Annotation modifications on human (and vertebrate) transcript and protein records – July 24, 2013

We have recently made some annotation changes to vertebrate, primarily human, RefSeq transcript and protein records. These changes appear on transcript and protein records that are in scope for manual curation, namely accessions that have the NM, NR, NP, NG accession prefix. These changes reflect recent work toward increased transparency with regard to curation decisions and reporting primary sequence evidence support for RefSeq records.

Recent modifications include:

- Exon features
- Reporting support evidence
- Reporting gene and sequence attributes
- Record Title (mitochondrial localization)

Exon features:

Exon features are still annotated on RefSeq transcript records with an NM or NR accession prefix; however, exon numbers are no longer reported. Exon numbers were not stable and were recalculated any time a new or updated RefSeq transcript was available for the gene. The instability of exon numbers lead to confusion and complaints and thus they have been removed.

Exon numbers continue to be provided only on RefSeqGene genomic records based on a subset of available transcript RefSeqs for a GeneID; namely, those selected by locus-specific databases as reference sequence reporting standards. For example, NG_029703.1 includes exon numbers for the GAS6 gene based only on NM_000820.2 but not the other transcript variants for the gene (NM_001143945.1 and NM_001143946.1). More information on the RefSeqGene project is available here: http://www.ncbi.nlm.nih.gov/refseq/rsg/.

Reporting support evidence:

Reporting of transcript support evidence for RefSeq transcript records (with NM or NR accessions) has been expanded and is now reported in a new structured comment with the header 'Evidence Data'. This comment appears on both the transcript and protein record.

We report evidence for the exon combination represented by the RefSeq transcript record based on interpretation of RefSeq and GenBank cDNA, EST, and SRA RNAseq alignments to the reference genome assembly. We report a maximum of two supporting evidence identifiers. The evidence categories reported include the following (presented here as the label and its definition):

- o Transcript exon combination – This category indicates a fully supported (all splice sites) exon combination, with two supporting accessions reported. This category is reported with the evidence code ontology ID ECO:0000332. For example, see NM_005589.3.

- CDS exon combination – This category indicates a fully supported (all splice sites) exon combination spanning the annotated CDS feature only, with two supporting accessions reported. The combination of UTR and CDS exons is supported by logical inference or publication reports. This category is reported with the evidence code ontology ID ECO:0000331. For example, see NM_002075.2.
- RNAseq introns, single sample – This category indicates that all intron positions are supported by the reported sample accessions. This category is reported with the ECO ID ECO:0000348. For example, see NM_001278586.1.
- RNAseq introns, mixed/partial sample support – This category indicates that some or all of the intron positions are supported by more than one sample and no single sample supports all of the intron positions. This category is reported with the ECO ID ECO:0000350. For example, see NM_005589.3
- Transcript is intronless – This category indicates that intronless transcript alignments support an intronless RefSeq transcript. This category is reported with the ECO ID ECO:0000345. For example, see NM_205823.2.

Reporting Gene and Sequence Attributes:

We are now reporting a set of gene and sequence attributes, for RefSeq transcript records (with NM or NR accessions), in a structured comment with the header 'RefSeq Attributes'. This comment appears on both the transcript and protein record.

We report information observed in the sequence, imported from other data sources, or stored in internal databases by NCBI staff scientists during the sequence analysis and gene curation process. The attributes provided are not comprehensive; in other words, the absence of an attribute is not considered an error but simply reflects that the information has not been stored in our internal database at this time.

- bicistronic transcript – A transcript that may produce two distinct proteins from non-overlapping open reading frames. For example, GDF1 and CERS1 (GeneIDs 2657 and 10715; NM_021267.3, NM_001492.4) as supported by PubMed ID 2034669.
- CDS uses downstream in-frame AUG –an upstream in-frame AUG codon exists but a decision was made to annotate the CDS from a downstream start codon. Support for the decision such as a publication, sequence conservation, or protein sequence considerations, is provided. The upstream alternate start codon is annotated as a miscellaneous feature. For example, NM_138448.3 or NM_001206932.1.
- imprinted gene – Transcripts for an imprinted gene are expressed only from the paternal or maternal chromosome. This gene attribute is reported in association with a PubMed ID. For example, NM_016352.3.
- inferred exon combination – This attribute is reported when the exon combination of the annotated CDS feature has been inferred based on a combination of partial transcripts, protein homology, sequence analysis, and publications. This attribute may be reported with a supporting PubMed ID when available. For example, NM_133379.4.
- gene product(s) localized to mito. – This gene attribute is reported on all transcript variants for a gene based on gene and protein nomenclature, homology data, publications, and reports imported into the NCBI Gene database from MitoCarta (PubMed ID 18614015). For example, NM_018394.3.

- non-AUG initiation codon – This attribute is reported for transcripts with a CDS annotated from a non-AUG translation initiation codon based on a published report or inferred from conservation. For example, NM_021182.1.
- polyA required for stop codon – This attribute is used when there is sequence conservation for a protein-coding locus and the stop codon is completed by post-translational polyadenylation. These loci may be pseudogenes rather than protein-coding genes. For example, see NM_001145051.2.
- protein contains selenocysteine – This attribute is reported for transcripts in which a selenocysteine amino acid is encoded by a UGA stop codon based on published reports or conservation with known (published) selenoprotein orthologs. For example, NM_000581.2.
- readthrough transcript – This attribute is reported on transcripts that share exons with two or more distinct genes, indicated by NCBI GeneIDs in the attribute report. For example, NM_007203.4.
- ribosomal slippage – This attribute is reported for transcripts that utilize programmed translational frameshift to encode the protein, based on published reports or conservation with orthologs that have publication support. For example, NM_204916.1.
- undergoes RNA editing – This gene attribute is reported on all transcript variants for a gene based on published reports or conservation with orthologs that have publication support. For example, NM_000826.3.
- unitary pseudogene – This gene attribute is reported for pseudogenes that have a functional ortholog in other species based on published reports or observed by NCBI staff during sequence analysis. The functional ortholog is indicated in the attribute. For example, NR_003227.1.

Transcript record title change (DEFINITION line):

Mitochondrial localization information is no longer included in record title (also called the DEFINITION line). This information is now provided in the RefSeq Attributes section and was potentially misleading in the record title location because it is a comment about the gene rather than the protein. For example, NM_018394.3. Using the Nucleotide Revision History display option (http://www.ncbi.nlm.nih.gov/nuccore/NM_018394.3?report=girevhist), one can see that the title updated on July 7, 2013.

from:

```
DEFINITION  Homo sapiens abhydrolase domain containing 10 (ABHD10), nuclear
            gene encoding mitochondrial protein, transcript variant 1, mRNA.
```

to:

```
DEFINITION  Homo sapiens abhydrolase domain containing 10 (ABHD10),
transcript variant 1, mRNA.
```