

MIT Open Access Articles

The Consensus Coding Sequence (Ccds) Project: Identifying a Common Protein-Coding Gene Set for the Human and Mouse Genomes

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Pruitt, K. D. et al. "The Consensus Coding Sequence (CCDS) Project: Identifying a Common Protein-coding Gene Set for the Human and Mouse Genomes." *Genome Research* 19.7 (2009): 1316–1323. Copyright © 2009 by Cold Spring Harbor Laboratory Press

Published Version: <http://dx.doi.org/10.1101/gr.080531.108>

Publisher: Cold Spring Harbor Laboratory Press

Permanent Link: <http://hdl.handle.net/1721.1/72151>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: <http://creativecommons.org/licenses/by-nc/3.0/>





The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes

Kim D. Pruitt, Jennifer Harrow, Rachel A. Harte, et al.

Genome Res. 2009 19: 1316-1323 originally published online June 4, 2009
Access the most recent version at doi:[10.1101/gr.080531.108](https://doi.org/10.1101/gr.080531.108)

Supplemental Material <http://genome.cshlp.org/content/suppl/2009/06/09/gr.080531.108.DC1.html>

References This article cites 26 articles, 17 of which can be accessed free at:
<http://genome.cshlp.org/content/19/7/1316.full.html#ref-list-1>

Article cited in:
<http://genome.cshlp.org/content/19/7/1316.full.html#related-urls>

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at <http://creativecommons.org/licenses/by-nc/3.0/>.

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

Correction A correction has been published for this article. The contents of the [correction](#) have been appended to the original article in this reprint. The correction is also available online at:
<http://genome.cshlp.org/content/19/8/1506.full.html>

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>

The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes

Kim D. Pruitt,^{1,9} Jennifer Harrow,² Rachel A. Harte,³ Craig Wallin,¹ Mark Diekhans,³ Donna R. Maglott,¹ Steve Searle,² Catherine M. Farrell,¹ Jane E. Loveland,² Barbara J. Ruff,⁴ Elizabeth Hart,² Marie-Marthe Suner,² Melissa J. Landrum,¹ Bronwen Aken,² Sarah Ayling,⁵ Robert Baertsch,³ Julio Fernandez-Banet,² Joshua L. Cherry,¹ Val Curwen,² Michael DiCuccio,¹ Manolis Kellis,^{6,7} Jennifer Lee,¹ Michael F. Lin,^{6,7} Michael Schuster,⁸ Andrew Shkeda,¹ Clara Amid,⁴ Garth Brown,¹ Oksana Dukhanina,¹ Adam Frankish,² Jennifer Hart,¹ Bonnie L. Maidak,¹ Jonathan Mudge,² Michael R. Murphy,¹ Terence Murphy,¹ Jeena Rajan,² Bhanu Rajput,¹ Lillian D. Riddick,¹ Catherine Snow,² Charles Steward,² David Webb,¹ Janet A. Weber,¹ Laurens Wilming,² Wenyu Wu,¹ Ewan Birney,⁸ David Haussler,³ Tim Hubbard,² James Ostell,¹ Richard Durbin,² and David Lipman¹

¹National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA; ²Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; ³Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; ⁴Zebrafish Information Network, University of Oregon, Eugene, Oregon 97403-5291, USA; ⁵The University of Manchester, Faculty of Life Sciences, Manchester Interdisciplinary Biocentre, Manchester M1 7DN, United Kingdom; ⁶Computer Science and Artificial Intelligence Laboratory, Institute of Technology, Cambridge, Massachusetts 02139, USA; ⁷Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA; ⁸European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom

Effective use of the human and mouse genomes requires reliable identification of genes and their products. Although multiple public resources provide annotation, different methods are used that can result in similar but not identical representation of genes, transcripts, and proteins. The collaborative consensus coding sequence (CCDS) project tracks identical protein annotations on the reference mouse and human genomes with a stable identifier (CCDS ID), and ensures that they are consistently represented on the NCBI, Ensembl, and UCSC Genome Browsers. Importantly, the project coordinates on manually reviewing inconsistent protein annotations between sites, as well as annotations for which new evidence suggests a revision is needed, to progressively converge on a complete protein-coding set for the human and mouse reference genomes, while maintaining a high standard of reliability and biological accuracy. To date, the project has identified 20,159 human and 17,707 mouse consensus coding regions from 17,052 human and 16,893 mouse genes. Three evaluation methods indicate that the entries in the CCDS set are highly likely to represent real proteins, more so than annotations from contributing groups not included in CCDS. The CCDS database thus centralizes the function of identifying well-supported, identically-annotated, protein-coding regions.

[Supplemental material is available online at www.genome.org. Data sets and documentation are available in the CCDS database at <http://www.ncbi.nlm.nih.gov/CCDS>.]

One key goal of genome projects is to identify and accurately annotate all protein-coding genes. The resulting annotations add functional context to the sequence data and make it easier to traverse to other rich sources of gene and protein information. Accurately annotating known genes, identifying novel genes, and tracking annotations over time are complex processes that are best

achieved through a combination of large-scale computational analyses and expert curation. These methods must (1) process repetitive sequences in multiple categories including retrotransposons, segmental duplications, and paralogs; (2) process variation including copy number variation (CNV) (Feuk et al. 2006) and microsatellites; (3) distinguish functional genes and alleles from pseudogenes; (4) define alternate splice products; and (5) avoid erroneous interpretation based on experimental error.

Genome annotation information is available from many sources including publications on the sequencing and annotation

⁹Corresponding author.

E-mail Pruitt@ncbi.nlm.nih.gov; **fax** (301) 480-2918.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.080531.108>.

of genes for whole genomes, individual chromosomes, and whole-genome annotation computed by multiple bioinformatics groups. Ensembl and the National Center for Biotechnology Information (NCBI) independently developed computational processes to annotate vertebrate genomes (Kitts 2002; Potter et al. 2004). Both pipelines predict genes, transcripts, and proteins based on interpretations of gene prediction programs, transcript alignments, and protein alignments. In addition, manual annotation is provided by the Havana group at the Wellcome Trust Sanger Institute (WTSI) and the Reference Sequence (RefSeq) group at the National Center for Biotechnology Information (NCBI).

The abundance of different data sources has been problematic for the scientific community since annotated models may change over time as more experimental data accumulate, or may differ among annotation groups owing to differences in methodology or data used. Differences in presentation can also compound the problem. Assigning a unique, tracked accession (CCDS ID) to identical coding region annotations removes some of the uncertainty by explicitly noting where consensus protein annotation has been identified, independent of the website being used. "Consensus" is defined as protein-coding regions that agree at the start codon, stop codon, and splice junctions, and for which the prediction meets quality assurance benchmarks. Thus, a distinguishing feature of the collaborative consensus coding sequence (CCDS) project compared to other protein databases is the integrated tracking of protein sequences in the context of the genome sequence. The current CCDS data sets for human and mouse can be accessed from several public resources including the members of the CCDS collaboration, namely: (1) the Ensembl Genome Browser, which is a joint project between the European Bioinformatics Institute (EBI) and WTSI (Birney et al. 2004); (2) the NCBI Map Viewer (Dombrowski and Maglott 2002); (3) the University of California Santa Cruz (UCSC) Genome Browser (Karolchik et al. 2008; Zweig et al. 2008); and (4) the WTSI Vertebrate Genome Annotation (Vega) Genome Browser (Ashurst et al. 2005). The CCDS collaborators provide access to the same reference genomic sequence and CCDS data set.

The CCDS set is built by consensus; each member of the collaboration contributes annotation, quality assessments, and curation. The collaboration pragmatically defines the initial focus on coding region annotations, rather than the annotated transcripts including untranslated regions (UTRs), because it is critical to identify encoded proteins and because there is more variation in UTR annotation. Protein-coding region annotations that do not

satisfy the criteria for assigning a CCDS ID are evaluated between releases so that the annotation continues to improve. The key goal of the CCDS project is to provide a complete set of high-quality annotations of protein-coding genes on the human and mouse genomes.

Results

We have developed the process flow, quality assurance tests, curation infrastructure, and web resources to support identification, tracking, and reporting of identical protein annotations. Table 1 summarizes the growth of the CCDS database since its first release in 2005.

Following a coordinated annotation update of the reference genome annotation, results are compared to identify identical protein-coding regions. Each coding sequence (CDS) annotation must then pass quality assessment tests before being assigned a CCDS ID and version number (see Methods; Supplemental Fig. 4). The CCDS ID is stable, and every effort is made to ensure that all protein-coding regions with existing CCDS IDs are consistently annotated with each whole-genome annotation update. The protein sequence defined for the CCDS ID is the predicted translation of the coding sequence that is annotated on the genomic reference chromosome. Thus it is identical to the sequence reported by Ensembl, UCSC, and Vega. The sequence may differ at individual amino acids from associated RefSeq records (Pruitt et al. 2009) because the latter are often based on translations of independently generated mRNA sequences that may be selected to represent a different allele.

Quality assessment

We assessed the content and quality of the current public CCDS collection using three metrics: (1) evaluation of NCBI HomoloGene clusters to determine the number of homologous pairs of human and mouse CCDS proteins, (2) comparison of CCDS proteins to the curated UniProtKb/SWISS-PROT protein data set (hereto after referred to as "SWISS-PROT"), and (3) evaluation of genome conservation. HomoloGene (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene>) (Wheeler et al. 2008) reports groups of related proteins annotated on reference chromosomes for select species and includes a consideration of local conserved synteny that is limited to an assessment of flanking genes. HomoloGene is a gene-oriented resource as calculation is based on

Table 1. Growth and current size of the CCDS set

NCBI build ^a	Ensembl release ^a	Existing CCDS	Reinstated ^b	New candidate	QA tests, rejection	QA tests, curation override	Withdrawn, curation	Withdrawn, other ^c	Total no. of CCDS	Total no. of genes
Hs35.1	23	0		16,068	1273	0	0	0	14,795	13,142
Hs36.2	39	14,795		4955	1033	5	243	189	18,290	16,008
Hs36.3	49	18,290	23	2808	692	40	281	29	20,159	17,052
Mm36.1	39	0		15,744	2370	0	0	0	13,374	13,014
Mm37.1	47	13,374		5716	1371	37	6	43	17,707	16,893

^aNCBI build numbers and Ensembl release numbers (e.g., build 35.1 and release 23, etc.) are displayed in the Map Viewer and Ensembl browser, respectively, and represent distinct whole-genome annotation runs. The values reported here reflect the input annotation data set used to calculate new candidates. (Hs) *Homo sapiens*; (Mm) *Mus musculus*.

^bIf unexpected losses, indicated in the "Withdrawn, Other" column, are found in a later build, the CCDS ID is reinstated with a "public" status. Reinstatement requires that the CDS structure be identical to the version that was previously lost, or, if the CDS structure has changed and is found as identical in both input data sets, then the CCDS version number is incremented. For example, see CCDS2672.

^cUnexpected loss of consistent CDS annotation includes changed or removed annotation that is not tracked by the CCDS database as curation-based change. The large accidental loss in human build 36.2 resulted in improved tracking of annotation input data by both the NCBI and Ensembl annotation pipelines. Robust CCDS tracking continues to be a goal of annotation pipelines.

a single longest protein per annotated locus and excludes additional annotated proteins that may be available from consideration (Sayers et al. 2009). There are 16,590 HomoloGene clusters that include proteins from both human and mouse; 15,963 of these have at least one protein with a CCDS ID, and 13,329 clusters include proteins with CCDS IDs from both human and mouse. When the latter subset is evaluated for length of protein product, 68% are within 30 amino acids and 25% are identical. Note that because there are fewer CCDS proteins in mouse, assessing the quality of the annotation by counts of HomoloGene clusters with both human and mouse CCDS members may underrepresent the conservation of annotations. Of the human and mouse protein-coding genes with an associated CCDS ID, 96.5% of the human genes and 95.4% of the mouse genes (16,461 and 16,114, respectively) are clustered by HomoloGene with at least one homolog from any other species (see Fig. 1).

A second approach to assess the CCDS data set is to compare the derived protein sequences to another curated protein data set, namely, the SWISS-PROT records available for human and mouse. Of the 35,505 human and mouse SWISS-PROT records available at the time of this analysis (release 55.5), 81% match a CCDS protein at or above 95% identity, of which 66% are identical. Similar numbers are found for the human and mouse CCDS data sets (see Fig. 2). A complete match between the CCDS data set and that of SWISS-PROT is not expected because these two resources are generated using very different data models; SWISS-PROT is a protein-oriented resource that doesn't require consistency with the reference genome sequence, whereas CCDS proteins do have

that requirement. Lack of a SWISS-PROT match may indicate differential representation of alternate splice products, differences in the gene type designation (protein-coding vs. non-coding), known gaps in CCDS for genes with limited sequence data, and differences that can be correlated to the reference genome sequence including small sequence differences (mismatches, insertions, or deletions), assembly gaps, copy number variation, or alternate haplotypes. Since SWISS-PROT is centered on proteins, it includes records for which a direct correlation to a CCDS cannot be expected because the categories are not in scope for CCDS. Among those are entries for human endogenous retrovirus (HERV) proteins, mature immunoglobulin proteins resulting from genomic rearrangement events, small physiologically active peptides that may result from enzymatic reactions, and putative uncharacterized proteins that may be alternatively interpreted by another group as a pseudogene or non-protein-coding RNA.

The third assessment uses the reading frame conservation (RFC) methodology (Kellis et al. 2003) to compare the protein-coding evolutionary signature of genes in CCDS with those from the RefSeq and Ensembl collections that do not contain CCDS proteins (see Supplemental Table 2). The RFC score is the percentage of nucleotides whose reading frame is evolutionarily conserved across species. Previous work has shown that 98% of the well-studied human genes have RFC scores >90 (Clamp et al. 2007). Figure 3 shows the distribution of the RFC scores of the CCDS, RefSeq, and Ensembl loci for human and mouse, along with a control set of non-protein-coding DNA for human. In the human set, 95.9% of the CCDS genes have an RFC score above 90.

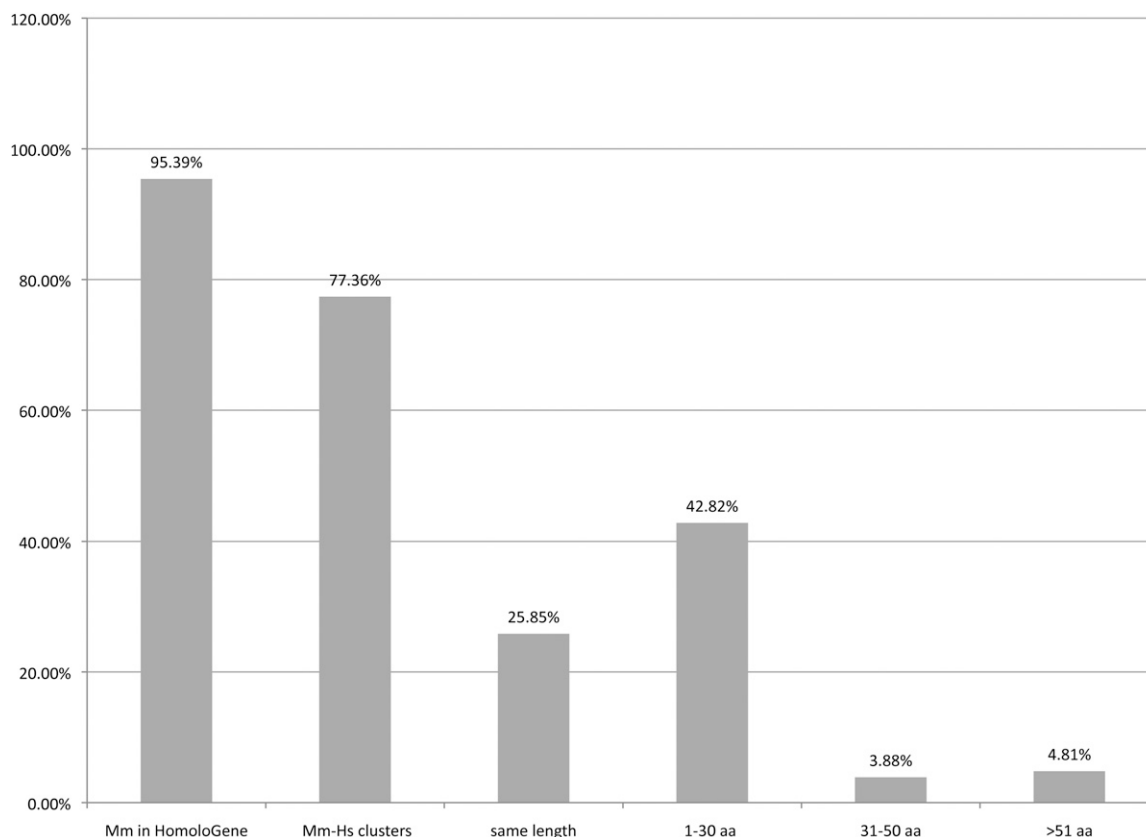


Figure 1. The percentage of mouse CCDS proteins that are found in any HomoloGene cluster versus those in a cluster that also contains a human CCDS protein (first two bars, respectively). For the latter category, results are further categorized based on protein length differences for the human and mouse homologous proteins.

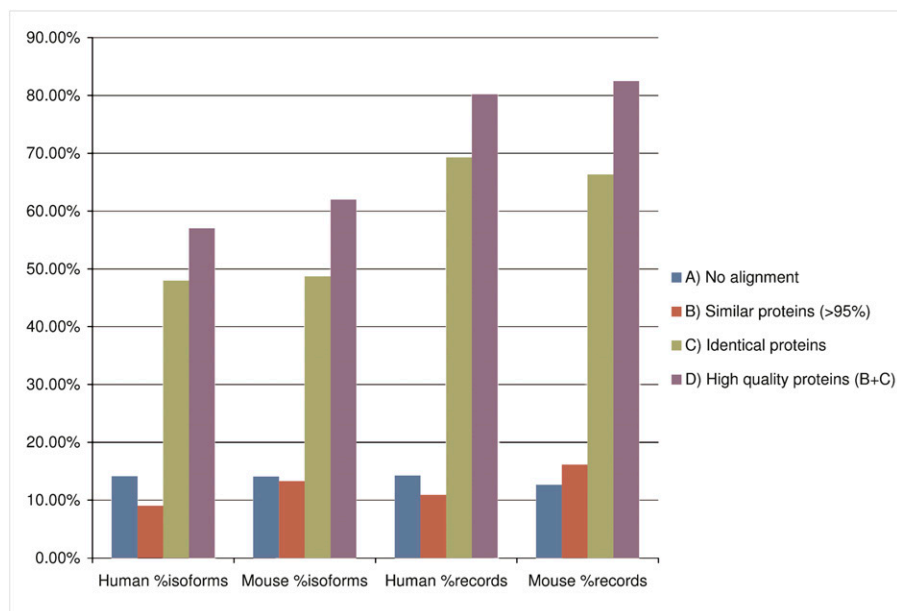


Figure 2. The percentage of human and mouse genes, with associated CCDS IDs for one or more proteins that are identical (C), similar (B), or unique (A) when compared to SWISS-PROT records and to SWISS-PROT isoforms that were extracted from record annotation (see Methods). (D) The total number of high-quality matches.

In contrast, 37.6% and 44.3% of the non-CCDS RefSeq and Ensembl loci, respectively, have scores above RFC 90, and only 1.2% of the control set has RFC scores >90. In mouse, 93.5% of the CCDS genes exceed the RFC 90 threshold, compared to 36.8% of the RefSeq non-CCDS genes and 46.3% of the Ensembl non-CCDS genes scoring above RFC 90. In both human and mouse, the CCDS gene set shows significantly stronger evidence of evolving in a manner consistent with protein-coding genes than the RefSeq and Ensembl loci that do not contain CCDS proteins. The weak-scoring genes tend to be the genes that require careful review by annotators. For instance, 35% of human RefSeq genes with low RFC scores (<90) are single-exon genes, while 6% of those with a high RFC score (≥ 90) are single-exon genes. Genes with low RFC

scores are also enriched for segmental duplications with 31% of low RFC RefSeq genes overlapping regions of segmental duplication (Bailey et al. 2001), compared to only 9% of the high RFC scoring genes being in segmental duplications. Similarly, 41% of the low RFC human RefSeqs are identified as originating by retrotransposition (Baertsch et al. 2008), while only 18% of high RFC ones are categorized as retro-copies.

Access to CCDS data

NCBI hosts a public website for the CCDS project (<http://www.ncbi.nlm.nih.gov/CCDS/>) that includes information about the collaboration, provides links to reports and FTP download, and provides a query interface to retrieve information about CCDS sequences and locations. The interface supports query by multiple identifiers including official gene symbols, CCDS ID, Entrez Gene ID (Maglott et al. 2007), or sequence ID (RefSeq, Ensembl ID, or Vega ID). Query results are presented in a table format with links provided to access the full report details

for each CCDS ID (see Supplemental Fig. 1). Multiple CCDS IDs are reported for a gene if both data sets consistently annotated more than one CDS location. The CCDS ID-specific report page, shown in Figure 4, provides a detailed report for the CCDS ID. Some records also include a Public Note, provided by a curator, summarizing the rationale for an update, withdrawal, or to explain representation choices. Reports may include an update history for associated sequence records when relevant (Supplemental Fig. 2).

All collaborating groups indicate the CCDS ID on gene and/or protein report pages, and Genome Browsers indicate when a CCDS is available for a locus using either display style (coloration), text labels, or by providing as a data track (Supplemental Fig. 3). For those interested in downloading the full CCDS

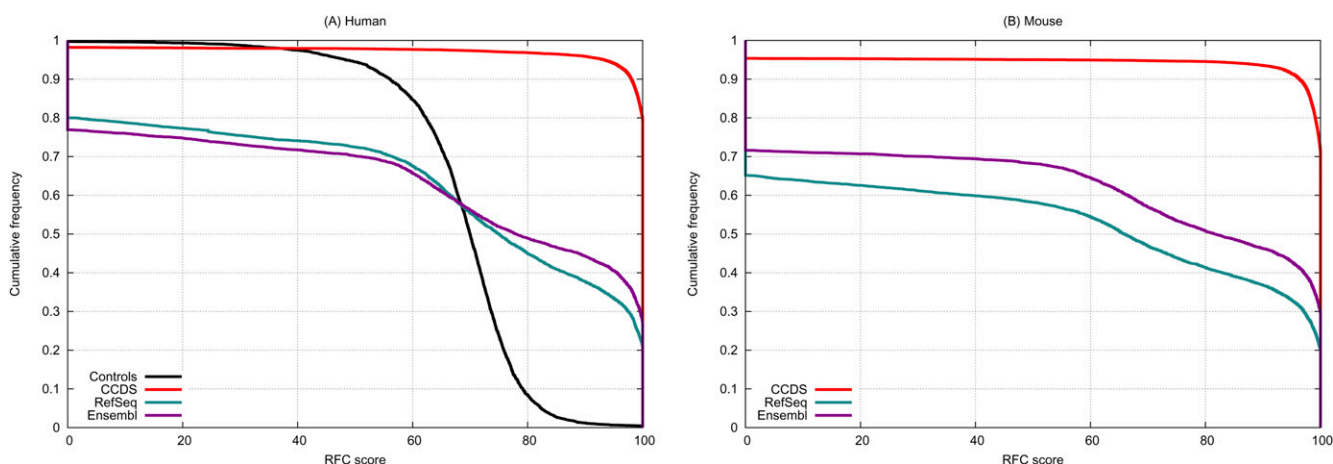


Figure 3. Cumulative distributions of RFC scores for human (A) and mouse (B). These graphs compare the RFC scores for CCDS loci with those of RefSeq and Ensembl loci that do not contain a CCDS protein, as well as a control data set for human. Since the controls were designed to have a similar alignment coverage to well-known genes, loci in other gene sets with less alignment coverage will score less than the controls.

Report for CCDS ID 7564

CCDS	Status	Species	Chrom.	Gene	NCBI Builds	Links
7564.2	Public	<i>Homo sapiens</i>	10	MXI1	35.1 - 36.3	

Public Note

The coding region has been updated to include an alternative exon that is more supported by the available transcript and protein data. The update completes the helix-loop-helix dimerization domain, which was previously truncated.

Sequence IDs included in CCDS 7564.2

Original	Current	Source	Nucleotide ID	Protein ID	Status in CCDS	Seq. Status	Links
✓	✓	EBI,WTSI	ENST00000239007	ENSP00000239007	Accepted	alive	
✓	✓	EBI,WTSI	OTTHUMT0000050316	OTTHUMP0000020469	Accepted	alive	
✓	✓	NCBI	NM_005962.4	NP_005953.4	Accepted	alive	

Chromosomal Locations for CCDS 7564.2

On '+' strand of Chromosome 10 (NC_000010.9)

Genome Browser links:

Chromosome	Start	Stop	Links
10	111975959	111976031	
10	111977937	111978069	
10	111994576	111994605	
10	112028928	112029042	
10	112029662	112029833	
10	112034572	112034735	

CCDS Sequence Data

Blue highlighting indicates alternate exons.

Red highlighting indicates amino acids encoded across a splice junction.

Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (687 nt):

ATGGAGCGGGTGAAGATGATCAACGTGCAGCGTCTGCTGGAGGCTGCCAGTTTTTGGAGCGCCGGGAGC
GAGAGTGTGAACATGGCTACGCCCTCTTCATTCCCGTCCATGCCGAGCCCCGACTGCAGCATTCAAAGCC

Figure 4. Detailed CCDS ID report page for a MXI1 protein. The CCDS report page presents three tables of information followed by nucleotide and protein sequences for the annotated CDS. The first table summarizes the status for the specified CCDS ID. Colored icons provide links to related resources, to a history report (orange H icon), or to re-query the CCDS database with a different type of identifier. (Red G icon re-queries the CCDS database by GeneID to return all CCDS IDs available for a gene.) A Public Note is provided for a subset of curated records to explain the nature of, and/or the reason for, an update or withdrawal. The sequence identifiers table reports sequences tracked as members of the CCDS ID. A checkmark in the first column (Original) identifies sequence identifiers represented on the annotated genome and included in the analyzed input data sets, and a checkmark in the second column (Current) identifies those that are considered current members (see Supplemental Fig. 2) The Chromosome Locations table reports the genomic coordinates of each exon of the CDS with links to view the annotation in different browsers—the violet icons (N) NCBI; (U) UCSC; (E) Ensembl; and (V) Vega. The nucleotide sequence of the CDS, derived from the genome sequence using the reported exon.

data set, the collaboration provides reports of the associated identifiers as well as the sequence data for anonymous FTP (<ftp://ftp.ncbi.nih.gov/pub/CCDS/>). Please refer to provided README files for descriptions of information provided and file formats.

Manual curation of the CCDS data set

Coordinated manual curation, a critical aspect of the CCDS project, is supported by a restricted-access website and a discussion e-mail list. The collaboration has generated standardized curation

guidelines for selection of the initiation codon and interpretation of upstream ORFs and transcripts that are predicted to be candidates for nonsense-mediated decay (NMD) (Lejeune and Maquat 2005). Curation occurs continuously, and any of the collaborating centers can flag a CCDS ID as a potential update or withdrawal. Planned updates that are either under discussion or have achieved consensus agreement are indicated in the public CCDS website by a change in the status of the CCDS ID (Supplemental Table 1). Conflicting opinions are addressed by consulting with scientific experts or other annotation curation groups such as the HUGO Gene Nomenclature Committee (HGNC) (Bruford et al. 2008) and Mouse Genome Informatics (MGI) (Eppig et al. 2005). If a conflict cannot be resolved, then collaborators agree to withdraw the CCDS ID until more information becomes available.

To date, we have reviewed more than 14,000 CCDS proteins and confirmed the existing CDS annotation with no change. Review also resulted in the removal of 530 CCDS IDs and suggested updates to 1014 proteins. If a CCDS protein is updated, the CCDS ID version number is incremented, and often a note is provided explaining the update. For example, review of the protein annotation (CCDS ID 10689) for the human *SRCAP* gene resulted in an N-terminal extension that adds an HSA domain that is found in DNA-binding proteins and is often associated with helicases. The HSA domain is consistent with the other domains found in the protein and with the presumed function of this protein as a component of the SRCAP chromatin remodeling complex (Johnston et al. 1999; Wong et al. 2007). This significant improvement to the protein annotation was immediately available in the RefSeq transcript and protein sequences and in the Vega and UCSC Genome Browsers. This improvement became available in the NCBI and Ensembl Genome Browsers following a recalculation of the annotation for the human genome. Note there is

no corresponding CCDS ID for the mouse *Srcap* locus yet, primarily owing to insufficient transcript data but confounded by the observation of differences in the exon definition compared to the human coding sequence.

Discussion

Although the human and mouse genome sequences have been of “finished” quality for several years, refinements to the annotation of protein-coding genes continue to take place. Until the

inception of the CCDS project, it was difficult to identify which protein annotations were represented consistently by the major browsers. The CCDS project solves this problem and establishes a framework to support well-supported, consistent, comprehensive annotation of the protein-coding content of the human and mouse genomes. The CCDS project has already identified 37,866 consistently annotated human and mouse CDSs for 33,945 genes and assigned them stable, versioned identifiers. By developing annotation standards, coordinating review of automated annotation, and documenting annotation decisions, the CCDS group continues to make a major contribution to the usability of the human and mouse genomic sequences.

The three independent methods used to assess the CCDS collection demonstrate that the genes included are highly likely to be protein-coding loci. Comparison to HomoloGene data indicates that at least 77% of the mouse genes with an associated CCDS ID have a homologous gene in human with an associated CCDS ID. Of these, the majority of the homologous CCDS proteins have a comparable length. Review of identified homologs with larger length differences indicates that the majority of them reflect valid differences due to alternate splicing. Comparison to another highly curated data set, SWISS-PROT, showed that 81% of the SWISS-PROT proteins are identical or highly similar to those encoded by CCDS, with similar results for mouse and human. The RFC analysis results indicate that 95% of the CCDS proteins do have an evolutionary signature that is consistent with their protein-coding designation.

The absence of a CCDS ID for a putative protein-coding gene annotation does not necessarily indicate that annotation is of poor quality; it indicates only that annotation is not yet consistent and requires additional review. Causes of annotation differences include resource-specific automatic annotation methods, timing of manual curation updates, conflicts between genomic and cDNA evidence, and incomplete curation guidelines on evaluating whether or not a locus is protein-coding, how much evidence is required to provide annotation, or where splice junctions should be annotated in repetitive regions. Until we have robust data from proteomic analyses, it is indeed a challenge to identify genes that are protein-coding, whether or not they are in the CCDS set. Supplemental Table 3 summarizes one approach to this problem, namely, classifying officially named genes thought to be protein-coding that have not yet been assigned a CCDS ID. Some have been assigned a RefSeq protein accession, many of which became available since the last CCDS analysis and are expected to gain a CCDS ID in the next analysis; some are associated with genome assembly issues that prevent representing the preferred protein from the genomic sequence; some are associated with protein sequence but are not in the RefSeq NM/NP set (often because the protein data appear to be partial); and some loci, perhaps historical, have no associated protein sequence at all. It is important to note that a CCDS ID represents consistency between annotation resources—it does not indicate that the annotation has been manually reviewed. We welcome feedback from the scientific community either regarding current annotations or to provide data and help with annotating new loci (see the CCDS website for contact information).

The benefits of the CCDS project extend beyond the CCDS data set currently available for the human and mouse genomes. The collaboration supporting the CCDS analysis process has resulted in improvements in automated annotation methods, quality assessment, and manual curation that are applied to many genomes. Discussions about evidence for annotation and pub-

lications between the RefSeq, Havana, and UCSC curation staff have resulted in re-evaluation of genomic sequence including assembly issues, correction of annotation errors, and identification of loci for which additional experimental validation is needed. Questions about the genomic sequence are reported to the Genome Reference Consortium (<http://www.ncbi.nlm.nih.gov/genome/assembly/grc/index.shtml>); annotation errors are resolved collaboratively ensuring consistent representation at all sites, and loci in need of experimental validation are reported to the GENCODE project. Experimental validation of transcripts and splice sites will occur as part of the GENCODE scale-up project (<http://www.sanger.ac.uk/encode/>), which builds on the successful GENCODE pilot project (Harrow et al. 2006). GENCODE is part of the extended human Encyclopedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium 2007). Annotated transcripts highlighted for validation will be confirmed in an array of tissues using RT-PCR or RNAseq. The resulting sequence will be fed back into the CCDS project as supporting evidence.

The CCDS group is thus a key participant in improving the representation of the human and mouse genomes. For example, we are collaborating with the HGNC to match loci they have named to the human genomic sequence. Curation is also focused on human–mouse homologous proteins for which one lacks a CCDS ID, and protein-coding loci with associated SWISS-PROT proteins for which there is no corresponding CCDS ID. An additional long-term goal is to add attributes that indicate where transcript annotation is also identical (including the UTRs) and to indicate splice variants with different UTRs that have the same CCDS ID. It is also anticipated that as more complete and high-quality genome sequence data become available for other organisms, annotations from these organisms may be in scope for CCDS representation.

Methods

Identifying the candidate CCDS groups and tracking updates

Following the release of a re-annotation of the human or mouse reference genomes by both NCBI and Ensembl, we compared the genome annotation data sets provided by NCBI and Ensembl (Supplemental Fig. 4) to identify protein-coding annotations that are identical (start codon, stop codon, splice junctions) and do not include in-frame stop codons or apparent frameshifts. The full-length translation of proteins that include the amino acid selenocysteine (identified as the codon UGA) is provided when collaborators are consistent in annotating an internal stop codon. Identical annotations are subject to quality assessment tests and assessed to identify whether they correspond to existing CCDS IDs or are novel. Existing CCDS IDs are tracked using the combination of sequence identifiers and chromosomal coordinates. The version number is incremented if the annotated exon coordinates and predicted protein product have changes, in which case it is required that they have been identically updated in both annotation sources owing to coordinated curation (see below and Supplemental Table 1). Novel entries are assigned a unique CCDS ID with an initial version of 1. Although mechanisms are in place in the Ensembl and NCBI annotation pipelines to ensure that existing CCDS entries are stably incorporated in the whole genome re-annotation process, it is possible that final annotation of a CCDS protein may not be included or may no longer match, and so the CCDS entry is determined to be “lost” by the comparative process described above, in which case, they are flagged for manual follow-up.

The primary data represented by a CCDS ID are the chromosome coordinates of the annotated protein-coding exons and the nucleotide and conceptually translated protein sequence obtained from those coordinates. Ancillary data associated with the CCDS ID include sequence IDs and gene IDs included in the NCBI and Ensembl data sets. Each CCDS ID includes at least one protein identifier from both NCBI and Ensembl; a CCDS ID can include additional protein identifiers from either data set when they are predicted from transcripts that differ only because of alternate splicing in the untranslated region (i.e., when the protein-coding annotation is identical).

Quality assessment for a CCDS build

We have implemented a series of tests that assess the protein sequence, conservation, and likelihood that annotation erroneously represents a pseudogene as protein-coding. The Ensembl and NCBI protein length and sequence are compared by alignment to identify and discard proteins that are discordant owing to annotation or processing error or insertion/deletion differences between NCBI RefSeq proteins and the protein annotated on the reference genome. Two types of protein comparisons are done: (1) protein sequences provided by each annotation source as FASTA files are compared to the conceptually translated CDS sequence that is extracted de novo from the genome annotation coordinates, and (2) proteins provided as FASTA by each annotation data source are compared to each other. Putative retrotransposed pseudogenes are identified from mRNAs [with poly(A) tails removed] that align to the genome using BLASTZ (Schwartz et al. 2003) at more than one location. Alignments are scored for a series of features to identify putative retrotransposed pseudogenes as previously described (Baertsch et al. 2008). Protein models are also evaluated for genome conservation patterns that may indicate that the gene is not functional in human. Analysis of BLASTZ cross-species alignments to the human genome gene annotation detects potential problems including: nonconserved start and stop codons, nonconserved splice sites, uncompensated insertion- or deletion-associated frameshifts, and in-frame stop codons. Cross-species alignments included chimpanzee, mouse, rat, dog, chicken, and rhesus; only syntenic alignments are used from the assembled genomes. Additional QA tests are applied by NCBI to the RefSeq sequences included in the CCDS database as previously described (Pruitt et al. 2007).

Exchange of curation data

NCBI maintains a relational database that tracks CCDS candidates; locations and identifiers; results of quality assurance tests, curator comments; and CCDS IDs and versions. A database extraction is distributed via a private ftp site to the members of the CCDS collaboration on a daily basis. A restricted-access website supports the collaboration, and a public access website and ftp site disseminate data for CCDS IDs.

Quality assessment of the current human and mouse CCDS data set

HomoloGene

The number of mouse and human CCDSs with homologs in other species was calculated by determining whether the current CCDS genes for human and mouse are members of a HomoloGene group (release 62) that has a gene from at least one other species as a member. Additional filtering identified HomoloGene clusters containing human and mouse genes where the corresponding human and mouse loci are both associated with a CCDS ID.

Comparison to SWISS-PROT

Manually curated SWISS-PROT records (Apweiler et al. 2008) (release 55.5) were obtained via the EBI ftp site (<ftp://ftp.ebi.ac.uk/pub/databases/uniprot/>). A SWISS-PROT accession number may represent more than one sequence isoform. For example, Q8NCE2 includes annotation indicating that amino acids 479 through 538 are missing in isoform 3. Therefore, the SWISS-PROT data set was processed with VARSPLIC (Kersey et al. 2000) to derive an expanded data set (isoforms) by extracting alternate splice products based on record annotation. Exonerate (Slater and Birney 2005) was used with the affine:local model and a sequence identity threshold of 95% to align UniProt sequences to CCDS proteins. Alignments were analyzed and binned into several different categories, with interpretations based on an evaluation of alignments to the expanded set of alternate protein variants versus alignments to the UniProt record where the record is defined as a CCDS protein alignment with the highest coverage score to any of the set of splice variants extracted from the UniProt record. Results were binned as follows: (1) alignment found or not, (2) overall sequence identity, (3) N terminus identical or not, (4) C terminus identical or not, and (5) the alignment coverage of the UniProt protein.

RFC conservation analysis

Conservation analysis was performed using genomic annotations of transcripts for human assembly NCBI 36 (UCSC hg18) and mouse assembly NCBI 37 (UCSC mm9) for CCDS along with the corresponding RefSeq and Ensembl protein-coding gene sets (Supplemental Table 2). Human transcripts were scored against mouse and dog, with mouse transcripts scored using human and dog. The score for a gene is the maximum RFC for any of its transcripts against either of the aligned genomes. A control data set of randomized human sequences was also scored, as previously described (Clamp et al. 2007). Controls are non-protein-coding regions of the genome that serve as a null model, with similar structure, GC content, alignment coverage, and mutation rates as for well-known protein-coding genes.

Acknowledgments

We thank the programmer, database, and curation staff at Ensembl, NCBI, WTSI, and UCSC for their contribution to the CCDS analysis, maintenance, and continuing curation efforts. We thank the UniProt Consortium, the HGNC, and MGI for many useful discussions that improve protein representation in all data sets. UCSC thanks the UCSC Genome Browser team for their tools, data, and assistance, and Michele Clamp (Broad Institute) for providing controls for conservation analysis. NCBI thanks Zev Hochberg for his contributions toward the initial CCDS database schema and CCDS build analysis. UCSC was funded for this work from subcontract no. 0244-03 from NHGRI grant no. 1U54HG004555-01 to the Wellcome Trust Sanger Institute. Work at the Wellcome Trust Sanger Institute was supported by the Wellcome Trust (grant nos. WT062023, WT077198) and by NHGRI grant no. 1U54HG004555-01. Work at NCBI was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

- Apweiler R, Bougueleret L, Altaïrac S, Amendolia V, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, et al. 2008. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**: D169–D174.
- Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33**: D459–D465.

- Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9**: 466. doi: 10.1186/1471-2164-9-466.
- Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res* **11**: 1005–1017.
- Birney E, Andrews T, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, et al. 2004. An overview of Ensembl. *Genome Res* **14**: 925–928.
- Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. 2008. The HGNC Database in 2008: A resource for the human genome. *Nucleic Acids Res* **36**: D445–D448.
- Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104**: 19428–19433.
- Dombrowski SM, Maglott D. 2002. Using the Map Viewer to explore genomes. In *The NCBI handbook*. National Library of Medicine, Bethesda, MD. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch20>.
- Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM, et al. 2005. The Mouse Genome Database (MGD): From genes to mice—a community resource for mouse biology. *Nucleic Acids Res* **33**: D471–D475.
- The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.
- Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7**: S4. doi: 10.1186/gb-2006-7-s1-s4.
- Johnston H, Kneer J, Chackalaparampil I, Yaciuk P, Chrivia J. 1999. Identification of a novel SNF2/SWI2 protein family member, SRCAP, which interacts with CREB-binding protein. *J Biol Chem* **274**: 16370–16376.
- Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**: D773–D779.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.
- Kersey P, Hermjakob H, Apweiler R. 2000. VARSPLIC: Alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics* **16**: 1048–1049.
- Kitts P. 2002. Genome assembly and annotation process. In *The NCBI handbook*. National Library of Medicine, Bethesda, MD. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch14>.
- Lejeune F, Maquat LE. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* **17**: 309–315.
- Maglott D, Ostell J, Pruitt KD, Tatusova T. 2007. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res* **35**: D26–D31.
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. 2004. The Ensembl analysis pipeline. *Genome Res* **14**: 934–941.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61–D65.
- Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37**: D32–D36.
- Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvermin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**: D5–D15.
- Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human mouse alignments with BLASTZ. *Genome Res* **13**: 103–107.
- Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31. doi: 10.1186/1471-2105-6-31.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvermin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36**: D13–D21.
- Wong MM, Cox LK, Chrivia JC. 2007. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. *J Biol Chem* **282**: 26132–26139.
- Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. 2008. UCSC Genome Browser tutorial. *Genomics* **92**: 75–84.

Received December 5, 2008; accepted in revised form April 20, 2009.

Erratum

Genome Research 19: 1316–1323 (2009)

The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes

Kim D. Pruitt, Jennifer Harrow, Rachel A. Harte, Craig Wallin, Mark Diekhans, Donna R. Maglott, Steve Searle, Catherine M. Farrell, Jane E. Loveland, Barbara J. Ruff, Elizabeth Hart, Mari-Marthe Suner, Melissa J. Landrum, Bronwen Aken, Sarah Ayling, Robert Baertsch, Julio Fernandez-Banet, Joshua L. Cherry, Val Curwen, Michael DiCuccio, Manolis Kellis, Jennifer Lee, Michael F. Lin, Michael Schuster, Andrew Shkeda, Clara Amid, Garth Brown, Oksana Dukhanina, Adam Frankish, Jennifer Hart, Bonnie L. Maidak, Jonathan Mudge, Michael R. Murphy, Terence Murphy, Jeena Rajan, Bhanu Rajput, Lillian D. Riddick, Catherine Snow, Charles Steward, David Webb, Janet A. Weber, Laurens Wilming, Wenyu Wu, Ewan Birney, David Haussler, Tim Hubbard, James Ostell, Richard Durbin, and David Lipman

The affiliation for Clara Amid was listed incorrectly. The correct affiliation is:

Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom

The authors apologize for any confusion this may have caused.