# Common Eukaryotic Annotation Errors to Avoid

**Sequence Problems**:

* Contamination
* Ns at the ends of sequences
* internal Ns that represent gaps but have not been converted to assembly gaps

**Annotation Rules that must be followed to avoid common annotation errors:**

<http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation>

<http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission>

[] Every CDS, rRNA, tRNA, ncRNA has a gene feature

[] Every gene has a unique locus\_tag

[] Every locus\_tag has the same prefix, which is registered at GenBank

[] No overlap of CDS and ribosomal RNA features. (check that any overlap of CDS and other RNA features is biologically correct, but we don’t generally validate those overlaps)

[] CDS begins at a start codon or is 5’ partial, and it ends at a stop codon or is 3’ partial

[] CDS has no internal stops

[] The product name is SwissProt-compliant (<http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#CDS>; <http://www.uniprot.org/docs/gennameprot>)

[] There’s only one gene for alternatively spliced coding regions

[] The annotation makes biological sense… eg, having two genes at the same location on opposite strands is a situation that is flagged for submitter review

[] Cannot have just a short intron to adjust for a frameshift… need to do one of these options:

* Change the gene to /pseudo, which suppresses the CDS translation
* Add the “low-quality sequence region” exception to the CDS, to keep the translation but have it flagged as LOW-QUALITY
* If the source is a heterogeneous population and that is causing the sequence issues, then add the “heterogeneous population” exception to the CDS, to keep the translation but have it flagged as LOW-QUALITY
* If you have mRNA or protein evidence, then add the “annotated by transcript or proteomic data” exception to the CDS AND provide the translation in .pep file so that it is included in the final submission file

[] Include an mRNA feature for each translated CDS (<http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#mRNA>). Several things to note are:

* The CDS and its mRNA have the same internal exon/intron junctions, but the mRNA extends beyond the start or stop codon when the 5’UTR or 3’UTR information is known
* Use the same product name for the mRNA and its corresponding CDS.
* If there is no UTR information, then the mRNA's location will agree with its CDS's location, but the mRNA will be partial at its 5' and 3' ends.
* Extend the gene feature to include the entire mRNA.
* If the mRNA is partial, then make the gene partial.

[] The gene location includes all of its features, so extends to the 5’-most and 3’-most ends of its mRNAs

[] The CDS and its mRNA share a protein\_id and transcript\_id, each of which has unique identifiers (<http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#protein_id> and <http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation#transcript_id>)

[] When annotating scaffolds, you need to be careful about crossing gaps-

* A CDS may not cross the gap if the gap size is unknown. Instead, you could have two partial CDS features (and mRNAs) abutting the gap, with a single gene over the whole locus. Alternatively, one of the partial CDS/mRNA features may be deleted if it is very short and there is little or no supporting evidence. If you have a single gene and two partial CDS/mRNA features, you should: (1) add a note to each CDS referencing the other half of the gene, (2) add a note to the gene and CDS features stating, "gap found within coding sequence."
* A CDS can cross the gap if the gap size is known. However, a CDS (or mRNA) should not cross a gap such that over 50% of the translation is X (ie, in the gap). This situation will generate an error. Again, the CDS/mRNA should either be partial up to the gap or split into two partial CDS/mRNA features on either side of the gap, depending upon your confidence in the translation on each side of the gap. In addition, no feature should end inside a gap. Instead, the feature should end at the gap and should be partial.