overlapping locus** to all relevant loci.

3′_overlapping_ncrna[#]: transcripts where ditag, TSS and/or published experimental data strongly supports the existence of short non-coding transcripts independently transcribed from the 3′ UTR. Add **Locus Attribute:** **overlapping locus** to all relevant loci.

NOTE: these biotypes are only for stand-alone loci, not for variants of coding loci.
NOTE: conversely, do not use Transcript biotypes for non-coding loci.
NOTE [0]: lincRNA is trumped by all other non-coding subtypes (Antisense, 3′_overlapping_ncRNA, Sense_overlapping, Sense_intronic).
NOTE [#]: a transcript can at the same time be Antisense and one or more sense subtypes, *e.g.* Sense_overlapping, in which case Antisense takes precedence over any other non-coding subtype.

See ***Figure 10*** for some examples of ncRNA use and the required attributes. Any spliced locus-specific evidence is sufficient, including a single spliced EST. For single exon evidence see next section (Single-exon mRNAs).
Make gene descriptions as informative as possible, for example:

Full name	novel transcript, antisense to ARHGAP1 and DST novel transcript, sense overlapping XBOX1 novel transcript, sense intronic to RPS3

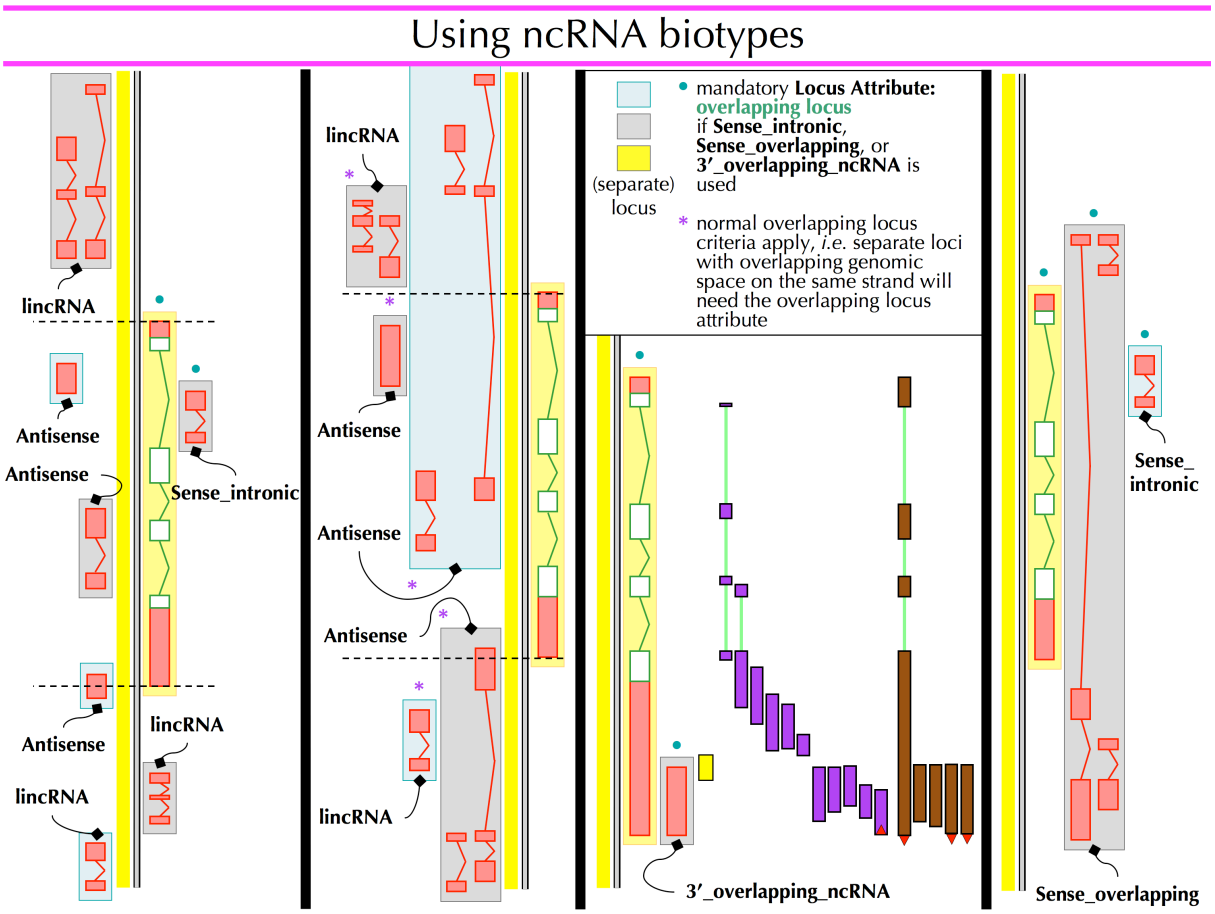


Figure 10: examples of ncRNA biotype use

Single-exon mRNAs

We strive to represent all locus-specific mRNAs, including single-exon. The following applies to intergenic, antisense and intronic transcripts, so where it says Non-coding, choose the subtype that is appropriate for the circumstances. See ***Figure 11*** for examples. For non-splicing transcripts overlapping 3′ UTR, see ***Figure 4***.

- If it involves a single-exon mRNA locus, annotate as TEC or Artifact biotype (see **Gene Classification** for criteria), or upgrade to a Non_coding subtype if the evidence is strong enough (see previous section).
- If it is a host for small RNAs, annotate as appropriate Non_coding subtype and add ncRNA host attribute See also **ncRNA hosts** for details.
- If it involves a single-exon variant of a coding locus, annotate as Retained_intron variant.

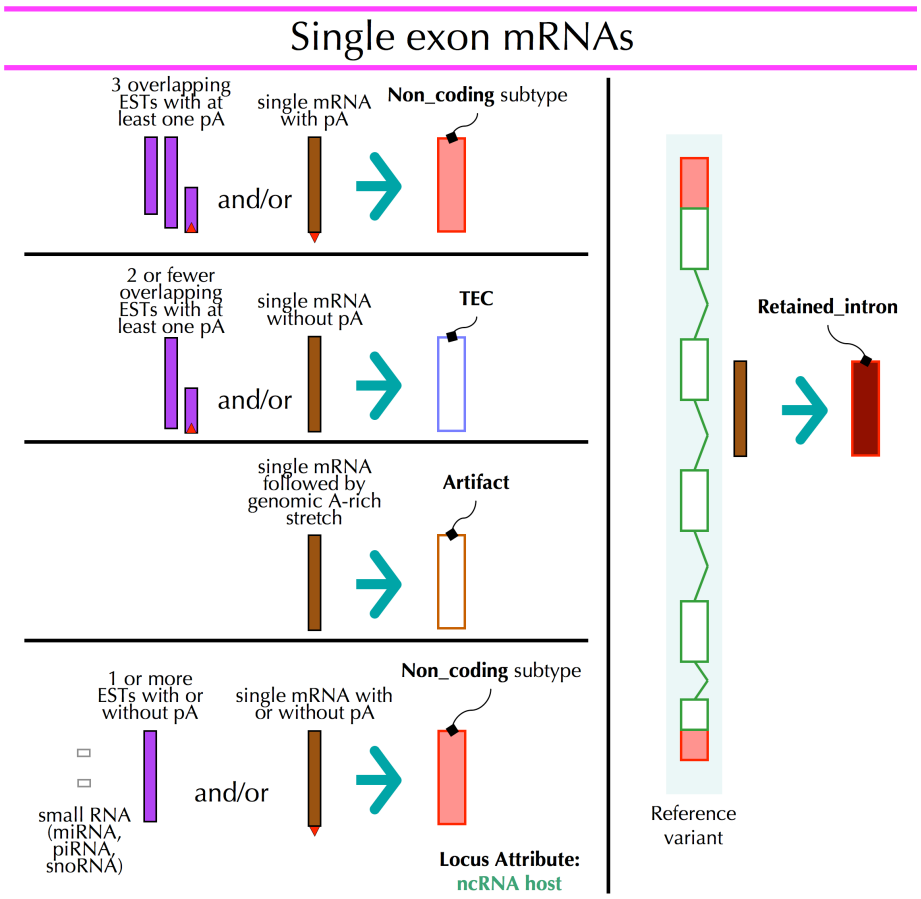


Figure 11: annotating single-exon mRNAs

Pseudogenes

We divide pseudogenes into five categories. Most pseudogenes are straightforward, 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 are 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 part of the annotated pseudogene, *i.e.* the pseudogene should be annotated as two "exons", but of course still labelled "processed" (see Figure 12). 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 where either of the two most common signals (AATAAA or ATTAAA) are visible in the genomic sequence or in aligned transcript evidence. Sometimes processed pseudogenes have an intact CDS similar or even identical to their unprocessed parent. If it does not have

locus-specific transcription evidence it will be annotated as a pseudogene; however, if there is locus-specific transcription evidence and the translation start, end and length are the same as the parent, this can be annotated as a "Putative_CDS", or "Known_CDS" if it is a known named retrogene.

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 receptor) and the prospective pseudogene has a slightly truncated 5′ or 3′ 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. By definition, pseudogenes that occur in a cluster with other family members (coding or pseudo) are unprocessed pseudogenes. Where possible annotate the proper exon boundaries. The easiest way to do this is to build the model based on mRNA homology (because it is easier to see the splice sites and the alignment shown is splice site aware) and then trim the ends to the extent of the protein coverage.

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 used 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 (we have used mouse as a reference for all unitary pseudogenes annotated to date). 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. Unitary pseudogenes are generally unprocessed pseudogenes and they can actually have more than one "orthologous parent". For example certain gene families (*e.g.* Mup, Vnr) 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 fixed (species-wide) pseudogenization event and not a polymorphism. Check that it is not a known validated SNP.

IG_pseudogene

Special category for pseudogene versions of Immunoglobulin gene building blocks.

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. In cases where a single transcript overlaps (on exon level) more than one pseudogene, annotate the transcript as a separate locus of the appropriate ncRNA biotype. Whichever scenario applies, all pseudogene loci that overlap transcripts use the transcribed pseudogene Locus Attribute:

Locus Attribute: transcribed pseudogene

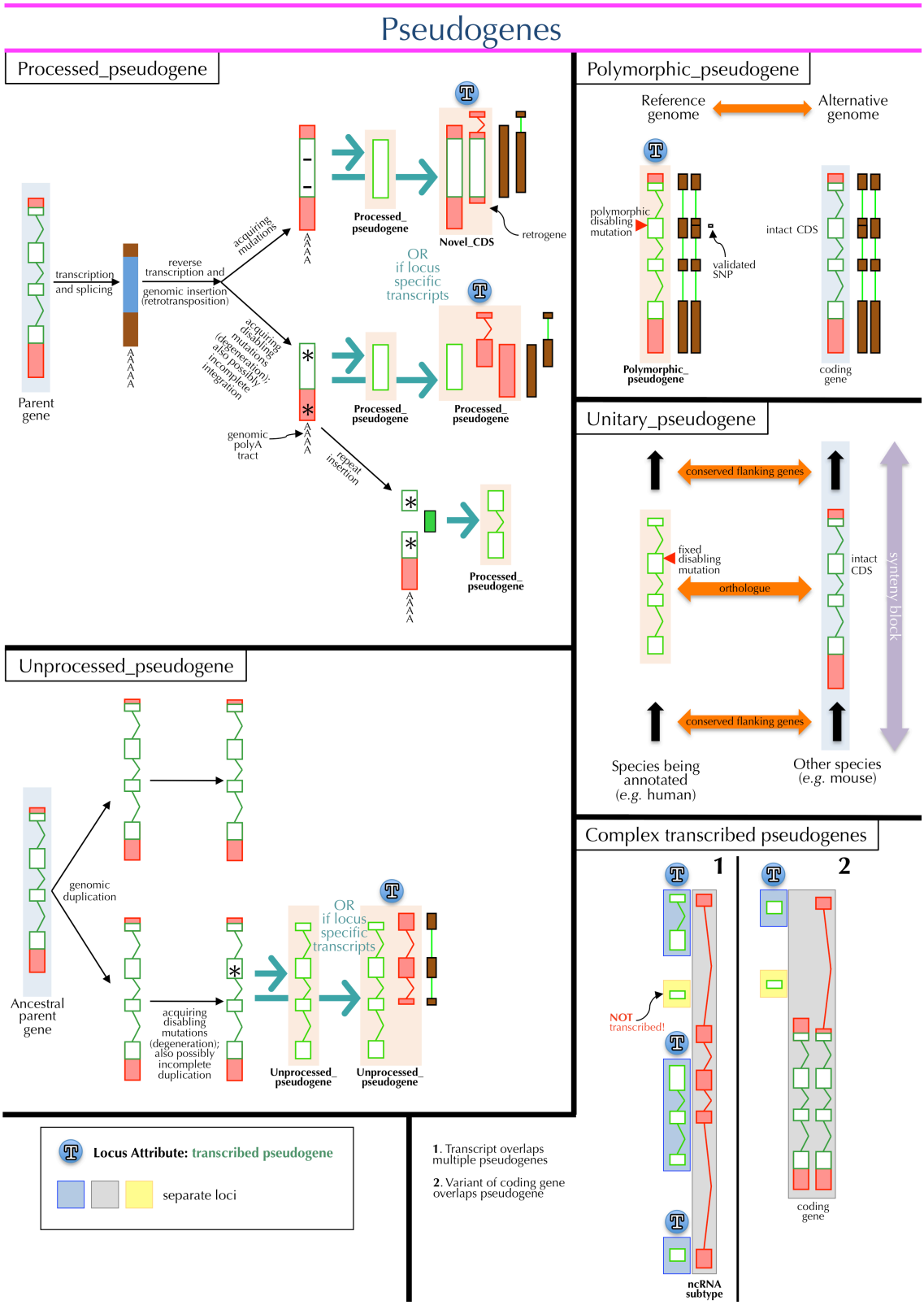


Figure 12: pseudogene annotation

Supporting evidence

In case of splice variation, the main variant receives 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. 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_seq 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, NP_123456

If supporting evidence has not been submitted, *i.e.* the model is based on literature or collaborator evidence, add the following:

Transcript Attribute: non-submitted evidence

Using RNA-seq data

Where there is paucity or absence of longer transcript evidence (ESTs and mRNAs), RNA-seq confirmed introns or ensembl RNA-seq based models can be used to:

- connect fragments of a fragmented locus together
- extend incomplete loci
- provide validation for non_organism_supported splice variants (where RNA-seq data validates all non-organism supported splice junctions, the "not organism supported" tag should be removed)

NOTE: Where loci are extended/linked they must produce sensible results: coding loci should be coding and not break pfam domains, non-coding loci should not break any structural motifs (if such information is available from rfam).
NOTE: Where an Ensembl RNA-seq based model extends a partial model supported by non_organism evidence the variant can be extended to full length based on the Ensembl model. The provisos on sensible model described above apply. RNA-seq data should be used conservatively where there is native transcript data.

Any objects modified on the basis of support provided by RNA-seq data need the non-submitted evidence attribute and a transcript annotation remark referencing the RNA-seq based gene model or confirmed intron:

Transcript Attribute: non-submitted evidence
Transcript Annotation Remark: RNA-seq supported
Transcript Annotation Remark: <IDs of RNAseq evidence>
tissue=ovary:SOLEXAG0000014410

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 chance that a variant has (an)other alternative event(s) outside the homology.

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. But, if the overlapping exons in the two transcript models have CDSs in different frames they should be annotated as separate loci.

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.

CDS or no CDS?

Many factors determine whether or not we annotate a CDS in a splice variant, mostly related to the structure of the variant compared to other, confidently annotated, coding variants of the locus. An important consideration is whether the variation affects the first or last coding intron. A coding intron is an intron flanked by two coding exons or coding parts of exons. See Figure 13, ***Figure 14*** and ***Figure 15***, and the section **Defining first and last coding introns** for guidance on the annotation of CDSs in variants.

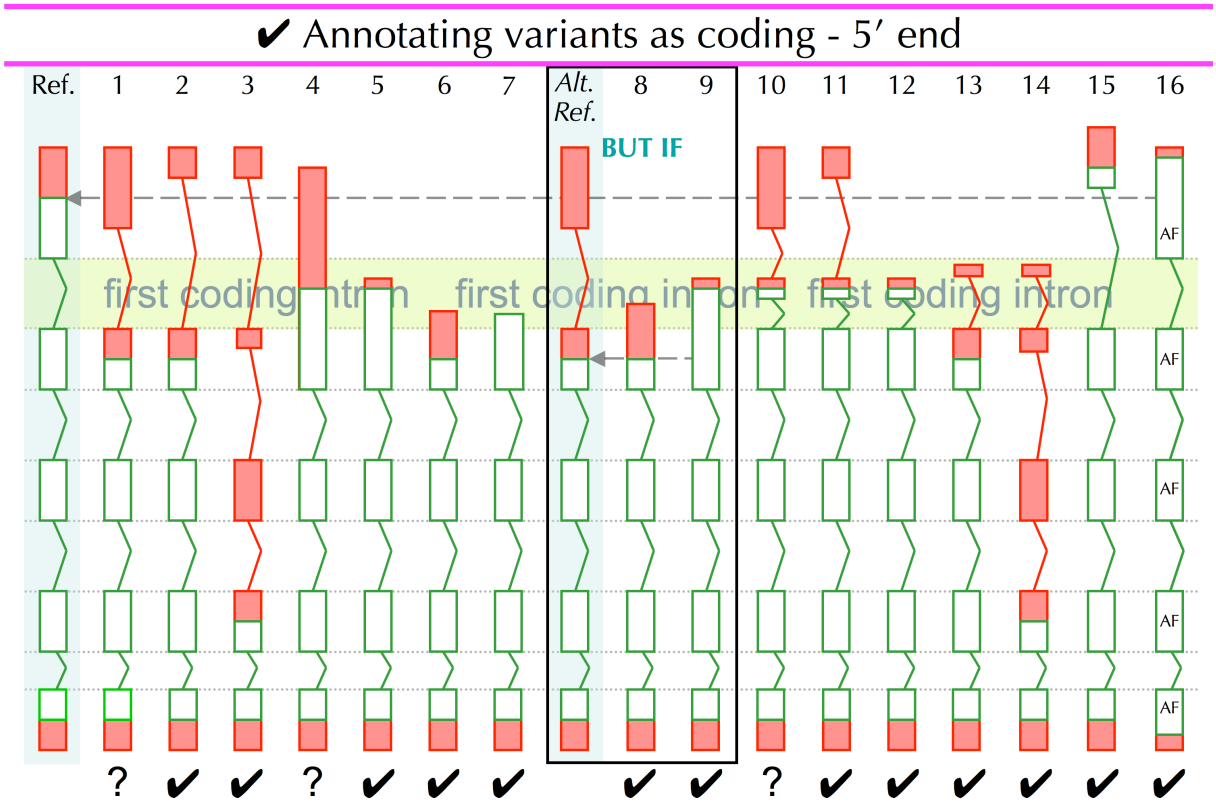


Figure 13: annotating variants as coding - 5' end

1 & **10** - ORF initiating at the same ATG as the 'reference' variant ≥35aa? Annotate as NMD; if <35aa consider coding variant reinitiating form downstream ATG.

4 - ORF initiating at the same ATG as the 'reference' variant ≥35aa? Annotate as Retained_intron; if <35aa consider coding variant reinitiating form downstream ATG.

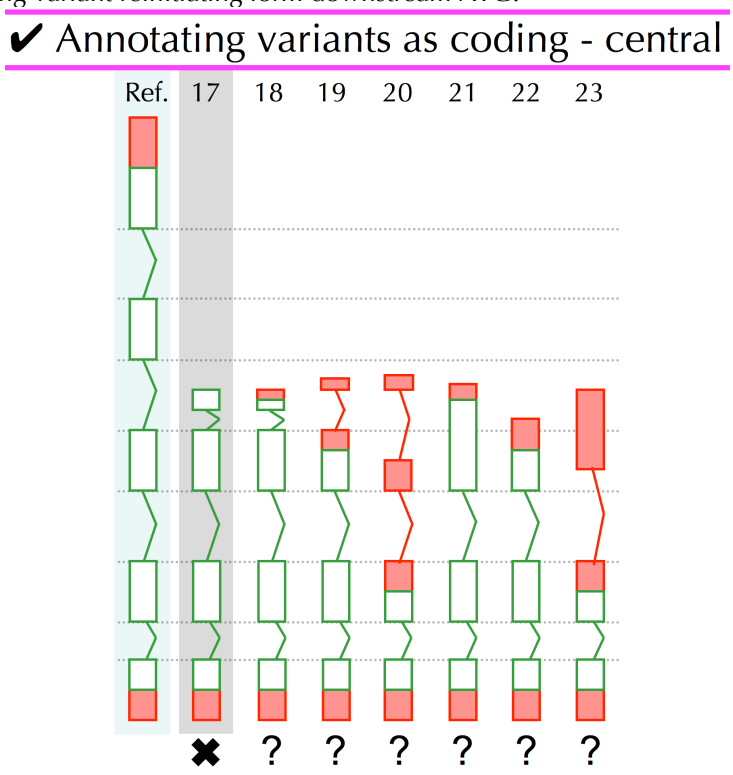


Figure 14: annotating variants as coding - central

18 - 23 - Annotate as coding if the presence of a novel TSS is supported by mRNA/EST cluster (≥3 independent transcripts; can be mix of ESTs and mRNAs) or CpG islands or CAGE-tag cluster or PE-tag cluster or clear TF binding evidence. Otherwise type as appropriate (transcript, retained intron).

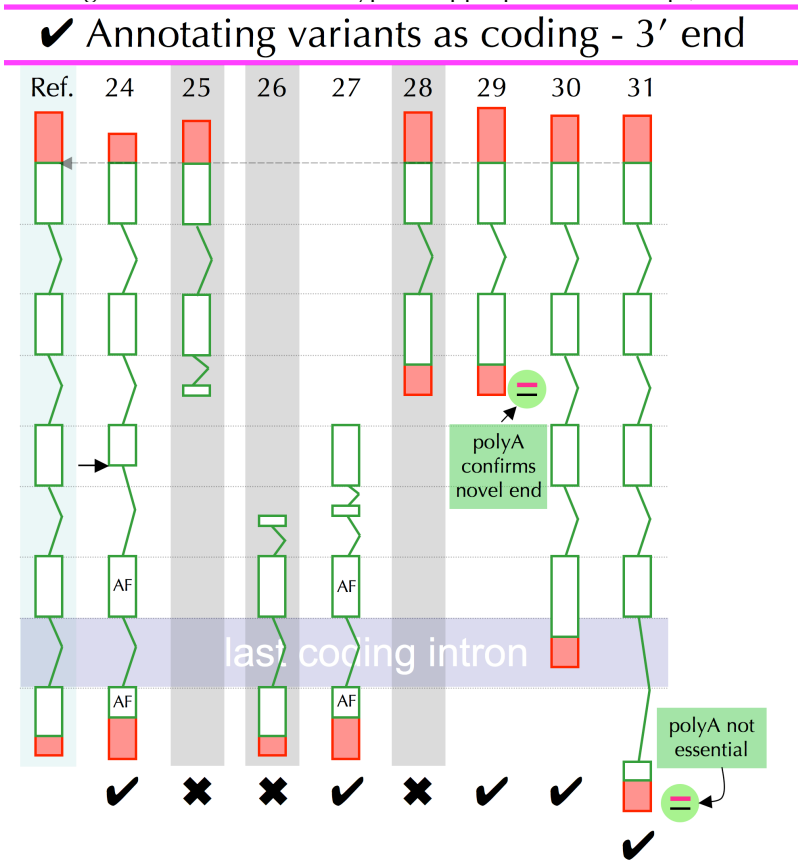


Figure 15: annotating variants as coding - 3' end

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 coding intron may well be utilizing an alternative promoter (or be under weak control of the promotor used by another proximal variant), which are more likely clustered at the 5′ end of genes (see Figure 13). 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 at least three independent ESTs/mRNAs or 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 coding exon is confirmed by polyA features, a novel last coding intron is created where normal final exon annotation rules apply. See ***Figure 16***.

If a variant has a novel first or last internal exon relative to a reference transcript and no polyA features, CpG island, TSS support, conservation, SwissProt, domains or paralog homology to support it as a true start/end, annotate as a transcript (***Figure 14***, ***Figure 15***).

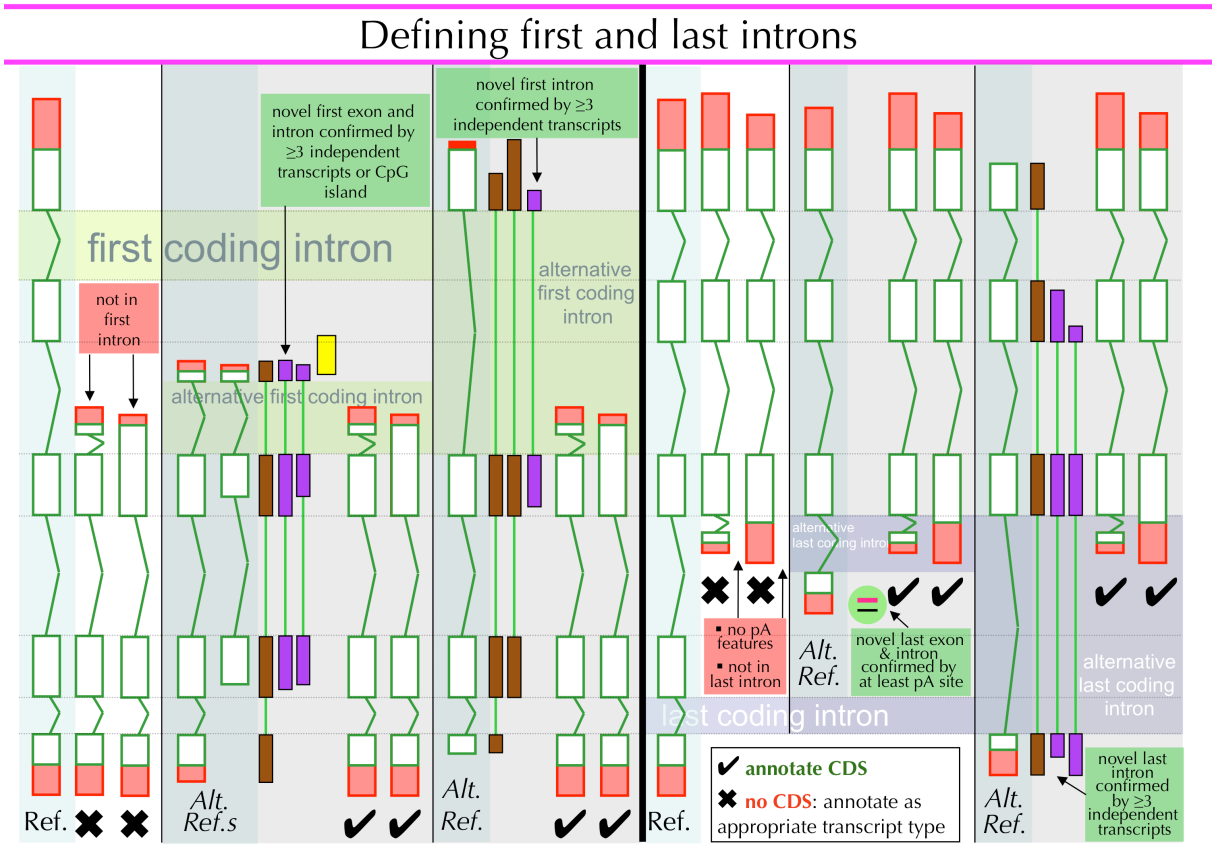


Figure 16: defining first and last introns

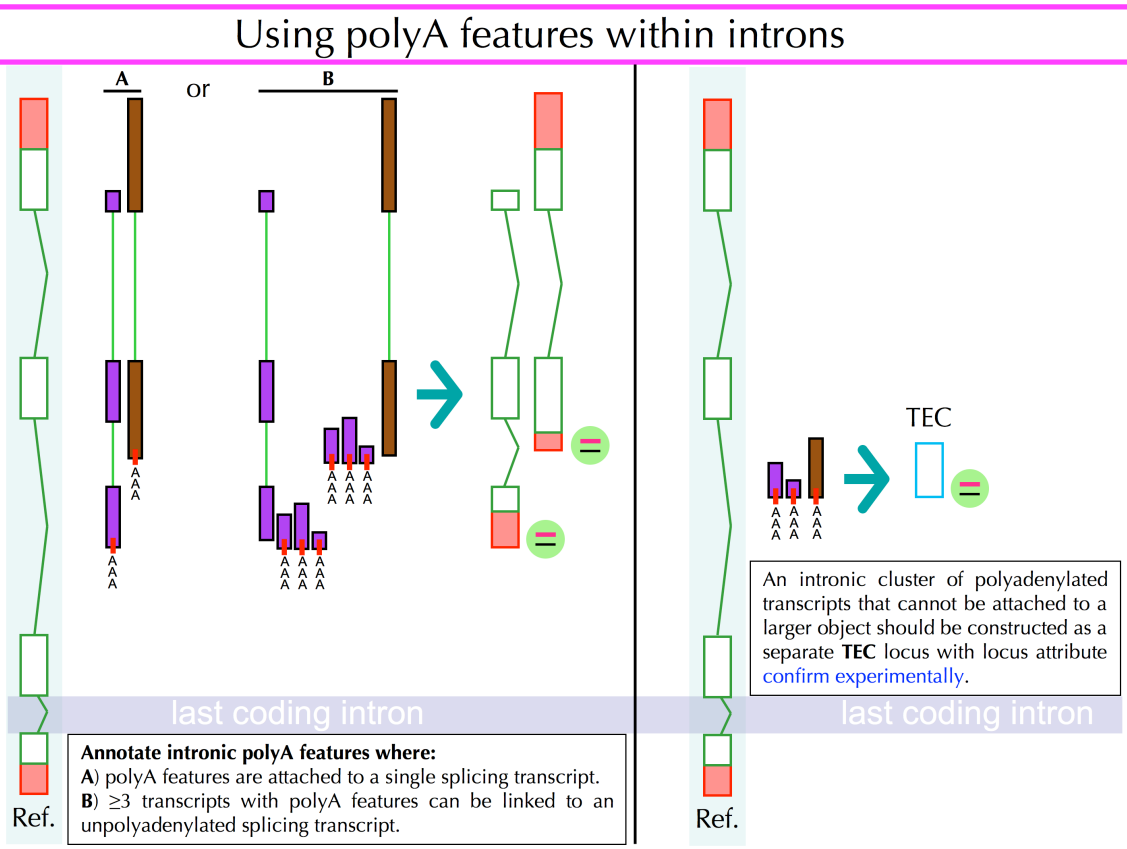


Figure 17: how and when to use polyA features within introns for variants

Retained introns in coding transcripts

Variants that have complete retained introns in the coding region should have the appropriate attribute set:

Transcript Attribute: retained intron first
retained intron CDS
retained intron final

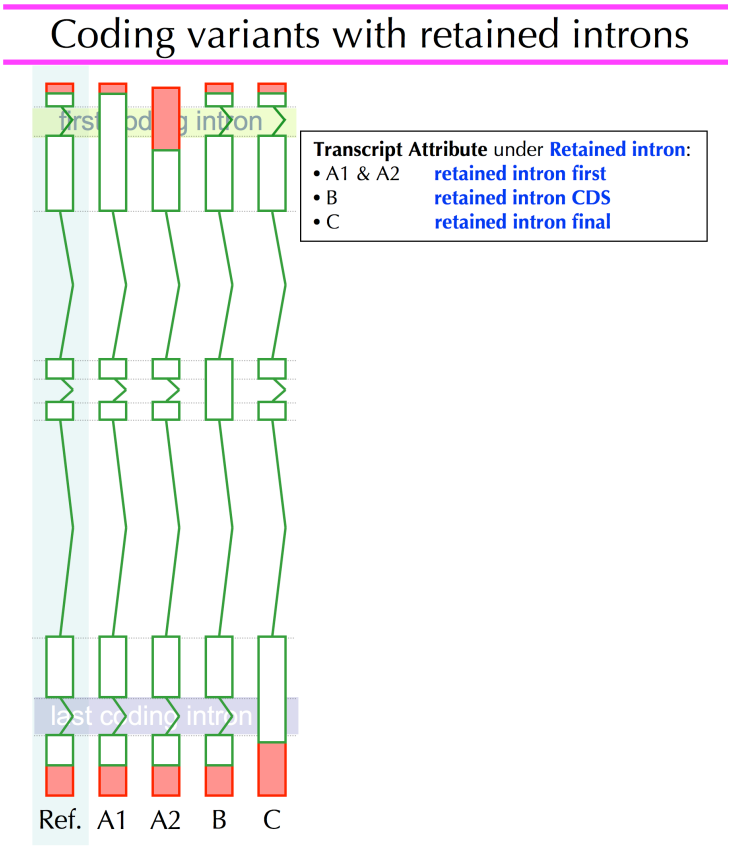


Figure 18: attributes for coding transcripts with retained intron

NMD

The presence of ***any*** splice site further than 50bp from the stop codon will be likely to render a transcript subject to degradation via nonsense-mediated decay (NMD). So 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 (***Figure 19***).
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). If, however, this cannot be determined (double cross in ***Figure 19***), annotate as Transcript and add:

Transcript Attribute: NMD likely if extended

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

Transcript Attribute: NMD exception
Transcript Annotation Remark: [PMID <id>, <publication reference>]

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

NOTE: A transcript with a retained intron after the NMD stop codon, is annotated as NMD (asterisk in ***Figure 19***). Unless NMD is a consequence of a retained intron in in which case it would be Retained_intron (diamond in ***Figure 19***).

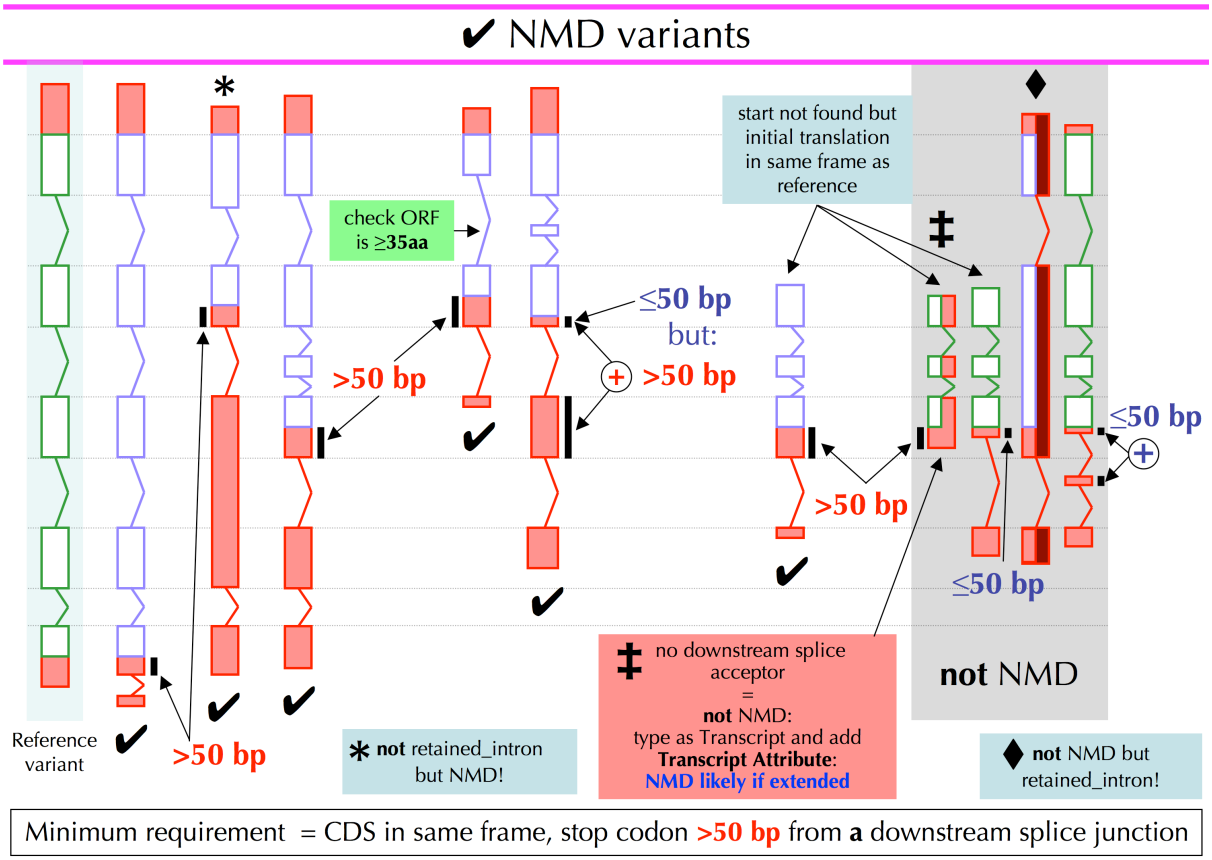


Figure 19: annotating NMD variants

NOTE: If according to homology evidence the only protein-coding variant appears to be subject to NMD because of an apparent intron in the 3′ UTR, check whether this intron actually represents a polymorphic repeat insertion or expansion. See ***Figure 20***.

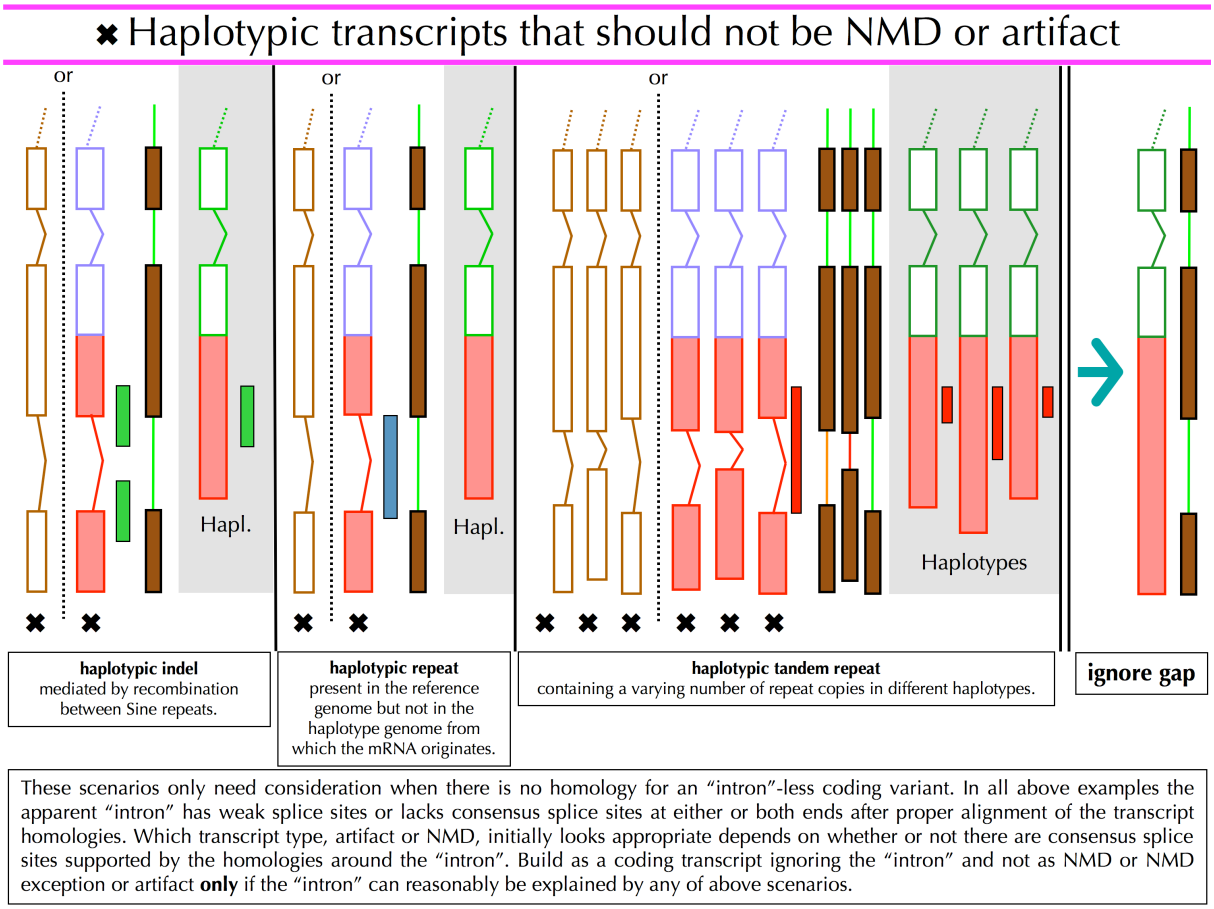
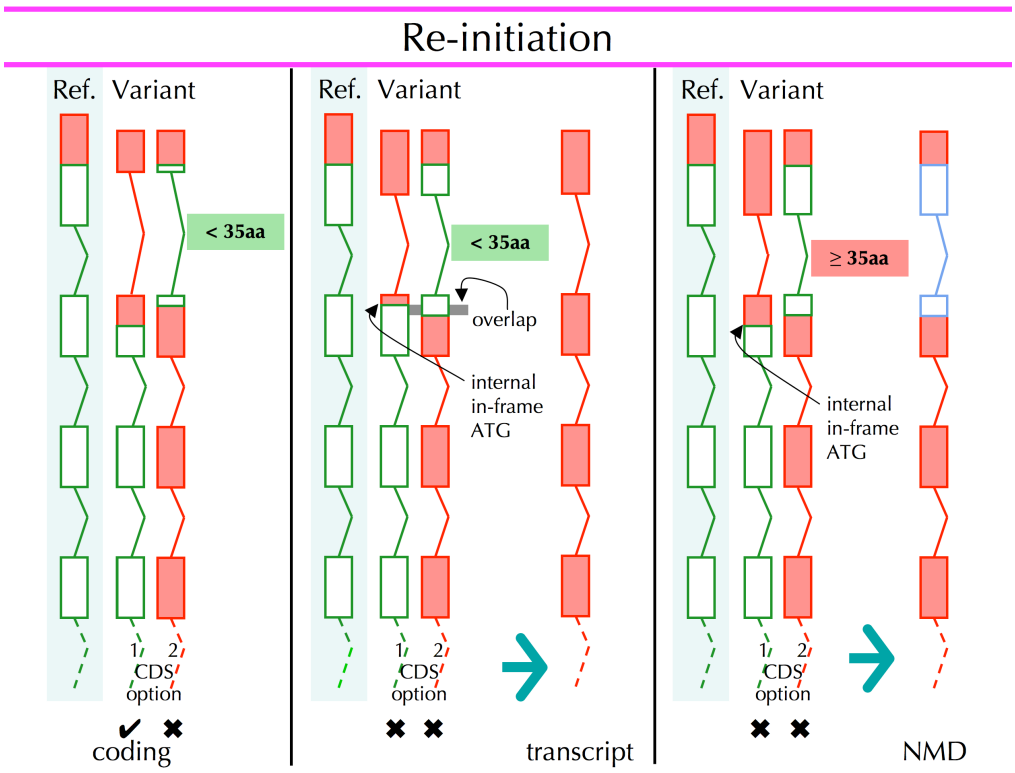


Figure 20: false haplotypic introns

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 21). If the uORF <35aa re-initiation will occur at the next ATG downstream 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 upstream of the stop codon of the uORF or 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 CDS or 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 21: effect of uORFs on re-initiation of translation

NSD

Nonstop decay is a process that affects transcripts that have polyA features (including signal) without a prior stop codon in the CDS, *i.e.* a non-genomic polyA tail attached directly to the CDS without 3′ UTR. These transcripts are subject to degradation. Their translation could give rise to harmful peptides with a poly-K (poly-lysine) stretch at the C-terminal end. Much like NMD transcripts, either aberrant splicing or SNPs can cause these transcripts to be generated. See Figure 22 for examples. The end not found tag should **not** be set.

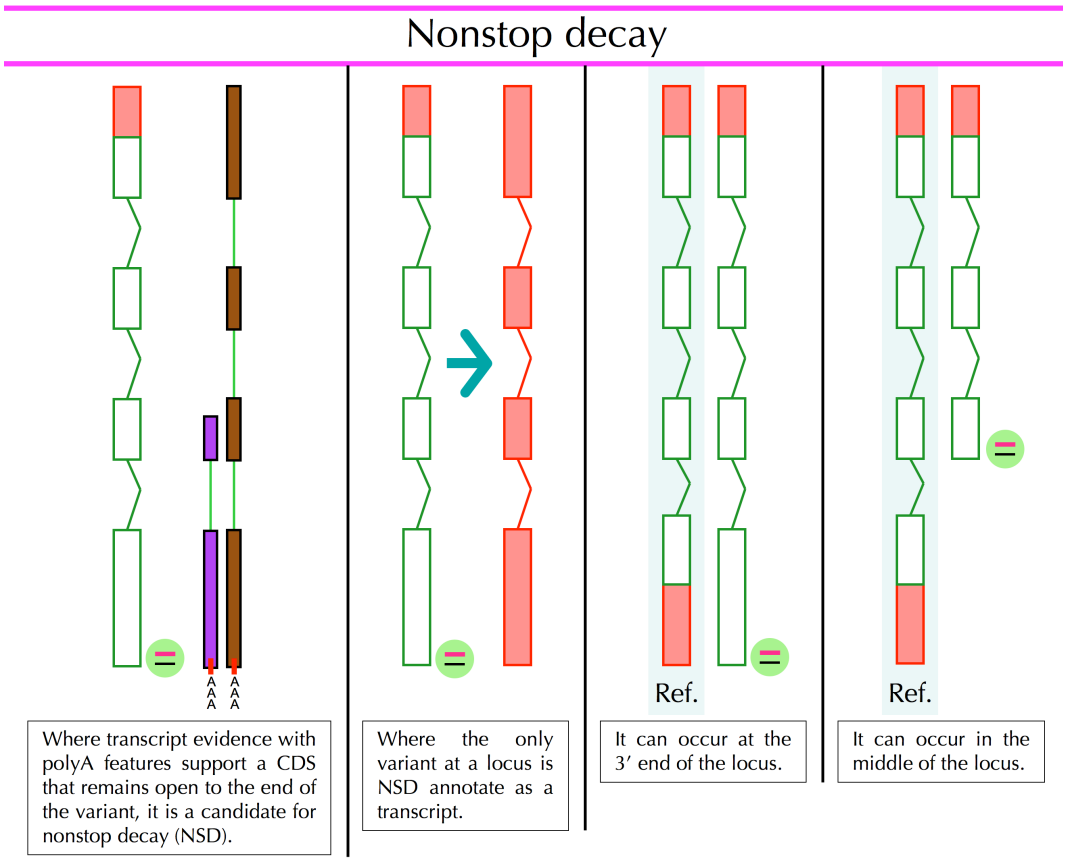


Figure 22: instances of nonstop decay

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 23 to decide on the use of this evidence.

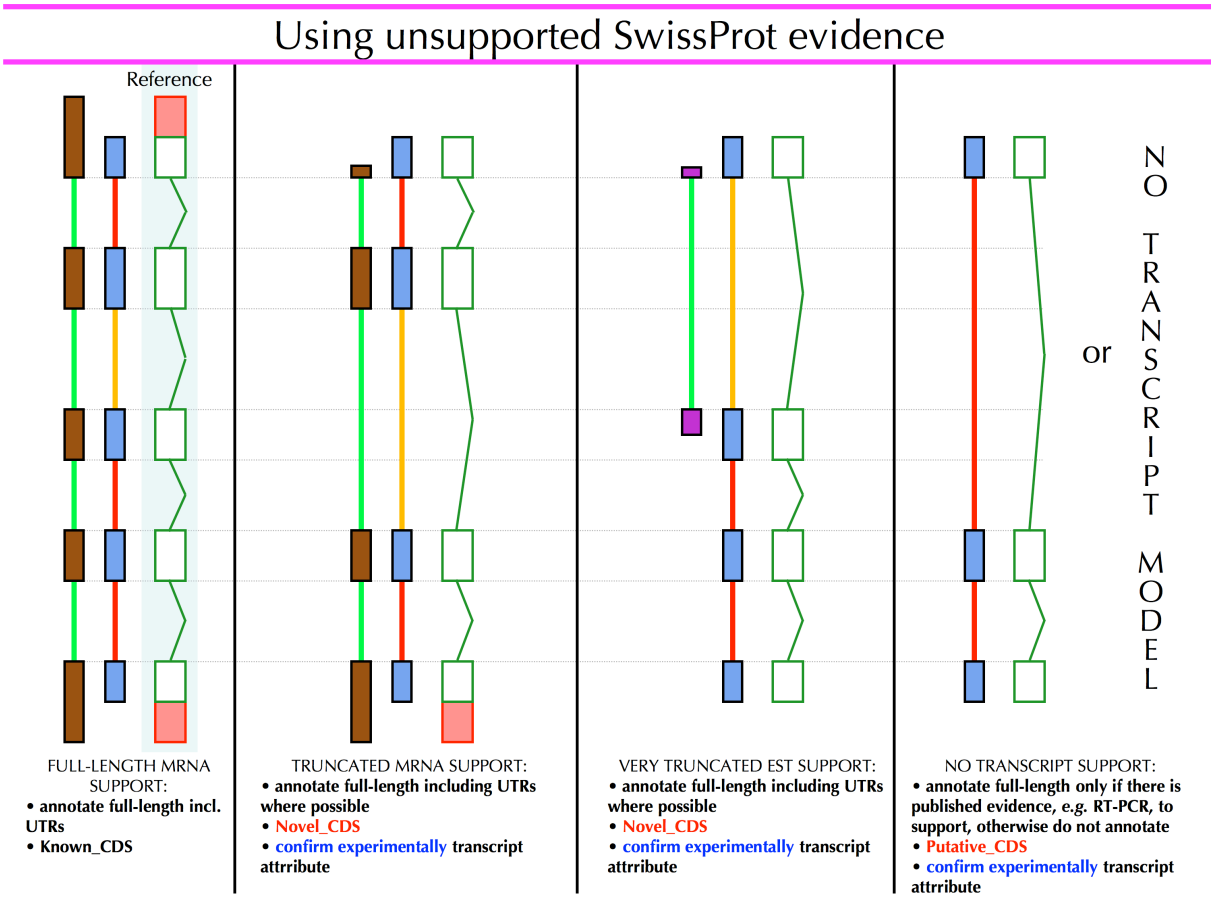


Figure 23: extending or building transcript models using unsupported SwissProt evidence

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.

Transcripts for experimental confirmation

For a variant with a CDS that breaks protein domain structure or is otherwise very different (truncated) from the reference CDS, add a "confirm experimentally" transcript attribute to flag it for possible future experimental confirmation of expression and investigation of expression pattern:

Transcript Attribute: confirm experimentally

Variants without CDS

Transcripts that do not have a CDS (*i.e.* no CDS is annotated because it would not fulfil CDS criteria mentioned earlier) are labelled with one of the following tags.
NOTE: a transcript that overlaps with an exon on the same strand but doesn't share a splice junction becomes a variant of the locus unless there is strong evidence that it should be annotated as a separate locus (Figure 27).

Transcript: the transcript does not fit any of the sub-categories below.

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 functional. 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 16***.
NOTE: especially, but not exclusively, with small genes with one or few (small) introns, a retained intron transcript can be single-exon.
NOTE: shifts in splice donor or acceptor resulting in an exon containing intronic sequence compared to another variant, does not qualify the transcript for retained intron status.
EXCEPTION: another variation upstream of retained intron induces NMD?
> Tag it NMD.
EXCEPTION: the retained intron is the last intron and gives rise to a novel stop?
> Tag it Putative_CDS.
EXCEPTION: the retained intron is a UTR intron and thus doesn't affect CDS?
> Annotate as coding.
EXCEPTION: the retained intron is only supported by other species evidence?
> Don't annotate (unless annotation in this species depends on evidence from closely related species, *e.g.* human transcripts in gorilla).

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

IG_gene: only for immunoglobulin gene building blocks.

NOTE *: only for variant, not for single-transcript stand-alone locus.

Artifact transcripts

Apart from being used for a locus, more often used for variants based on cDNAs with artifactual "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 artifacts are annotated as a variant of the locus. If the alignment around the junction shown in Blixem is incorrect, please adjust the junction so it is fully supported; by their very nature these junctions can generally be annotated at various different supported positions across several bases so the annotator can choose one at random.

NOTE: Artifacts should only be made from species- and locus-specific mRNAs, not from ESTs, nor from any transcript from other species or similar loci.

NOTE: If according to homology evidence the only protein coding variant appears to be an artifact because of an non-splicing "intron" in the 3′ UTR, check whether this "intron" actually represents a polymorphic repeat insertion or expansion (***Figure 20***).

NOTE: mRNAs from the NEDO project are apparently sometimes not completely sequenced, resulting in a submitted sequence that basically represents the 5′ and 3′ sequences falsely joined into one mRNA sequence. Aligned against the genome this will present as an artifact but not likely with the repeated sequence around the junction commonly found in other artifact transcripts.

EXCEPTION: if an artifact transcript has both an artifact event and a genuine variation that is not represented by other evidence, annotate the section of the transcript containing the genuine event up to the artifact event as a normal transcript variant. But still build an Artifact typed transcript representing the entire artifactual cDNA (see Figure 24).

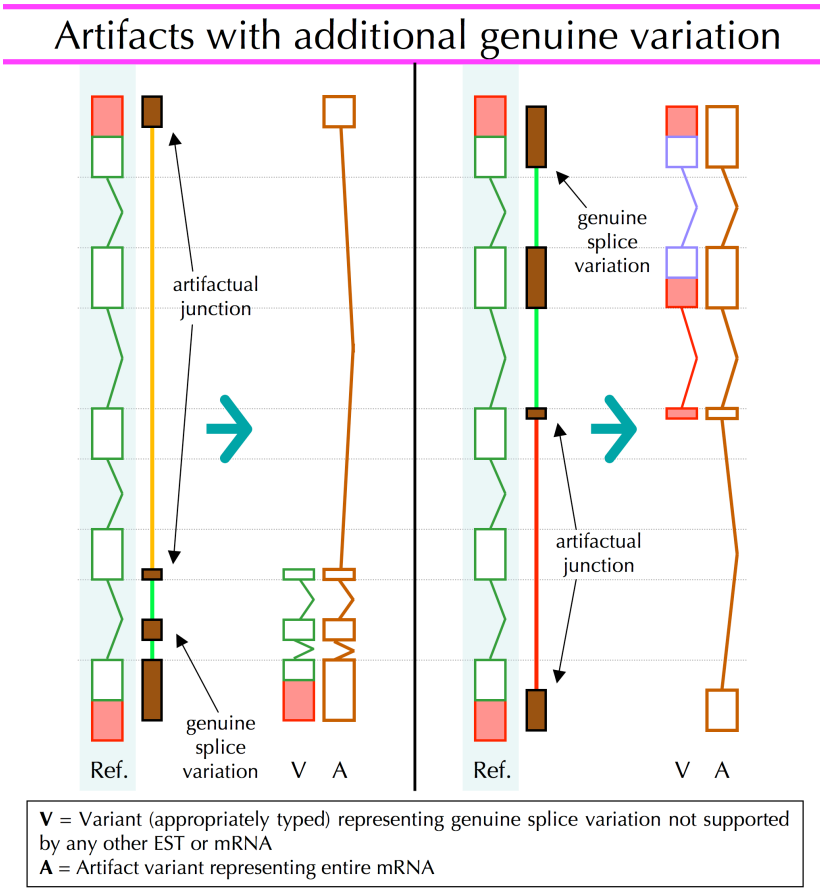


Figure 24: duplication of genuine variation in cases of artifacts with genuine variation

Complex loci

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 confirm experimentally locus attribute 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 Attribute: fragmented locus
Locus Attribute: confirm experimentally

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 Attribute: 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.

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. In these cases the loci are annotated as three separate loci: upstream, downstream and readthrough. For other cases follow the flowchart below (Figure 25). In summary, annotating a separate locus for the readthrough is the default and only a few scenarios deviate from that. Transcripts that share at least one splice junction unique to the readthrough locus (but outside the other loci) will be variants of that locus even if they are not strictly readthrough themselves. Of course any transcript that reads through is a variant of the readthrough locus.
Use the approved symbol and description if available, otherwise use the format as in this example:

Full name novel protein
novel transcript

NOTE: A readthrough locus can consist of a single NMD transcript (Figure 26***D***), in which case the full name is "novel protein".

The following Locus Attribute is added to **all** overlapping loci:

Locus Attribute: overlapping locus

Also add the following Transcript Attribute to any transcript from the readthrough locus that overlaps two or more loci:

Transcript Attribute: readthrough

See also Figure 26 for examples of the various scenarios. It shows only a selection of the numerous permutations that are possible and is meant to be viewed in conjunction with the decision diagram Figure 25). Note the different treatment of cases of readthrough with 5′ and 3′ non-coding loci (C and D) and cases of readthrough transcripts overlapping 5′ UTR (B).

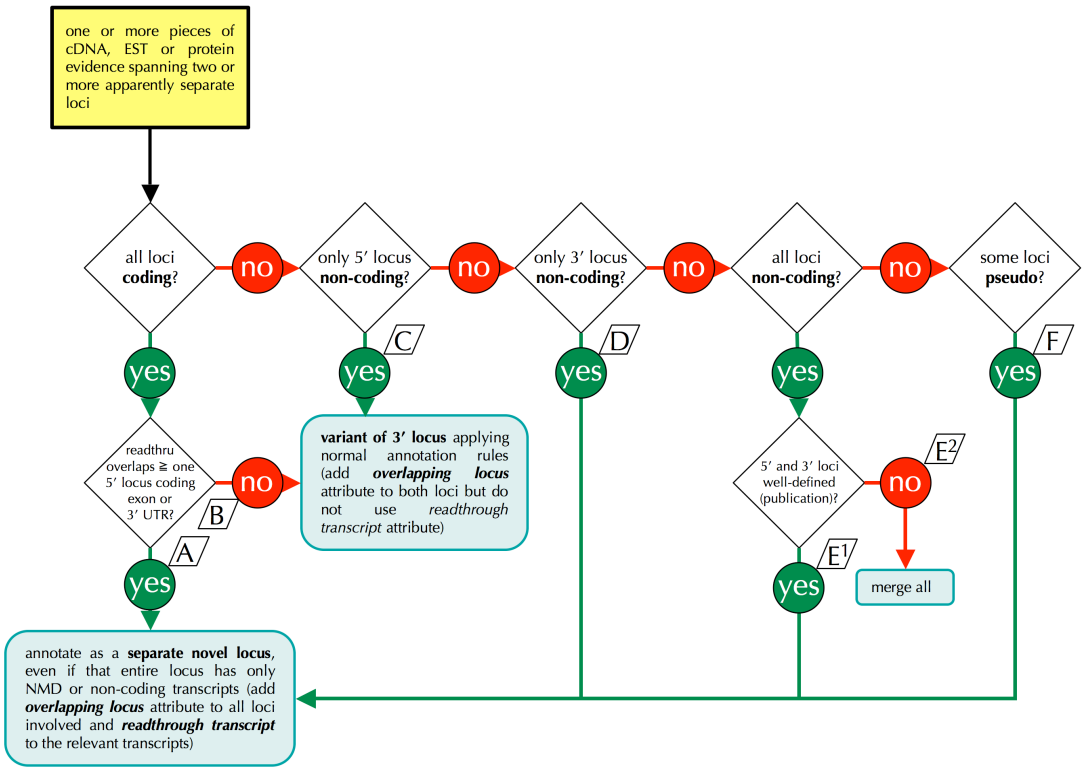


Figure 25: readthrough flowchart

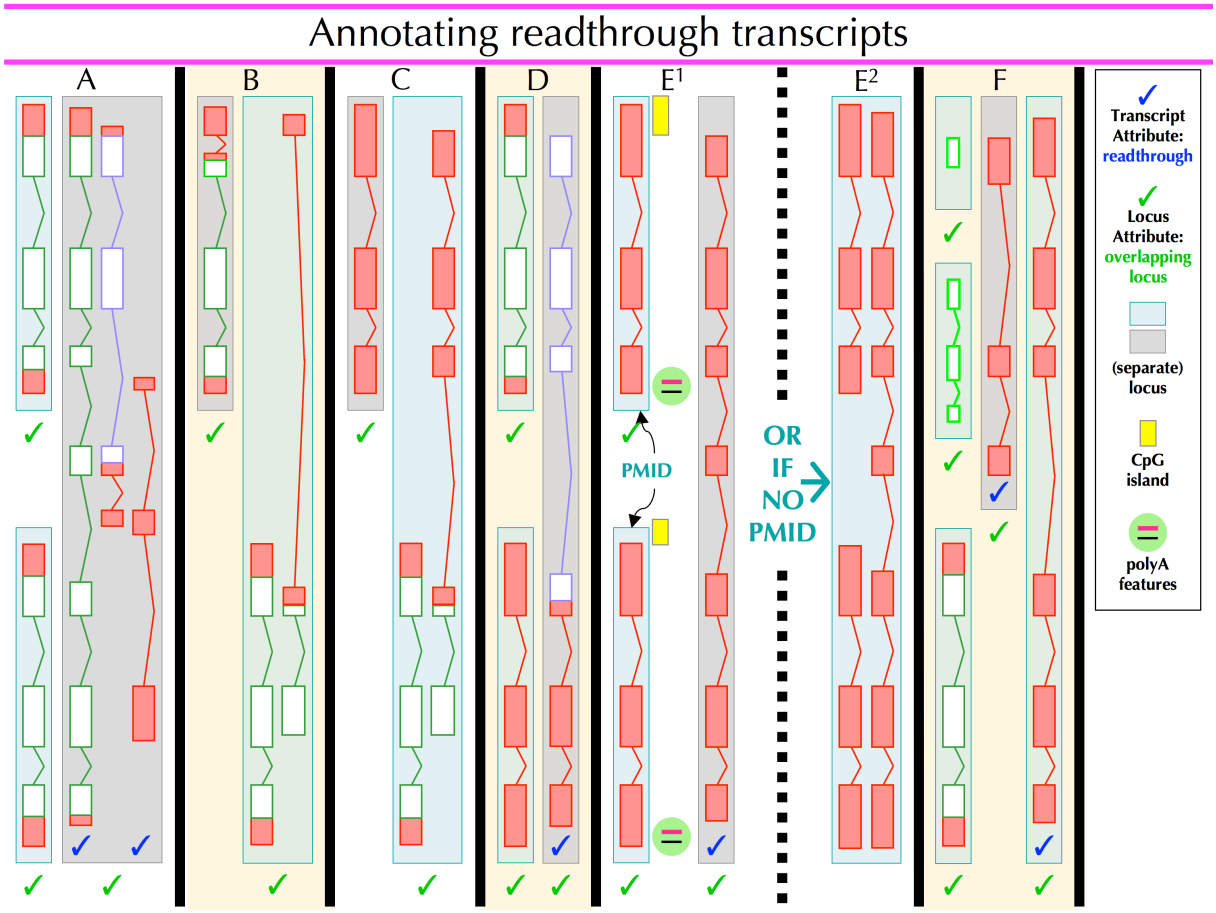


Figure 26: examples of different readthrough scenarios

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 27. If there is overlap, even in the absence of shared splice junctions, the transcript is annotated as a variant of the reference locus. If loci are nested, add the following to all overlapping loci:

Locus Attribute: overlapping locus

If a nested locus is non-coding, it will be biotype Sense_intronic (see the **Non-coding loci**

Loci where none of their variants have a CDS are annotated with one of the following ncRNA biotypes: for more details). A nested locus can be coding, in which case the coding biotype rules apply, but in practice the majority will be non-coding.

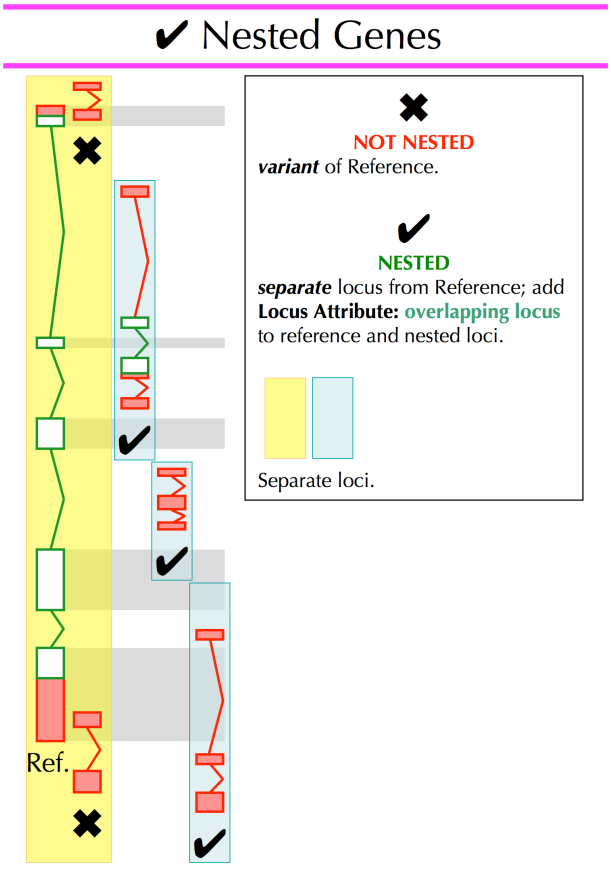


Figure 27: nested genes as separate loci

Naming Genes

This section describes gene nomenclature. Below, the locus **Symbol** is only shown when it needs to be changed. If not shown keep the automatically generated symbol.

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 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 symbols we use is the Corf types and KIAA types.

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

Symbol KIAA0146
Full name KIAA0146

Symbol C22orf23
Full name chromosome 22 open reading frame 23

Known ✔

Known anonymous genes

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

Full name novel protein (FLJ10034)
novel C2H2 type zinc finger protein (0610007P08Rik)

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 FLJ20090)

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 named gene in another species (*i.e.* very high cross-species homology to that one type of protein from different species, gene with the same genomic neighbours in bothe species), in which case it is acceptable to call it an orthologue.

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

Homology to model organism predicted/hypothetical genes

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

Full name novel protein

Novel genes with non-informative matches or non-coding

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 subtype"Putative"), or subtype of "ncRNA":
Full name novel transcript
novel transcript, antisense to Argaph1 and Rpl17

Pseudogenes

Full name of pseudogenes is 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". Never use the parent gene symbol for the pseudogene symbol!

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

For known pseudogenes use their given description and symbol, but do not tag "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 "artifact gene" and "the gene for a TEC" 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 member.

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 + TEC locus

```
Contains a actin, beta pseudogene 8, part of the HIRIP3
gene for HIRA interacting protein 3, the gene for a TEC
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.
```

Biotypes

In the annotation database every transcript annotated is associated with a specific *transcript_biotype*. The hierarchy of these transcript biotypes determines the *gene_biotype* for the locus. See the following VEGA link: http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html

Literature References

polyA signals: • Beaudoing E, ….. , Gautheret D. *Patterns of variant polyadenylation signal usage in human genes.* Genome Res. 2000; 10(7):1001-10. PMID: 10899149

Start codons: • Kozak M. *Regulation of translation via mRNA structure in prokaryotes and eukaryotes.* Gene. 2005; 361:13-37. PMID: 16213112

• Kozak M. *Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes.* Nucleic Acids Res. 1981; 9(20):5233-52. PMID: 7301588

NMD: • Green RE, ….., Brenner SE. *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics.* 2003; 19 Suppl 1:i118-21. PMID: 12855447.

• Lewis BP, Green RE, Brenner SE. *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A. 2003; 100(1):189-92. PMID: 12502788.

• Lareau LF, ….., Brenner SE. *The coupling of alternative splicing and nonsense-mediated mRNA decay.* Adv Exp Med Biol. 2007; 623:190-211. Review. PMID: 18380348.

• Hansen KD, ….., Brenner SE. *Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila.* PLoS Genet. 2009; 5(6):e1000525. PMID: 19543372.

Pseudogenes: • Khachane AN, Harrison PM. *Assessing the genomic evidence for conserved transcribed pseudogenes under selection.* BMC Genomics. 2009; 10:435. PMID: 19754956.

Nonstop decay: • Sundermeier T, ….., Karzai AW. *Studying tmRNA-mediated surveillance and nonstop mRNA decay.* Methods Enzymol. 2008; 447:329-58. PMID: 19161851.

• Akimitsu N. *Messenger RNA surveillance systems monitoring proper translation termination.* J Biochem. 2008; 143(1):1-8. Review. PMID: 17981821.

• Isken O, Maquat LE. *Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function.* Genes Dev.; 21(15):1833-56. Review. PMID: 17671086.

Reference Tables, Figures and Lists

Codon table

non-polar	polar	basic	acidic	(stop codon)

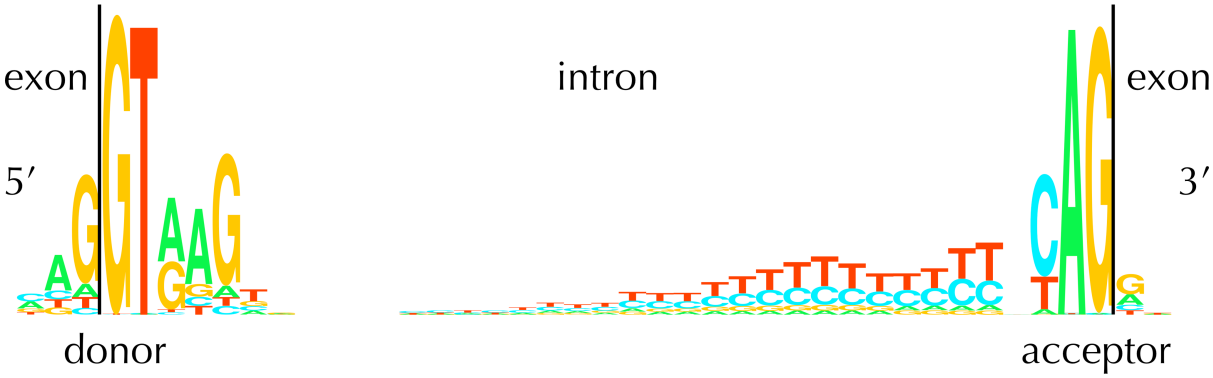
1st base	2nd base: T	2nd base: C	2nd base: A	2nd base: G
T	**TTT** (Phe) Phenylalanine (F)	**TCT** (Ser) Serine (S)	**TAT** (Tyr) Tyrosine (Y)	**TGT** (Cys) Cysteine (C)
	TTC (Phe) Phenylalanine (F)	**TCC** (Ser) Serine (S)	**TAC** (Tyr) Tyrosine (Y)	**TGC** (Cys) Cysteine (C)
	TTA (Leu) Leucine (L)	**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)

Nucleotide degenerate code table

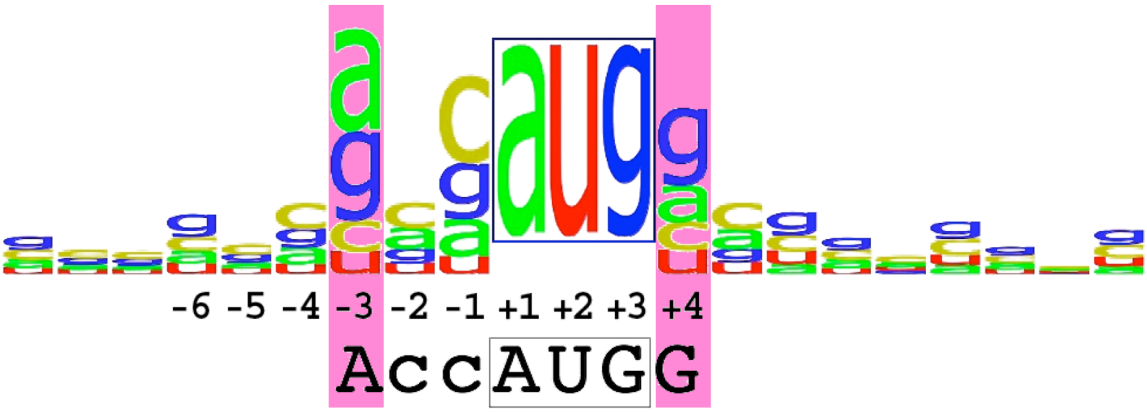
Symbol	Bases represented				Description
W	A	T			weak
S			G	C	strong
M	A			C	amino
K		T	G		keto
R	A		G		purine
Y		T		C	pyrimidine
B		T	G	C	not A (B after A)
D	A	T	G		not C (D after C)
H	A	T		C	not G (H after G)
V	A		G	C	not T/U (V after U)
N	A	T	G	C	any

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)

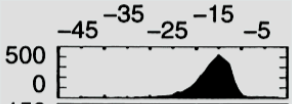
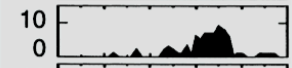
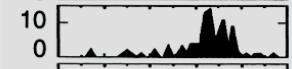
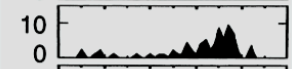
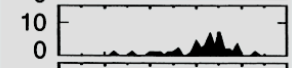
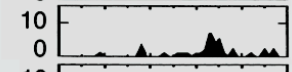
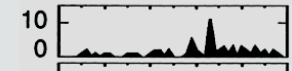
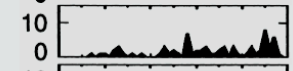
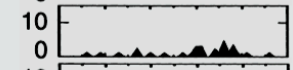


Start codon Kozak sequence



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]
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	
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	

Attributes and controlled vocabulary remarks

Transcript Attribute: upstream ATG
downstream ATG
non-ATG start codon

Transcript Attribute: NMD exception
Transcript Annotation Remark: [PMID <id>, publication reference]
PMID 12345678, Wilming et al. (2007) Nature 447
Transcript Attribute: NMD likely if extended

Transcript Attribute: selenocysteine
Locus Visible Remark: selenoprotein

Transcript Attribute: alternative 5′ UTR

Transcript Attribute: not organism-supported

Transcript Attribute: not best-in-genome evidence
non-submitted evidence

Transcript Annotation Remark: RNA-seq supported
tissue=ovary:SOLEXAG0000014410

Locus Attribute: fragmented locus

Transcript Visible Reamark: 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.

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.

Transcript Attribute: confirm experimentally
Locus Attribute: confirm experimentally

Transcript Attribute: sequence error

Transcript Attribute: readthrough
Locus Attribute: overlapping locus

Locus Attribute: orphan

Transcript Attribute: retained intron 5′ UTR
retained intron CDS
retained intron final

Transcript Attribute: NAGNAG splice site
non canonical U12
non canonical TEC
non canonical conserved
non canonical genome sequence error
non canonical other
non canonical polymorphism

Transcript Attribute: bicistronic

Transcript Attribute: overlapping uORF
upstream uORF