1-4-2018

Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation.

Shashikant Pujar

Nuala A O'Leary

Catherine M Farrell

Jane E Loveland

Jonathan M Mudge

See next page for additional authors

Authors
Shashikant Pujar, Nuala A O'Leary, Catherine M Farrell, Jane E Loveland, Jonathan M Mudge, Craig Wallin, Carlos G Girón, Mark Diekhans, If Barnes, Ruth Bennett, Andrew E Berry, Eric Cox, Claire Davidson, Tamara Goldfarb, Jose M Gonzalez, Toby Hunt, John Jackson, Vinita Joardar, Mike P Kay, Vamsi K Kodali, Fergal J Martin, Monica McAndrews, Kelly M McGarvey, Michael Murphy, Bhanu Rajput, Sanjida H Rangwala, Lillian D Riddick, Ruth L Seal, Marie-Marthe Suner, David Webb, Sophia Zhu, Bronwen L Aken, Elspeth A Bruford, Carol J Bult, Adam Frankish, Terence Murphy, and Kim D Pruitt

Published online 6 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D221–D228***
doi: 10.1093/nar/gkx1031

Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation

Shashikant Pujar[1], Nuala A. O'Leary[1], Catherine M. Farrell[1], Jane E. Loveland[2], Jonathan M. Mudge[2], Craig Wallin[1], Carlos G. Girón[2], Mark Diekhans[3], If Barnes[2], Ruth Bennett[2], Andrew E. Berry[2], Eric Cox[1], Claire Davidson[2], Tamara Goldfarb[1], Jose M. Gonzalez[2], Toby Hunt[2], John Jackson[1], Vinita Joardar[1], Mike P. Kay[2], Vamsi K. Kodali[1], Fergal J. Martin[2], Monica McAndrews[4], Kelly M. McGarvey[1], Michael Murphy[1], Bhanu Rajput[1], Sanjida H. Rangwala[1], Lillian D. Riddick[1], Ruth L. Seal[5], Marie-Marthe Suner[2], David Webb[1], Sophia Zhu[4], Bronwen L. Aken[2], Elspeth A. Bruford[5], Carol J. Bult[4], Adam Frankish[2], Terence Murphy[1,*] and Kim D. Pruitt[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [3]University of California Santa Cruz Genomics Institute, Santa Cruz, CA 95064, USA, [4]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME 04609, USA and [5]HUGO Gene Nomenclature Committee, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 20, 2017; Revised October 13, 2017; Editorial Decision October 16, 2017; Accepted October 20, 2017

ABSTRACT

The Consensus Coding Sequence (CCDS) project provides a dataset of protein-coding regions that are identically annotated on the human and mouse reference genome assembly in genome annotations produced independently by NCBI and the Ensembl group at EMBL-EBI. This dataset is the product of an international collaboration that includes NCBI, Ensembl, HUGO Gene Nomenclature Committee, Mouse Genome Informatics and University of California, Santa Cruz. Identically annotated coding regions, which are generated using an automated pipeline and pass multiple quality assurance checks, are assigned a stable and tracked identifier (CCDS ID). Additionally, coordinated manual review by expert curators from the CCDS collaboration helps in maintaining the integrity and high quality of the dataset. The CCDS data are available through an interactive web page (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) and an FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). In this paper, we outline the ongoing work, growth and stability of the CCDS dataset and provide updates on new collaboration members and new features added to the CCDS user interface. We also present expert curation scenarios, with specific examples highlighting the importance of an accurate reference genome assembly and the crucial role played by input from the research community.

INTRODUCTION

Accurate and comprehensive whole genome annotation of the human and mouse reference genomes is essential to support many areas of scientific inquiry, including medical research. RefSeq (1) annotation from NCBI (National Center for Biotechnology Information) and Ensembl (2) annotation from EMBL-EBI (The European Molecular Biology Laboratory's European Bioinformatics Institute) provided for these species are the primary reference resources through which biological data are interpreted and reported by the scientific community. The analytical workflows used by these separate projects are conceptually similar in that they both produce gene annotations based on a combination of computational pipelines and manual curation, largely based on the interpretation of transcriptomics and other experimental data. Ultimately, however, Ensembl and NCBI have developed different rules and guidelines for au-

*To whom correspondence should be addressed. Tel: +1 301 402 0990; Fax: +1 301 594 5166; Email: murphyte@ncbi.nlm.nih.gov

tomated and manual annotation or curation, and this has led to differences between the sets of genes, transcripts and proteins currently found in these datasets. Such inconsistencies can present a challenge to the scientific community in their efforts to interpret biological data; for example, when a disease-associated variant is found to occur in a protein-coding transcript in one dataset, but a non-coding model in another. Furthermore, the regular emergence of new data types and methodologies with which to identify novel transcripts and to gain insights into their functionality mean that these datasets have the potential to include additional divergence with each new release.

The Consensus Coding Sequence (CCDS) collaboration was formed in 2005 to address the issue of discrepancies between Ensembl and NCBI genome annotations by producing a consensus dataset of protein-coding regions with identical coding sequence (CDS) coordinates on the human and mouse reference genomes in both annotations. Consensus protein-coding regions, identified by stable and tracked identifiers (CCDS IDs), and related metadata, are accessible through a public search page (www.ncbi.nlm.nih.gov/CCDS). In addition, data are available for bulk download from an FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). A detailed description of methods to access CCDS data, the CCDS workflow, curation processes and quality assurance (QA) tests involved in creating the dataset, were published previously (3–5).

The core of the collaboration relies on consensus building across members, including NCBI who provide the RefSeq annotation, the Ensembl Genebuild and Ensembl HAVANA (Human and Vertebrate Analysis and Annotation) groups at EMBL-EBI who produce the GENCODE (6) gene set, University of California Santa Cruz (UCSC), and the two nomenclature authorities, HUGO Gene Nomenclature Committee (HGNC) (7) and Mouse Genome Informatics (MGI) (8) who provide standardized human and mouse gene symbols and names, respectively. To maintain high-quality annotation, expert curators from the collaborating groups continuously review and discuss CCDS IDs flagged for review by QA tests, collaboration members and users. Curators also review regions where there are differences in protein annotation between the NCBI and Ensembl genome annotations, to try to reach consensus using all available data types through a voting process, which is described in an earlier publication (4). The combination of CDS annotation concordance across two groups with different annotation methods and policies, and the regular review of the resource by expert curators make CCDS data, a stable and high-confidence option for users focused on protein-coding genes that are annotated consistently in the two major annotation databases. Therefore, CCDS data have been used in genome analyses, such as comparison of whole-genome sequencing and whole-exome sequencing for effective detection of disease-causing mutations (9), large-scale evaluation of proteomics data to determine if genes have a dominant protein isoform (10) and high-throughput exome coverage analysis of clinically relevant cardiac genes (11). In addition, CCDS data are used to design commercial exome microarrays (12).

In this manuscript, we present the current status of the CCDS collaboration, describe the updates that have been made to the CCDS resource since our last report, and we highlight some of the challenges and future plans of this ongoing and active collaboration.

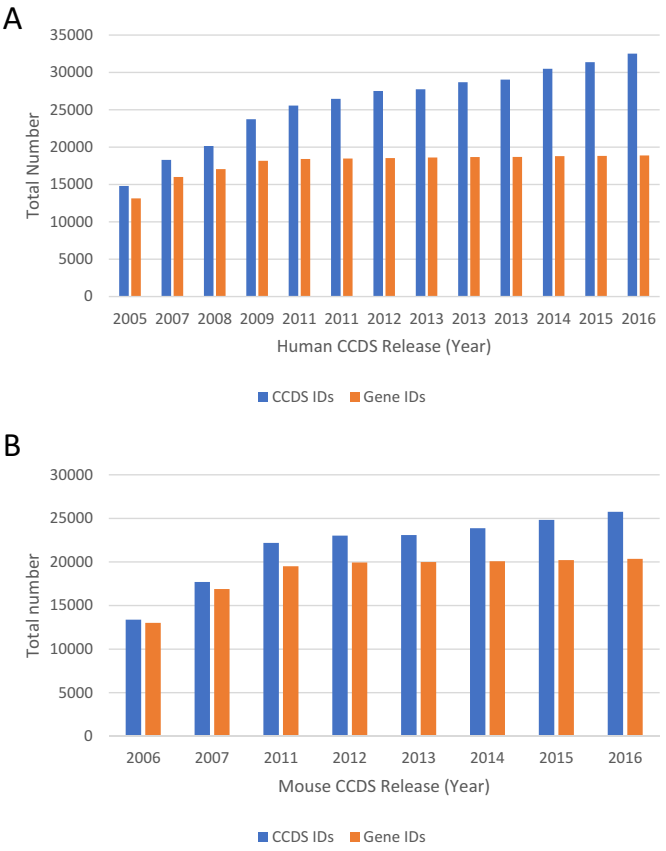

Figure 1. Number of CCDS IDs and genes represented in the human (**A**) and mouse (**B**) CCDS releases. The *X*-axis indicates the year in which a CCDS dataset was made public. Details about CCDS releases are available on the CCDS Releases and Statistics web page (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi?REQUEST=SHOW_STATISTICS).

GROWTH OF THE CCDS DATASET

The CCDS collaboration releases an update of the human and mouse CCDS datasets approximately once each year, following annotation updates of the reference genomes by either NCBI or Ensembl, or both annotation groups. Since the first human and mouse data releases in 2005 and 2006, respectively, the CCDS dataset has seen a steady growth in the number of new coding regions (CCDS IDs) as well as the number of genes that have at least one CCDS ID in both the human (Figure 1A) and mouse (Figure 1B) datasets. This growth reflects increasing concordance in NCBI and Ensembl protein-coding region annotations over the years. The most recent CCDS releases in human (Release 20) and mouse (Release 21) contain 32 524 and 25 757 CCDS IDs represented by 18 892 and 20 354 genes, respectively, and cover ∼33.2 Mb (1.03%) of the human and 33.7 Mb (1.24%) of the mouse reference genomes. Notably, the growth in recent years is largely due to the addition of new CCDS IDs representing alternatively spliced transcripts within existing protein-coding genes, and this trend has continued un-

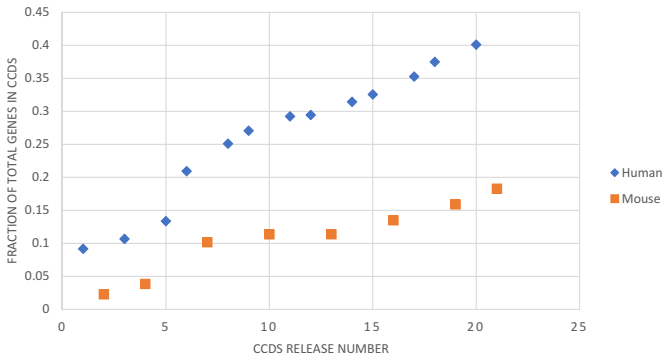

Figure 2. Fraction of all genes in a CCDS release that are represented by at least two current CCDS IDs.

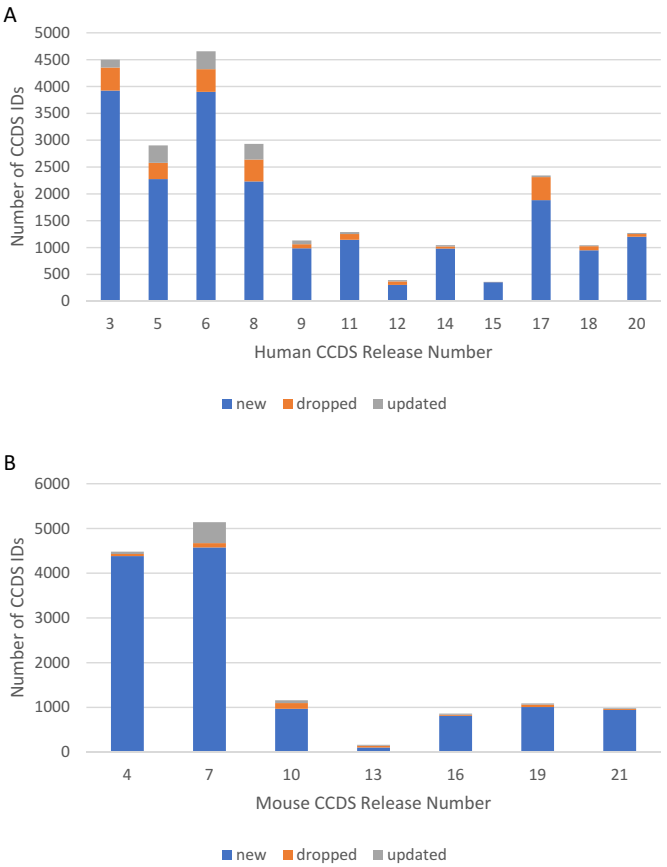

Figure 3. Changes in the human (**A**) and mouse (**B**) datasets with every new CCDS release. 'New' = new CCDS IDs added; 'dropped' = CCDS ID present in the previous release but withdrawn in the subsequent release; 'updated' = CCDS IDs that have an incremented accession version compared to the previous release, indicating a sequence update in the coding region.

abated up to the current human and mouse releases. This is evident from Figure 2, which shows a growing trend in the fraction of total genes in the CCDS dataset that are represented by at least two protein isoforms in both the mouse and the human CCDS sets. The increase in detectable splicing variation can be attributed to the identical annotation of alternatively spliced transcript variants resulting from the incorporation of new data types into automated pipelines, as well as curation workflows of the two annotation groups. It also likely reflects an end to the locus-by-locus 'growth phase' of human annotation as both annotation groups have manually annotated most human genes in their respective annotation catalogs, and are expected to reach the same goal in mouse annotation in the next few years. Nonetheless, the protein-coding gene count remains dynamic in both annotation catalogs; we anticipate that both Ensembl and NCBI will continue to add and remove protein-coding genes in future releases.

Notably, despite a greater number of protein-coding genes in the mouse than in the human dataset (as seen in Figure 1), the Figure 2 graph shows lower numbers for multiple isoform representation in the mouse data. This is likely due to the prioritization of curating the human annotation by both annotation groups. In recent years, however, mouse curation has received increasing focus at both Ensembl and NCBI, and this is expected to narrow the gap in the number of CCDS IDs between the human and mouse datasets in future CCDS releases. New human and mouse CCDS releases are planned before the end of 2017.

The general slow-down in growth over recent releases (Figures 1 and 2) also suggests increasing stability in the dataset. To evaluate dataset stability, we assessed the rates of change for CCDS IDs between releases, including additions of new CCDS IDs, updates to existing CCDS IDs (indicated by an increment in the CCDS ID version) and withdrawals of CCDS IDs. Our analyses (Figure 3) show that larger numbers of CCDS updates and withdrawals occurred between earlier CCDS releases than recent CCDS releases. These results indicate that the CCDS dataset is becoming increasingly stable, not only at the level of new additions, but also at the level of individual CCDS ID alteration.

CCDS DATABASE UPDATES

Collaborators

In May 2014, the official human and mouse nomenclature groups, HGNC and MGI, respectively, replaced UCSC as voting members of the collaboration. UCSC is no longer an active participant in CCDS curation, although they still provide QA input on pseudogenes, orthology and conservation during automated CCDS analysis. Representatives from HGNC and MGI are involved in CCDS policy decisions and in review of specific annotation cases. As voting members in the CCDS annotation review process, they provide input on all cases of conflict between NCBI and Ensembl annotations, highlight individual genes for consideration and raise policy issues for discussion. Their involvement also facilitates timely nomenclature updates prompted by CCDS curation.

Guidelines

Coordinated curation of human and mouse coding region annotation by experts in the CCDS collaboration is crucial for maintaining the high quality of the CCDS dataset. As different collaborating groups may follow di-

verse rules for curation, a common set of guidelines was established for consistent and efficient curation within the CCDS collaboration. These guidelines are available on the CCDS web page (https://www.ncbi.nlm.nih.gov/projects/CCDS/docs/CCDS_curation_guidelines.pdf) and are described in detail in a previous report (4). Curation guidelines need to be updated in response to discussions among collaborators, feedback from research groups and emerging data types. Several updates to CCDS guidelines have been made since our last publication (5) and they reflect changes in CCDS curation policies, which are described in previous reports (4,5). The updates include rules pertaining to the selection of translation start codons and guidelines for the representation of nonsense-mediated mRNA decay (NMD), inferred exon combination and readthrough gene representations. Section 2A, which contains guidelines for start codon selection, was updated to add rules to allow the use of an in-frame downstream start codon instead of the first start codon in the open reading frame (ORF). Specific case examples were included to illustrate scenarios and the support needed to choose the downstream start codon.

The existence of upstream ORFs (uORFs) in protein-coding transcripts were previously considered to be evidence for the removal of the protein-coding capacity of a transcript. In literature-led discussions within the collaboration, it was decided that uORFs are predominantly regulatory rather than deterministic, so they are no longer considered in annotating the CDS.

A new section (Section 2B) was added to the guidelines to describe rules pertaining to CDS annotation in cases where the location of the translation start codon suggests that the transcript may be subject to NMD. Another section (Section 2D) was added to describe the CDS annotation for transcripts that lack full-length support, i.e. a transcript archived in an International Nucleotide Sequence Database Collaboration (INSDC) (13) database and includes at least the entire coding region. In some of these cases, a full-length protein can be inferred from homology, orthology or publications. In other cases, the coding region is inferred when the gene contains cassette exons that are supported individually by transcript, conservation or published data, but full-length support is lacking. A good example of this is *TTN* (CCDS59435.1) which has 363 exons.

A readthrough transcript is annotated when neighboring genes share transcripts that overlap at least one exon per gene. Historically a '3-gene model' was used where the readthrough transcript is considered a part of a separate gene. A CCDS was generated when there was NCBI and Ensembl consensus annotation of a coding readthrough transcript. In some cases, where the overlapping region included only the untranslated region (UTR) of the upstream gene, the readthrough transcript could share the CDS (and hence, the CCDS ID) with the downstream genes, which caused confusion. As transcriptional data are increasing at an exponential rate, we are finding that the existence of readthrough transcripts is very widespread and most readthrough transcripts are likely to be non-functional. Hence, they will now only be considered for inclusion in the CCDS dataset when there is strong experimental evidence for their existence and the CDS is unique.

User interface

A CCDS report page includes links to genome browsers in the 'Chromosomal Locations' section (purple icons) that display the genomic span of the coding sequence ('Genome Browser links') or genomic span of individual coding exons (browser links in the exon table). In 2014, an additional link (purple 'S' icon) was included in the last column of the report table at the top of the CCDS report page. This link opens an interactive combined graphical display (Figure 4) of the NCBI and Ensembl annotations associated with the CCDS ID, using the NCBI Sequence Viewer tool**.** The graphical display offers several options to customize the browser view, including the 'Tracks' button, which allows a user to load additional data tracks to the default view.

The curated SwissProt subset of the UniProtKB (14) database provides additional data relevant to a protein, including function, protein features, subcellular location, expression and structure. To provide access to this data to CCDS users, since January 2014, CCDS reports include a section that displays the UniProtKB/SwissProt accession, including the specific isoform, that matches the CCDS (example CCDS4565.1). The hyperlinked accession number in the 'Related UniProtKB/SwissProt' column provides the user a direct link to the UniProtKB/SwissProt record. These CCDS:SwissProt accession matches are also available in the CCDS2UniProtKB.[release_date/current].txt files in the CCDS FTP site.

Review status

While the CCDS dataset results from concordant annotation in two independent annotation sets, it also reflects manual review by curators from the Ensembl and NCBI annotation groups and by curators in the CCDS collaboration. The greatest value of manual annotation is in the insights it can provide into the functionality of a given transcript, i.e. our understanding of what that transcript does (for example, encodes a protein) and the ability to critically assess the validity of all primary data. In 2017, a 'Review Status' section was added to the CCDS report page above the 'Sequence IDs' table to convey to users if a CCDS ID has been reviewed. Table 1 summarizes the different categories of review status with a brief description of each category. The review status depends on manual review carried out at two levels. The first level of review is performed by curators in the individual annotation groups who review genes, transcripts and proteins as a part of their manual annotation process, and this review is independent of the CCDS workflow. Transcripts and proteins are then flagged within each annotation set to indicate their review status. NCBI-RefSeq accessions are flagged as 'validated' or 'reviewed'. In the Ensembl annotation set, a transcript that is manually annotated has a VEGA (Vertebrate Genome Annotation) (15) accession (with 'OTT' prefix) in addition to an Ensembl ('ENS' prefix) accession, which is indicated in the Transcript window of the Ensembl genome browser. When the two annotations are compared to generate the CCDS set, transcripts and proteins that have gone through this level of review, and are annotated identically in the Ensembl and NCBI sets, result in CCDS IDs with a 'Reviewed (by RefSeq and HA-

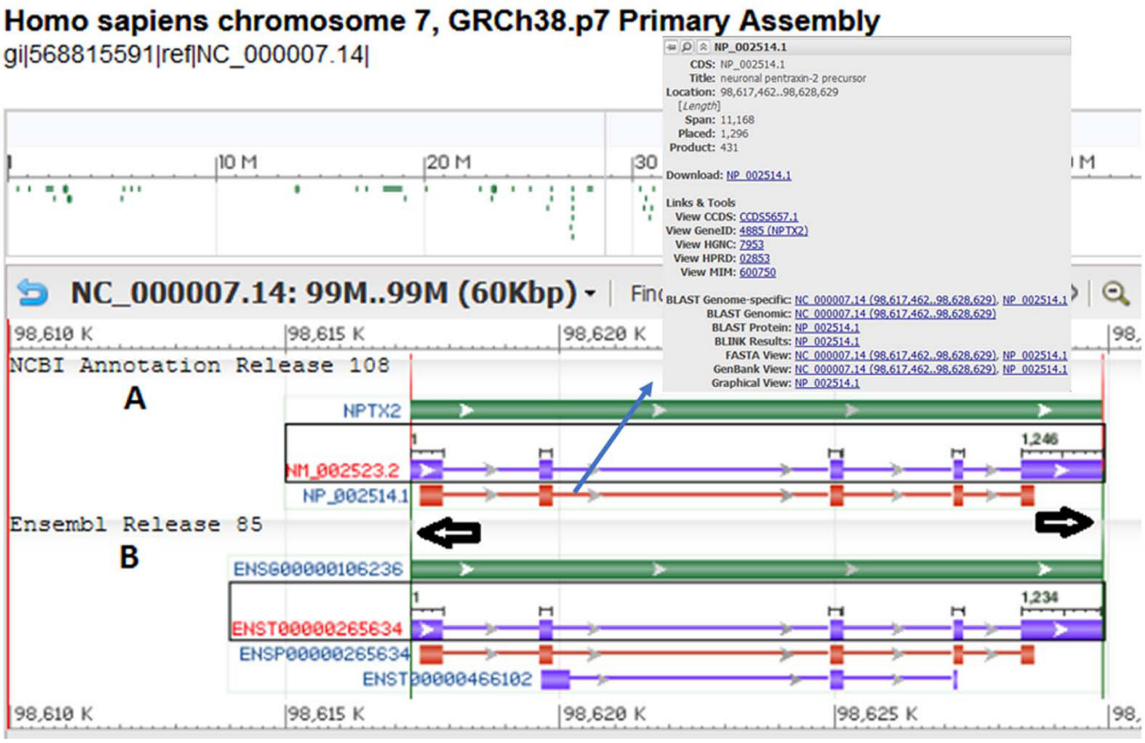

Figure 4. A view of the graphical display accessed from the report page of CCDS3542.1 (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi?REQUEST=ALLFIELDS&DATA=CCDS3542&ORGANISM=0&BUILDS=CURRENTBUILDS) using the purple 'S' icon. (**A**) Transcripts and proteins from NCBI Annotation Release 108. (**B**) Transcripts and proteins from Ensembl Release 85. The green bar indicates the gene; transcripts are shown in purple and proteins are shown in red color. Positioning the cursor over any of these objects (gene, transcript or protein) opens a tool tip which includes additional information and links. Proteins in the NCBI annotation display that are in the CCDS set include a link to the CCDS ID in the tool tip. The gray box to the right (indicated by vertical arrow) is the tool tip corresponding to the protein accession NP_002514.1. Differences between any two objects can also be revealed as vertical lines (indicated by horizontal arrows) when the objects (NM_002523.2 and ENST00000265634 in the figure) are selected using the 'Control' or 'Command' button on the keyboard.

VANA)' status (example CCDS83093.1). It is noteworthy that the VEGA genome browser was retired and archived early in 2017 but manually annotated transcripts continue to be accessible in the Ensembl genome browser. In Ensembl release files, manually annotated transcripts are indicated as ensembl_havana or havana.

A second level of manual review includes the review of CCDS IDs flagged by QA tests for errors or inconsistencies. In addition, specific review cases are raised by individual collaborators or external databases (e.g. UniProt/SwissProt). CCDS users can also request the review of CCDS IDs through the user contact interface (https://www.ncbi.nlm.nih.gov/CCDS/UserRequest/UserRequest.cgi). Internally, these cases either involve a discussion and voting process described in an earlier report (4) or have a 'Public Note' on the CCDS report page explaining an update that was made to the CCDS record. This level of review results in the CCDS ID being designated as 'Reviewed (by CCDS collaboration)' (example CCDS48347.1). CCDS IDs that meet both above-mentioned levels of review get the 'Reviewed (by RefSeq, HAVANA and CCDS collaboration)' label (example CCDS16957.2). Conversely, CCDS IDs that do not meet any of these levels of review are assigned the 'Provisional' review status (example CCDS45069.1).

Figure 5 shows the distribution of human and mouse

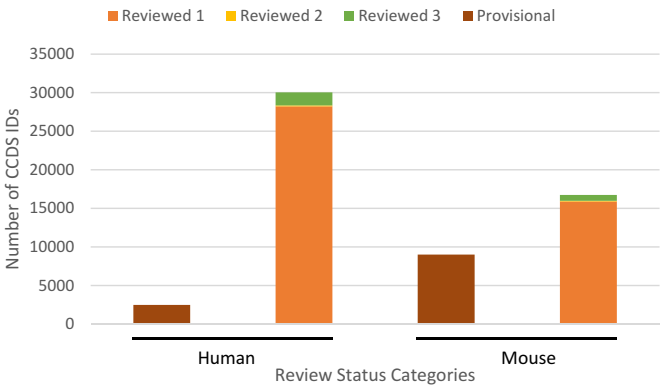

Figure 5. Distribution of human and mouse CCDS IDs by their 'Review status' in the current human (Release 20) and mouse (Release 21) CCDS releases at the time of data freeze. Details of the review status categories and sub-categories are provided in Table 1. Reviewed 1 = CCDS IDs reviewed 'by RefSeq and HAVANA', Reviewed 2 = CCDS IDs reviewed 'by CCDS collaboration', Reviewed 3 = CCDS IDs reviewed 'by RefSeq, HAVANA and CCDS collaboration'.

CCDS IDs among these review categories. Review status is available only for the CCDS IDs of the current release. At the time of the data freeze for the CCDS release, the current human CCDS set (Release 20) had 30 181 (out of a 32 524 total) 'Reviewed' CCDS IDs, while the current mouse set

Table 1. Description of CCDS 'Review Status' categories

Review status category	Public description of category	Detailed description
Provisional	'this record has not been manually reviewed by the collaboration'	The CCDS ID does not have a 'validated' or 'reviewed' RefSeq or a VEGA accession associated with it; nor was it reviewed by the CCDS collaboration
Reviewed	• 'by RefSeq and HAVANA'	• The CCDS ID is associated with at least one 'validated' or 'reviewed' RefSeq AND at least one VEGA accession, which are manually reviewed by curators at NCBI and Ensembl-HAVANA groups, respectively.
	• 'by CCDS collaboration'	• The CCDS ID was reviewed by curators in the CCDS collaboration.
	• 'by RefSeq, HAVANA and CCDS collaboration'	• The CCDS ID meets both 'by RefSeq and HAVANA' and 'by CCDS collaboration' review requirements.

(Release 21) had 16 740 (out of a 25 757 total) 'Reviewed' CCDS IDs.

Expert curation

The CCDS collaboration continues to provide expert curation support to the human and mouse datasets. Since our last publication, the CCDS collaboration reviewed several genes which led to improved annotation that is more consistent with the supporting data. For example, a new gene, *ASDURF* (*ASNSD1* upstream reading frame) was defined based on transcript data and conservation of the 96 amino-acid protein, which was included earlier as a product of the *ASNSD1* gene (Human CCDS2300.1; Mouse CCDS14954.1, CCDS69888.1). The CCDS collaboration reviews genes that are currently annotated as non-coding genes, but where recent evidence indicates they may encode small proteins. For example, the human gene *SMIM27* (NCBI GeneID:100129250, ENSG00000235453, formerly *TOPORS-AS1*) and its mouse ortholog *Smim27* (NCBI GeneID:100504309, ENSMUSG00000028407) are now annotated as protein-coding genes, based on orthology, ribosome profiling elongation and PhyloCSF (16) data.

On several occasions, the CCDS collaboration has worked with scientific research groups to improve annotation of genes. A notable example of the benefit of cooperative expert curation is reflected in the updated annotation of human *MIA2* and *CTAGE5* and their mouse orthologs, following a discussion of these genes by the CCDS collaboration, prompted by a review request from the UniProt group. These genome features were annotated as independent genes by NCBI (Annotation Release 108) and Ensembl (Release 85). CCDS collaborators merged the two genes into a single gene, *MIA2*, based on published data on the human (17) and mouse (18) genes, direct consultations with the research groups and new data submitted to INSDC (KX388743.1, representing the TANGO1-like transcript, TALI, which contains both *MIA2*-specific and *CTAGE5*-specific exons). The *CTAGE5* protein was retained as a splice variant of *MIA2* and a new transcript variant (RefSeq transcript NM_001329214.2; Ensembl transcript ENST00000640607.1), based on TALI, was created. Similar changes were made in the representation of the mouse *Mia2* gene as well. Following these changes, the nomenclature groups agreed to merge their records and retain the symbol *MIA2* and *Mia2*, for the human and mouse genes, respectively. Such annotation changes made by CCDS curators are reflected in the subsequent updates of the NCBI and Ensembl genome annotations. The CCDS dataset reflects the change in the new CCDS build, which, as mentioned earlier, is typically released after NCBI and Ensembl produce updated versions of their gene sets.

In addition to the manual review of specific genes or CCDS IDs, the collaboration regularly reviews targeted lists with a common goal. For example, curators recently reviewed a list of around 250 genes that are annotated by both Ensembl and NCBI, but have different gene biotypes (typically protein-coding versus pseudogene versus long non-coding RNA). As a result of this review, both annotation groups agreed on consistent gene biotypes for about 70 genes. Where disagreements remain, a common factor is the lack of sufficient transcriptomic or proteomic evidence to confirm or confidently suggest a biotype. These cases will be subject to iterative review as additional data become available. A second task involved the review of genes (86 human and 130 mouse genes) that were represented in the CCDS database, but had differing gene symbols in the NCBI and Ensembl databases. This review led to the correction of gene symbols in both annotation sets based on standard names provided by the respective nomenclature authorities. Last year, the CCDS collaborators undertook a review of human-mouse orthologs to ensure that both orthologs were represented in the CCDS dataset when supporting data were available. Such targeted curation on a regular basis ensures the accuracy and consistency of the CCDS dataset.

Sometimes, accurate annotation of coding regions is limited by factors that are beyond the control of annotation pipelines and expert curators. Despite several rounds of improvement, human and mouse reference genome assembly errors continue to pose a challenge for accurate CCDS annotations. CCDS curators continue to report genome problems to the GRC (Genome Reference Consortium) (19), who provide curation support to improve the reference genome assemblies of select organisms, including human and mouse. The resolution of genome assembly errors in a new assembly version results in the addition of new CCDS IDs due to improved gene representations in the NCBI and Ensembl annotations. For example, CCDS75980.1 (human gene *DGKK*) was gained following consensus annotation on the GRCh38 (20) assembly after a single nucleotide deletion which existed in the GRCh37 assembly was fixed. In total, about 150 human CCDS IDs were gained based on the GRCh38 assembly following the resolution of genome assembly problems that existed in the GRCh37 assembly.

Data types used in manual curation

New, improved or more highly robust data types that support gene, transcript or protein existence continue to emerge through biological research. Expert curators in the CCDS

Table 2. Data types used in CCDS manual curation decisions

Data type	Curation decisions
RNA-seq (24)	Determination of transcript or gene structure or extent, inferred exon combination, splice variant existence
CAGE tags (25)	Determination of transcription start sites, 5′ UTR extension
H3K4me3 methylation	Determination of general 5′ completeness of transcripts or genes
CpG islands	Determination of general 5′ completeness of transcripts or genes (in conjunction with other data)
Long read transcriptome data	Splice variants; especially useful for genes with poor INSDC transcript support
Proteomics	Determination of gene biotype, novel exons, novel protein termini.
Ribosome profiling	Determination of translation start codons or the coding status of genes with questionable biotypes
Conservation in other species	Determination of gene biotype, annotation of proteins with little or no data about gene function, determination of translation start codon
Conserved protein domains	Determination of gene biotype, annotation of proteins with little or no data about gene function
PhyloCSF	Determination of gene biotype, annotation of uncharacterized proteins
polyA-seq (26)	Determination of 3′ completeness

collaboration adapt to such data types and incorporate them in curation/annotation workflows to provide accurate coding region annotations. Table 2 lists the more recent data types that have been adopted by CCDS curators to make key decisions. Most of the listed data types are used in manual review and are not yet incorporated in the automated annotation pipelines (except short-read and long-read transcriptome data). Typically, these data types are considered for genes that lack traditional support such as mRNAs and ESTs which are archived by INSDC databases, or they may lack information about gene function. Hence, such genes are reviewed on a case-by-case basis with additional support gleaned from newer data types where appropriate.

CONTRIBUTIONS FROM RESEARCH COMMUNITIES

Almost all the annotation included in the CCDS dataset is evidence-based and is supported by experimental data, including transcript, protein and other data types (Table 2) submitted by the research community to public databases such as INSDC. The curation examples cited in this paper underline the importance of sequence data submitted by research groups to public archives, as well as data published in peer-reviewed journals, for improving gene annotation. A small percentage of protein-coding genes remain excluded from the CCDS dataset owing to the lack of sequence data and lack of any information that would indicate the function of the gene. New data may lead to the consensus annotation of these genes in the NCBI and Ensembl annotation sets, and thus, their inclusion in the CCDS dataset. Therefore, it is important that research groups submit sequence data generated by them to public archives. Further, studies in hitherto uncharacterized genes will serve as a crucial resource and help CCDS curators improve their annotation. The CCDS collaboration also welcomes direct input from the research community for the annotation of specific genes and gene families. Such input can be communicated via the user contact email link, as mentioned above.

FUTURE DIRECTIONS

Although dataset analyses indicate the addition of new annotation and stability in the existing consensus annotation of protein-coding regions in the Ensembl and NCBI annotations (Figure 1), a small percentage of genes still lack consistent annotation. These may be cases where one group has annotated a coding gene and the other has not; the latter may have annotated a gene with a different gene type—i.e. non-coding or pseudogene—or may not have described any gene at all. Such cases are naturally a top priority for review by the CCDS curators. In particular, the interpretation of genes predicted to encode small proteins or small open reading frames (smORFs) and non-coding RNAs pose challenges for gene annotation (21,22), and thereby pose challenges for the completion of the CCDS dataset. Interpretation of new datasets including ribosome profiling, cross-species conservation, evolutionary conservation of synonymous codons in ORFs and mass spectrometry promises to help resolve some of the uncertainty about these genes. Nonetheless, gene-level biotype differences may persist between the Ensembl and NCBI annotation sets even after some of these genes have been debated. Such cases ultimately reflect differences between the annotation guidelines of these projects, specifically on how to judge the balance of probability when the evidence for annotation is limited or ambiguous.

While the CCDS project assists navigation between the Ensembl and NCBI databases, users may be confused by inconsistent terminology used by different annotation and nomenclature groups to describe gene and transcript biotypes. To provide consistent terms and further enhance navigation across different genome annotation resources, all members of the CCDS collaborative group have, or plan to, implement Sequence Ontology (23) terms to label genome feature biotypes.

Data pertaining to the 'Review status' (Figure 5) indicate that there are still a significant number of mouse CCDS IDs and a smaller number of human CCDS IDs that lack a 'Reviewed' status. It is our aim to review all provisional CCDS IDs with a goal of eventually providing completely reviewed human and mouse CCDS datasets.

Although the primary focus of the CCDS collaboration is the representation of protein-coding regions, it is also a platform for members of major bioinformatics resources to discuss new ideas, share strategies about using emerging data types for genome annotation and predict user needs based on the latest research trends. For example, user interest has spurred recent discussions in topics such as reconciling the UTRs of transcripts across independent annotation sets and assigning one representative transcript per gene. Further discussions and analyses are needed to explore providing potential deliverables based on these ideas. The CCDS collaboration continues in its pursuit to provide agreement in the annotation of human and mouse protein-coding genes in reference gene sets while constantly adapting to emerging data types and user needs. With the help of

new data in public archives and input from research groups, we will continue working toward the long-term goal of providing consistent and stable annotation of protein-coding genes on the human and mouse reference genome assemblies.

ACKNOWLEDGEMENTS

The authors wish to thank all programmers and other staff at NCBI, EMBL-EBI, UCSC, HGNC and MGI, who contributed to CCDS analyses and maintenance of the CCDS database and its internal curation interface. The authors also thank UniProtKB curators, scientific experts and CCDS users for their helpful inputs.

FUNDING

Work performed at NCBI is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. Work related to Ensembl annotation is supported by the Wellcome Trust [WT098051, WT108749/Z/15/Z], the National Human Genome Research Institute (NHGRI) [U41HG007234, 2U41HG007234] and the European Molecular Biology Laboratory. Work performed at HGNC is supported by an NHGRI grant [U41HG003345] and a Wellcome Trust grant [099129/Z/12/Z]. Work performed at the MGI group is supported in part by an NHGRI grant [U41HG000330]. Work performed at UCSC is supported by the NHGRI grant for the GENCODE project [U41HG007234].
Conflict of interest statement. None declared.

REFERENCES

1. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
2. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., Garcia Giron,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, 1–19.
3. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
4. Harte,R.A., Farrell,C.M., Loveland,J.E., Suner,M.M., Wilming,L., Aken,B., Barrell,D., Frankish,A., Wallin,C., Searle,S. *et al.* (2012) Tracking and coordinating an international curation effort for the CCDS Project. *Database*, **2012**, bas008.
5. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
6. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
7. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
8. Eppig,J.T., Smith,C.L., Blake,J.A., Ringwald,M., Kadin,J.A., Richardson,J.E. and Bult,C.J. (2017) Mouse Genome Informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Methods Mol. Biol.*, **1488**, 47–73.
9. Belkadi,A., Bolze,A., Itan,Y., Cobat,A., Vincent,Q.B., Antipenko,A., Shang,L., Boisson,B., Casanova,J.L. and Abel,L. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 5473–5478.
10. Ezkurdia,I., Rodriguez,J.M., Carrillo-de Santa Pau,E., Vazquez,J., Valencia,A. and Tress,M.L. (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, **14**, 1880–1887.
11. Manase,D., D'Alessandro,L.C., Manickaraj,A.K., Al Turki,S., Hurles,M.E. and Mital,S. (2014) High throughput exome coverage of clinically relevant cardiac genes. *BMC Med. Genomics*, **7**, 67.
12. Chen,R., Im,H. and Snyder,M. (2015) Whole-exome enrichment with the Agilent SureSelect human all exon platform. *Cold Spring Harb. Protoc.*, **2015**, 626–633.
13. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and International Nucleotide Sequence Database, C. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
14. Pundir,S., Martin,M.J. and O'Donovan,C. (2017) UniProt Protein Knowledgebase. *Methods Mol. Biol.*, **1558**, 41–55.
15. Harrow,J.L., Steward,C.A., Frankish,A., Gilbert,J.G., Gonzalez,J.M., Loveland,J.E., Mudge,J., Sheppard,D., Thomas,M., Trevanion,S. *et al.* (2014) The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Res.*, **42**, D771–D779.
16. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
17. Santos,A.J., Nogueira,C., Ortega-Bellido,M. and Malhotra,V. (2016) TANGO1 and Mia2/cTAGE5 (TALI) cooperate to export bulky pre-chylomicrons/VLDLs from the endoplasmic reticulum. *J. Cell Biol.*, **213**, 343–354.
18. Pitman,J.L., Bonnet,D.J., Curtiss,L.K. and Gekakis,N. (2011) Reduced cholesterol and triglycerides in mice with a mutation in Mia2, a liver protein that localizes to ER exit sites. *J. Lipid Res.*, **52**, 1775–1786.
19. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
20. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
21. Pauli,A., Valen,E. and Schier,A.F. (2015) Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *Bioessays*, **37**, 103–112.
22. Makarewich,C.A. and Olson,E.N. (2017) Mining for Micropeptides. *Trends Cell Biol.*, **27**, 685–696.
23. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
24. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
25. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15776–15781.
26. Derti,A., Garrett-Engele,P., Macisaac,K.D., Stevens,R.C., Sriram,S., Chen,R., Rohl,C.A., Johnson,J.M. and Babak,T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183.