Table 1. Management of CCDS IDs

Type of change	**Interim Status**[1]	**Final Status**[2]	**CCDS ID**	**version number**
Placement and CDS structure unchanged	Public	Public	unchanged	unchanged
CDS structure modified with partial change in genome annotation placement [3]	Under review, update Reviewed, update pending	Public	unchanged	incremented
New annotation match	NA	Public	created	version 1
Deemed invalid by curation	Under review, withdrawal Reviewed, withdrawal pending	withdrawn	unchanged	unchanged
Withdrawn, other [4]	Public	Withdrawn, inconsistent annotation	unchanged	unchanged

[1] Under review status types indicate that curation is still in progress. Reviewed status types indicate that all collaborators have agreed with the proposed change and finalization is dependent on confirmation following a genome annotation update and subsequent CCDS analysis.

[2] Final status is applied following a genome annotation update and CCDS re-analysis confirms that the expected annotation change is represented in both input datasets.

[3] The CDS structure has been modified in some manner by curation resulting in changes to the genomic exon coordinate data.

[4] Unexpected loss of consistent CDS annotation (not curation-based).

Table 2. Gene sets used in computing RFC scores.

Gene set	Human		Mouse	
	version	loci	version	loci
CCDS	March 30, 2008	16,992	November 28, 2007	16,893
Ensembl*	v49	5,371	v47	6,668
RefSeq*	Build 36.3	5,658	Build 37.1	7,868
Controls	na	4,582	na	na

*The Ensembl and RefSeq sets are the same versions used in constructing the corresponding CCDS sets. The Ensembl and RefSeq loci are the number of loci that do not have CCDS transcripts.

Table 3. Example categories of officially named protein-coding genes lacking CCDS IDs

Category*	Human	Mouse
1. NP_ RefSeq is available for the Gene	1580	2277
2. Reported Genome Assembly problem	205	117
3. No NP_ RefSeq available for the Gene	181	4612
4. Insufficient protein data for RefSeq use	118	3718
5. No protein data associated with the Gene	14	3504

*This represents one approach to define categories of interest that are not included in the CCDS dataset. There are also loci predicted by NCBI and Ensembl pipelines that do not have official nomenclature (not reported here). The subset with official nomenclature has undergone some review by those groups and thus is more likely to be validated as protein coding. The omission from CCDS reflects several factors, reported in rows 1-5. 1) More NP_ RefSeqs are available that lack a CCDS ID due to the the slow update cycle of the CCDS dataset relative to manual curation updates; 2) Some loci lack a CCDS ID because the protein cannot be accurately represented on the reference genome due to assembly gaps or other sequence differences; 3) a NP_ RefSeq record is not yet provided for the locus (so it is automatically out-of-scope for CCDS) for which there are subcategories of loci; 4) loci which do not have sufficient data to confidently annotate a protein (e.g., partial sequence data); or 5) do not have any associated protein data at all (potentially not protein coding loci). Note, there is still uncertainty as to the ‘correct’ number of protein coding genes that should be annotated on the human and mouse reference chromosomes. A recent Ensembl build predicts over 21,000 protein-coding genes, NCBI's spring 2008 annotation release (build 36.3) predicted over 22,000 protein-coding genes, and the publication by Clamp et al. (2007) predicted 20,500 protein-coding genes. We anticipate that the protein-coding loci not currently included in the CCDS database will be validated (or discounted) as additional data becomes available from GENCODE validation, deep RNA sequencing, or proteomics projects.