Curation at the NCBI: Genomes, Genes, & Sequence Standards

Garth Brown, Catherine Farrell, Jennifer Hart, Melissa Landrum, Donna Maglott, Bonnie Maidak, Michael Murphy, Terence Murphy, Kim Pruitt, Bhanu Rajput, Lillian Riddick, David Webb, Janet Weber, Wendy Wu

National Center for Biotechnology Information, National Institutes of Health

Abstract:

The National Center for Biotechnology Information (NCBI) provides curation support for many genomes, and disseminates information in several resources including Entrez Gene, reference sequences (RefSeq), the Consensus CDS (CCDS) database, and the Genome Reference Consortium (GRC). These projects are supported by several collaborations to provide:1) support to the international consortium maintaining the assemblies for human and mouse (GRC); 2) sequence standards for chromosomes, genes, transcripts and proteins (RefSeq); 3) reports of integrated information including nomenclature, publications, phenotypes and diseases, sequences, ontologies, interactions (Gene); and 4) identification of proteins that are consistently annotated on the human and mouse reference genomes, and consistently updated by collaborating members (CCDS).

NCBI curation of any one data type (e.g., a gene) is closely integrated with evaluation of the genome assembly, and determining annotation by way of RefSeq transcript and protein sequences. Database and work-flow infrastructure is designed to support reporting and tracking issues with the assembly, gene, or evidence data to collaborating groups, and to support collaborative review and discussions of issues that arise. Curation depends on publicly available information to represent the gene extent, alternatively spliced transcripts, and protein isoforms. Scientific consults occur regularly and wet-bench validation needs are supported by some of the collaborations. Curation of genome annotation results in improved data presentation at the three major genome browser sites (Ensembl, NCBI, UCSC) and has resulted in efforts to define common curation guidelines to maximize consistency and minimize conflicts.

The presentation focuses on curation of the human genome, genes, and RefSeq sequence standards.

RefSeq curation includes review of the genome sequence, gene data, and definition of transcript and protein products as RefSeq sequence records. Curation is highly collaborative and areas of concern are shared with multiple groups including the Genome Reference Consortium, the HUGO Gene Nomenclature Committee, and Consensus CDS collaborators. These panels illustrate some areas of communication.

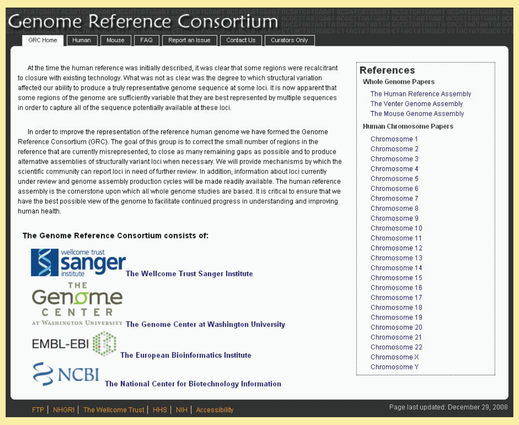
Maintaining the human genome assembly

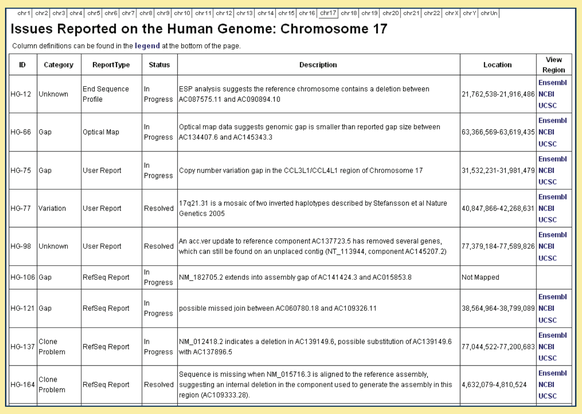
NCBI is a member of the Genome Reference Consortium (GRC). Consortuim goals include providing periodic assembly updates to close gaps, provide an improved representation for some regtions. The GRC recently submitted an update for the human genome (CM000663- CM000686).

RefSeq curation contributes to the GRC goals by reporting potential assembly issues. RefSeq transcript alignments to the human genome were analyzed to identify all records that demonstrate an unaligned region or mismatch vs. the genome. Curation staff reviewed results and either revised the RefSeq record, flagged the difference as a valid polymorphism intentionally retained in the RefSeq record, or reported the sequence conflict to the GRC for further evaluation. This resulted in 322 reports to the GRC including: 1) known, annotated assembly gaps that impact annotation of tracked genes; 2) observed insertions or deletions relative to human cDNAs; and 3) haplotype differences that impact genome annotation results.

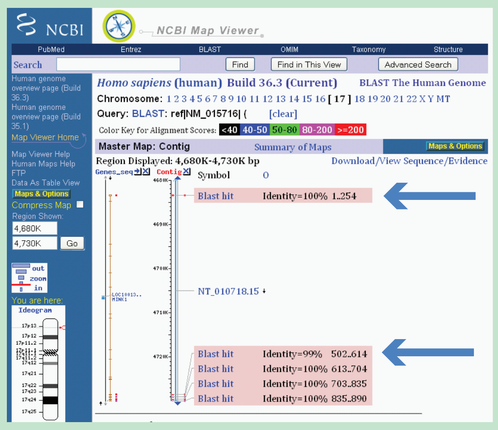
GRC site: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

The GRC web site reports regions under review.



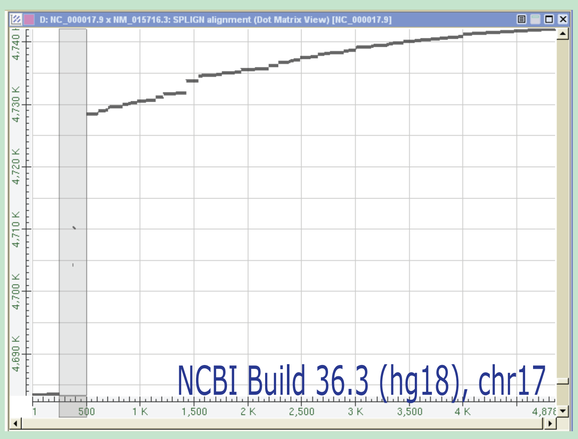


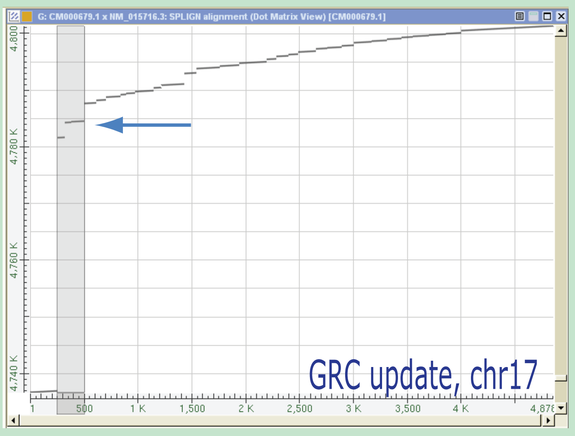
ID	Category	ReportType	Status	Description	Location	View Region
HG-12	Unknown	End Sequence Profile	In Progress	ESP analysis suggests the reference chromosome contains a deletion between AC087575.11 and AC090894.10	21,762,539-21,916,486	Ensembl NCBI UCSC
HG-66	Gap	Optical Map	In Progress	Optical map data suggests genomic gap is smaller than reported gap size between AC134407.6 and AC145343.3	63,366,569-63,619,435	Ensembl NCBI UCSC
HG-75	Gap	User Report	In Progress	Copy number variation gap in the CCL3L1/CCL4L1 region of Chromosome 17	31,532,231-31,981,479	Ensembl NCBI UCSC
HG-77	Variation	User Report	Resolved	17q21.31 is a mosaic of two inverted haplotypes described by Stefansson et al Nature Genetics 2005	40,847,866-42,268,631	Ensembl NCBI UCSC
HG-98	Unknown	User Report	Resolved	An acc.ver update to reference component AC137723.5 has removed several genes, which can still be found on an unplaced contig (NT_113944, component AC145207.2)	77,379,184-77,589,826	Ensembl NCBI UCSC
HG-106	Gap	RefSeq Report	In Progress	NM_182705.2 extends into assembly gap of AC141424.3 and AC015853.8	Not Mapped	
HG-121	Gap	RefSeq Report	In Progress	possible missed join between AC060780.18 and AC109326.11	38,564,964-38,799,089	Ensembl NCBI UCSC
HG-137	Clone Problem	RefSeq Report	In Progress	NM_012418.2 indicates a deletion in AC139149.6, possible substitution of AC139149.6 with AC137896.5	77,044,522-77,200,683	Ensembl NCBI UCSC
HG-164	Clone Problem	RefSeq Report	Resolved	Sequence is missing when NM_015716.3 is aligned to the reference assembly, suggesting an internal deletion in the component used to generate the assembly in this region (AC109333.28).	4,632,079-4,810,524	Ensembl NCBI UCSC



(left) Map Viewer display showing BLAST results on the genome that suggest a problem with the genome assembly in the location of the MINK1 gene (GeneID:50488). Note the gap in the alignment for the RefSeq query sequence (NM_015716.3) on reference chromosome 17 (NC_000017.9). The full extent of the RefSeq record is supported by other abundant data. This issue was reported to the GRC.

(below) Genome Workbench dot matrix displays illustrating the alignment gap found between NM_015716.3 and chromosome 17 (left panel). The right panel shows an improved alignment is found for the updated version of chromosome 17 (CM0000679.1). This update is being processed in NCBI's genome annotation pipeline.



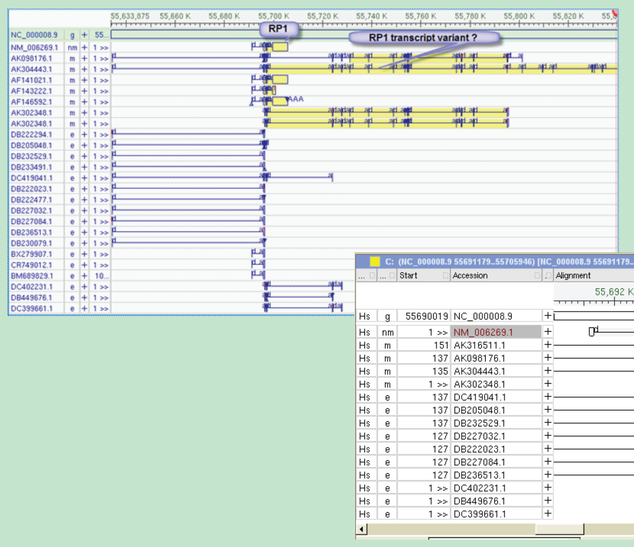


Genome Workbench is available for download: http://www.ncbi.nlm.nih.gov/projects/gbench/

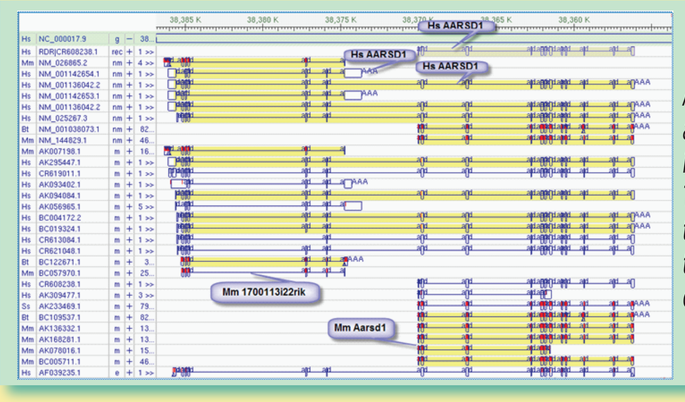
Genes

Extensive communication and coordination occurs between NCBI's RefSeq curation staff, the HUGO Gene Nomenclature Committee (HGNC), the Sanger Institute's Havana curation group, and other groups to maintain human gene information. One focus for RefSeq curation is correct definition of the gene location, extent, exon structure, and gene type (e.g., pseudogene) for the human genome. Defining the extent and products for a gene is not trivial for regions with complex transcription profiles. RefSeq uses the following guidelines:

1. *Treat transcripts having any identical exons as a single gene (default).*
   (may include different promoters (different first exons) and distinct proteins)
2. *Exception: treat complex transcribed locations as >2 genes if:*
   a. *discrete loci have been described historically & definitions continue to support scientific communication.*
   b. *abundant data define two non-overlapping complete transcript units and long connecting transcripts are very rare in comparison.*



A Genome Workbench display, customized for RefSeq, showing transcripts aligned to the genome in the location of the RP1 gene (GeneID 6101). The historical definition of the functional protein and transcript unit is represented by NM_006269. Other data suggest that the locus is complex.



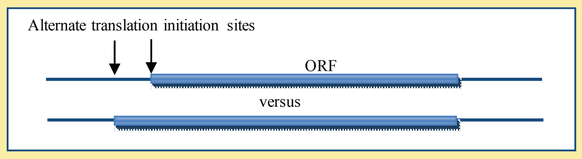
A Genome Workbench display showing a mix of human and non-human transcripts aligned to the human genome in the location of the AARSD1 gene (GeneID:80755). This region is treated as a single locus in human. In contrast, the homologous region of the mouse genome is tracked as two distinct genes as there is no evidence for the long connecting transcript data.

Transcripts, Proteins, & CCDS

NCBI is a member of the Consensus CDS (CCDS) collaboration which identifies high-quality identically annotated protein coding regions annotated on the human genome by both NCBI's and Ensembl's genome annotation pipelines. Modifications to CCDS proteins are done by collaborative agreement in order to maintain consistency. Proteins that are not included in the CCDS database are also reviewed to reach agreement on the correct protein definition to represent on the annotated genome.

The development of common curation guidelines used by NCBI's RefSeq curators, the Sanger Instutute's Havana curators, and the UCSC CCDS curation staff has improved efficiency and reduced the need for lengthy discussions to resolve conflicting annotation. These curation guidelines are now available on the CCDS web site.

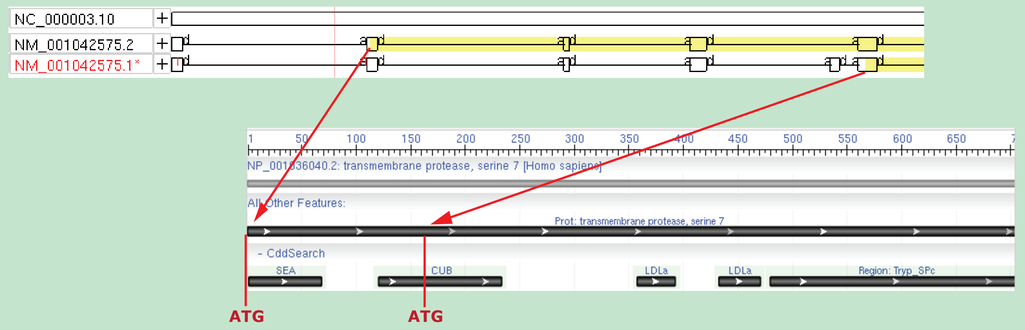
CCDS web site: http://www.ncbi.nlm.nih.gov/CCDS/



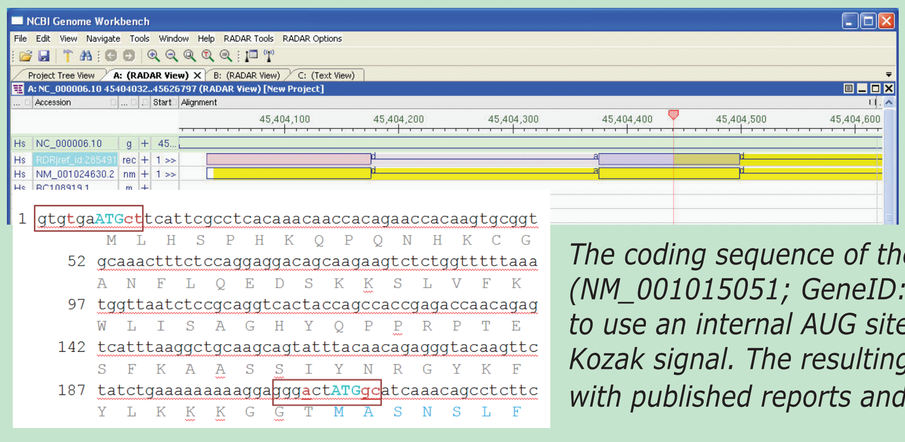
Selection of the translation initiation site:

1. *Always annotate the CDS from the first AUG*
2. *Exceptions:*
   a. *experimental evidence supports use of a downstream AUG*
   b. *downstream AUG is consistent with historical definitions that continue to support communication and there isn't stronger support for the upstream site.*
   c. *the upstream site has a poor Kozak signal and is not conserved in any species*
      --and the downstream site has a strong Kozak, is strongly conserved
      --and no additional support for use of the upstream site (e.g. gaining domains)

RefSeq & CCDS curation of TMPRSS7
(CCDS43129.1, NM_001042575.2, NP_001036040.2)



Genome workbench display of NM_001042575.2 (top; RefSeq mode) and Entrez Graphics display of the updated protein (NP_001036040.2) illustrating the N-terminal extension made in the RefSeq record and agreed upon by the CCDS collaboration. The 145 amino acid extension completes a CUB domain and adds a SEA domain.



The coding sequence of the RUNX2 RefSeq (NM_001015051; GeneID:860) was updated to use an internal AUG site that has a stronger Kozak signal. The resulting protein is consistent with published reports and clinical testing use.

Nature Precedings : doi:10.1038/npre.2009.3287.1 : Posted 26 May 2009