Ensembl gene annotation project (*e!76*)
Homo sapiens (human, GRCh38 assembly)

This document describes the annotation process of the high-coverage human assembly, described in Figure 1. The first stage is assembly loading where databases are prepared and the assembly loaded into the database.

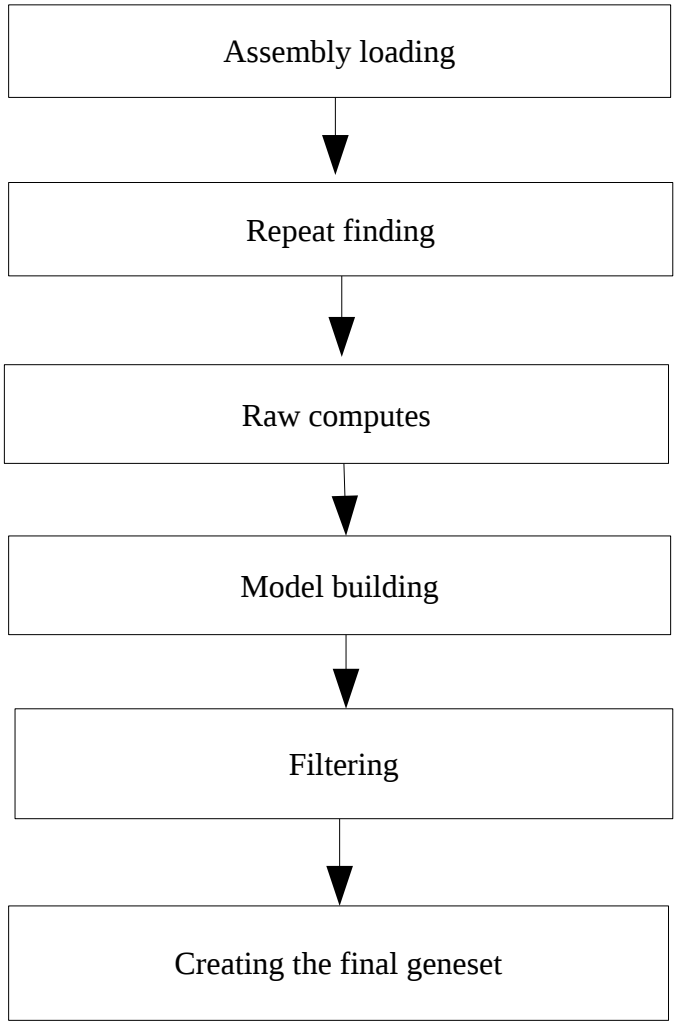

Figure 1: The gene annotation pipeline

Repeat finding

After loading into a database the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.3.0 with parameters '`-nolow -species "homo" -s`'), Dust [2] and TRF [3].

Both executions of RepeatMasker and Dust combined masked 50.45% of the species genome.

Raw computes

Transcription start sites were predicted using Eponine–scan [4] and FirstEF [5]. CpG islands [Micklem, G.] longer than 400 bases and tRNAs [6] were also predicted. The results of Eponine-scan, FirstEF, CpG, and tRNAscan are for display purposes only; they are not used in the gene annotation process.

Genscan [7] was run across repeat-masked sequence and the results were used as input for UniProt [8], UniGene [9] and Vertebrate RNA [10] alignments by WU-BLAST [11]. Passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required. This resulted in 451,431 UniProt, 348,921 UniGene and 346,165 Vertebrate RNA sequences aligning to the genome.

cDNA and EST alignments

Human cDNAs and ESTs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate. The cDNA alignments provide supporting evidence for models.

Species	cDNA/EST	Sequences Downloaded	Sequences Aligned
human	cDNA	300,648	159,081
	EST	8,705,408	3,976,554

Table 1: cDNA/EST alignments

Both cDNA and EST alignments were at a cut-off of 90% coverage and 97% identity.

Model generation

Various sources of protein data were investigated and used to generate gene models using a variety of techniques. The data and techniques employed to generate models are outlined here. The numbers of gene models generated are described in Table 2.

Pipeline	Source	Number of Models
Targeted	61,068 UniProt human proteins (PE 1 and 2, excluding fragments, minimum length = 15 aa) 36,861 RefSeq human proteins (NP, minimum length = 15 aa) 108,207 annotated cDNA sequences	131,118

Table 2: Gene model generation overview

Targeted pipeline: generating coding models using species specific proteins

Protein and cDNA sequences for human were downloaded from public databases (UniProt SwissProt/TrEMBL [8] and RefSeq [9] for proteins and ENA/Genbank/DDBJ and RefSeq [9] for cDNAs). We filtered the human protein and cDNA input sequences, for example by excluding sequences labelled as PE 3,4, or 5 by UniProt and sequences submitted by the NEDO and Genoscope projects. The resulting set contained 74,356 UniProt and RefSeq proteins and 108,207 cDNAs. The proteins were mapped to the genome using Pmatch set at a low threshold (-T 14). Three sets of coding models were then produced from the proteins using Exonerate [12] and Genewise [13]. The latter was run with two different sets of parameters to accommodate for cases where some coding models contain non-canonical (non GT/AG) splice sites. In parallel to the Genewise step, human cDNAs with known CDS start and end coordinates were aligned to the genome using Exonerate to generate a third set of coding models. Because all cDNAs used in this step had known pairing with proteins (e.g. RefSeq cDNAs with accession prefix “NM_” matching RefSeq proteins with “NP_” prefix), it allowed the comparison of coding models generated by Exonerate for a given cDNA to those generated by Genewise using its counterpart protein.

Where protein and cDNA sequences had generated more than one coding model at a locus, the BestTargeted module was used to select the coding model that most closely matched the source protein to take through to the next stage of the gene annotation process. This pipeline is shown in Figure 2.

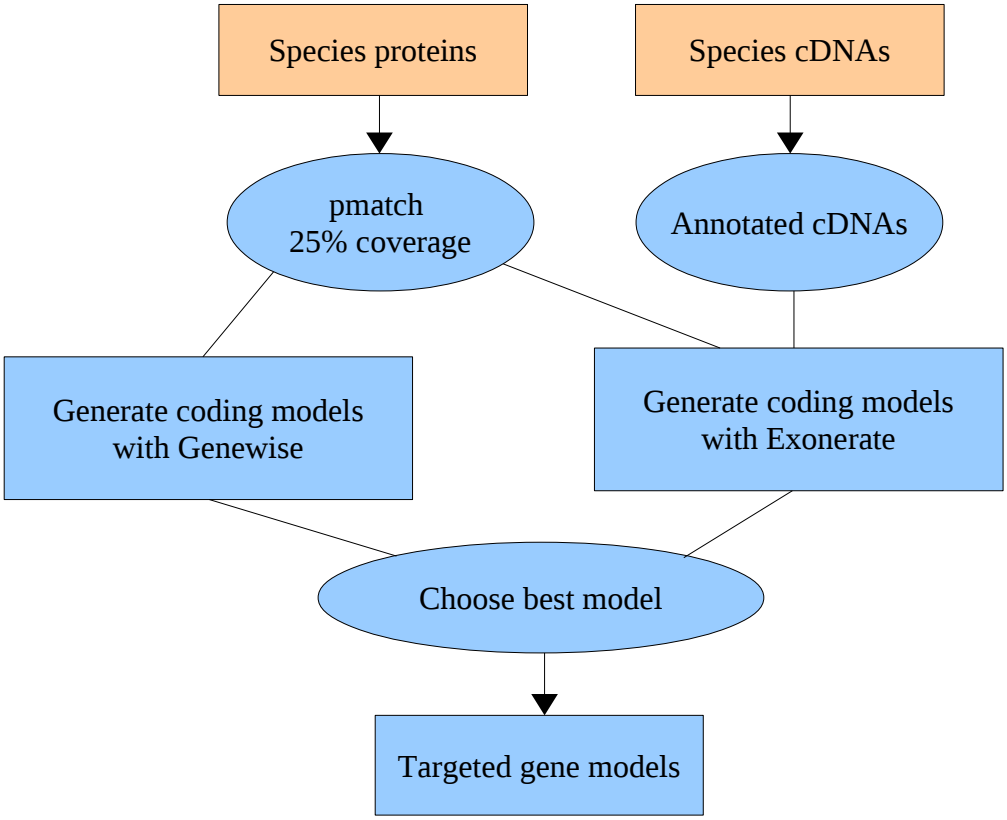

Figure 2: Targeted pipeline

Filtering the models

The filtering phase decided the subset of protein-coding transcript models, generated from the model-building pipelines, that comprise the final protein-coding gene set.

Models were filtered using the TranscriptConsensus and GeneBuilder modules. Additionally, some models based on reported dubious protein or cDNA evidence were manually removed at this stage.

Apollo software [15] was used to visualise the results of filtering.

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using cDNA sequences. At the UTR addition stage 30,843 gene models out of 106,060 had UTR added.

Generating multi-transcript genes

The above steps generated a large set of potential transcript models, many of which overlapped one another. Redundant transcript models were collapsed and the remaining unique set of transcript models were clustered into multi-transcript genes where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene.

At this stage the gene set comprised 22,315 genes with 39,711 transcripts.

Pseudogenes

The Pseudogene module was run to identify pseudogenes from within the set of gene models. A total of 618 genes were labelled as pseudogenes or processed pseudogenes.

Creating the Ensembl gene set

ncRNAs

Small structured non-coding genes were added using annotations taken from RFAM [16] and miRBase [17]. WU-BLAST was run for these sequences and models built using the Infernal software suite [24].

Cross-referencing

Before public release the transcripts and translations were given external references (cross-references to external databases). Translations were searched for signatures of interest and labelled where appropriate.

Stable identifiers

Stable identifiers were assigned to each gene, transcript, exon and translation. When annotating a species for the first time, these identifiers are auto-generated. In all subsequent annotations for a species, the stable identifiers are propagated based on comparison of the new gene set to the previous gene set.

As human has been previously released in Ensembl a comparison was made to the previous gene set.

Ensembl gene set summary

The Ensembl gene set consists of 21,694 protein coding genes (excluding 13 mitochondrial genes). These contain 39,106 transcripts. A total of 618 pseudogenes were identified. 10,312 ncRNA genes were added by the ncRNA pipeline.

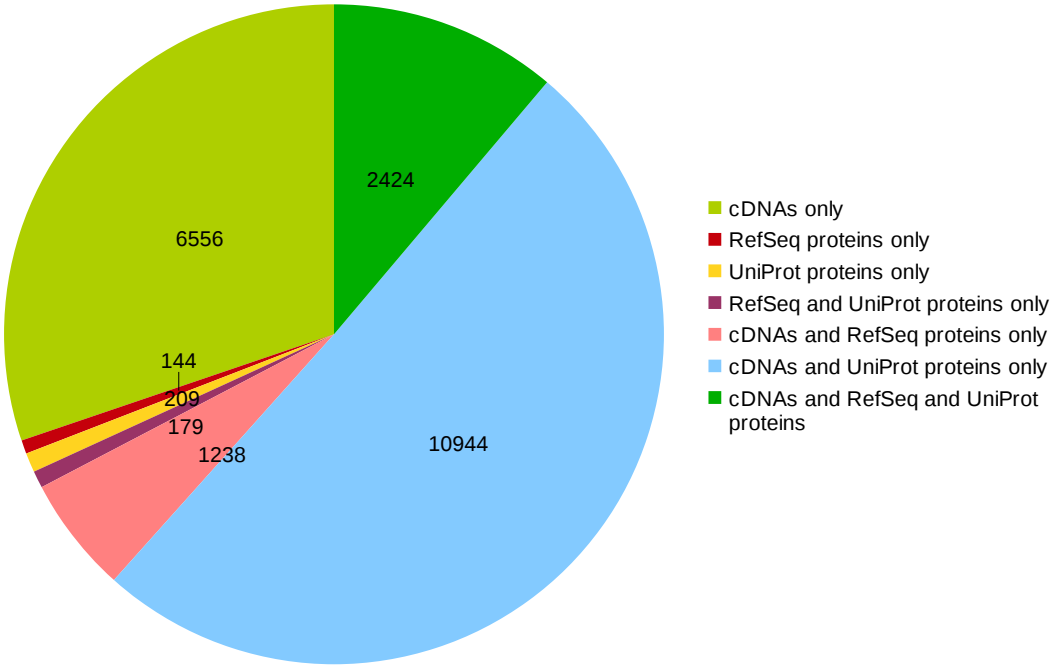

Figure 3: Supporting evidence for the protein coding gene models

Further information

The Ensembl gene set is generated automatically, meaning that gene models are annotated using the Ensembl gene annotation pipeline. The main focus of this pipeline is to generate a conservative set of protein-coding gene models, although non-coding genes and pseudogenes may also annotated.

Every gene model produced by the Ensembl gene annotation pipeline is supported by biological sequence evidence (see the “Supporting evidence” link on the left-hand menu of a Gene page or Transcript page); *ab initio* models are not included in our gene set. *Ab initio* predictions and the full set of cDNA and EST alignments to the genome are available on our website.

The quality of a gene set is dependent on the quality of the genome assembly. Genome assembly can be assessed in a number of ways, including:

1. Coverage estimate
    - A higher coverage usually indicates a more complete assembly.
    - Using Sanger sequencing only, a coverage of at least 2x is preferred.
2. N50 of contigs and scaffolds
    - A longer N50 usually indicates a more complete genome assembly.
    - Bearing in mind that an average human gene may be 10-15 kb in length, contigs shorter than this length will be unlikely to hold full-length gene models.
3. Number of contigs and scaffolds
    - A lower number toplevel sequences usually indicates a more complete genome assembly.
4. Alignment of cDNAs and ESTs to the genome

- A higher number of alignments, using stringent thresholds, usually indicates a more complete genome assembly.

More information on the Ensembl automatic gene annotation process can be found at:

- Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res.* 2004, **14(5):**942-50. [PMID: 15123590]
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M: **The Ensembl analysis pipeline.** *Genome Res.* 2004, **14(5):**934-41. [PMID: 15123589]
- http://www.ensembl.org/info/docs/genebuild/genome_annotation.html
- http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembl-doc/pipeline_docs/the_genebuild_process.txt?root=ensembl&view=co

Merging Ensembl and HAVANA gene sets, annotating long intergenic non-coding RNA genes and generating the GENCODE gene set.

Approximate time: 10 weeks

Following the completion of the Ensembl gene set, Ensembl annotations and manual annotations (primarily generated by the HAVANA team at the Wellcome Trust Sanger Institute) from the Vega database [18, 19] were merged at the transcript level to create the final gene set. The Vega database (as of 2 April 2014) contained 53,696 genes and 192,145 transcripts. In the merge process, Ensembl and HAVANA transcripts were merged if they had identical intron chains. If transcripts from the two annotation sources matched at all internal exon-intron boundaries, i.e. had identical splicing pattern, the Ensembl model was merged into the HAVANA model and the resulting merged transcript would adopt the exon-intron structure of the HAVANA transcript. Transcripts which had not been merged, either because of differences in internal exon-intron boundaries or presence of transcripts in only one annotation source, were transferred from the source to the final gene set intact.

Biotype conflicts between Ensembl and HAVANA were always reported to the HAVANA team for investigation, and when resolved, could improve the merged gene set in the future. As for supporting evidence, the merge of Ensembl and HAVANA transcripts also involved merging of protein and cDNA supporting evidence associated with the transcripts to ensure the basis on which the annotations were made would not be lost.

Following the merge, the long intergenic non-coding RNA genes (lincRNAs) annotated by the Ensembl lincRNA pipeline [20] on the human GRCh37 assembly were projected onto the GRCh38 assembly and incorporated in the final gene set.

An important feature of the merged gene set is the presence of all HAVANA source transcripts. This has been made possible by allowing HAVANA annotation to take precedence over Ensembl's when merging transcripts which do not match at their terminal exons or have different biotypes. Of all HAVANA transcripts, 14.4% of them were merged with Ensembl transcripts. The vast majority of merged transcripts (97.3%) are of protein-coding biotype. HAVANA transcripts which were not merged (85.6% of HAVANA source transcripts) were mostly alternative splice variants, pseudogenes or non-coding. These transcripts were fully transferred into the final gene set. The final Ensembl-HAVANA set consisted of 63,263 genes and 206,771 transcripts. Of these transcripts, 12.9% (26,731) were the result of merging Ensembl and HAVANA annotations, 10.2% (21,128) originated from Ensembl, 76.6% (158,287) originated from HAVANA, and the remaining 0.3% were incorporated from other sources (e.g. mithochondrial genes and LRG genes [21]).

As a quality-control measure, Ensembl translations of protein-coding transcripts in the final merged gene set were aligned against the NCBI RefSeq and Uniprot/SwissProt sets of public curated protein sequences (which were used in the “Targeted” stage of the gene build) to calculate the proportion of curated sequences covered by the merged gene set. Over 99%

of RefSeq and SwissProt proteins were represented in the merged gene set, and in the majority of cases, there was a 100% match between the curated protein and Ensembl translation.

Since Ensembl release 56 (September 2009), the Ensembl-HAVANA gene set has exactly corresponded to a GENCODE release [22]. The gene set in release 76, which this document describes, corresponds to GENCODE release 20. Each GENCODE release also contains the full annotation of the consensus coding sequence (CCDS) transcript models [23]. All CCDS models are included in each release of the human gene set.

References

1 Smit, AFA, Hubley, R & Green, P: **RepeatMasker Open-3.0.** 1996-2010. www.repeatmasker.org

2 Kuzio J, Tatusov R, and Lipman DJ: **Dust.** Unpublished but briefly described in: Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology* 2006, **13(5):**1028-1040.

3 Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res.* 1999, **27(2):**573-580. [PMID: 9862982] http://tandem.bu.edu/trf/trf.html

4 Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res.* 2002 **12(3):**458-461. http://www.sanger.ac.uk/resources/software/eponine/ [PMID: 11875034]

5 Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet.* 2001, **29(4):**412-417. [PMID: 11726928]

6 Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res.* 1997, **25(5):**955-64. [PMID: 9023104]

7 Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol.* 1997, **268(1):**78-94. [PMID: 9149143]

8 Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R: **A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res.** 2010, **38 Suppl:**W695-699. http://www.uniprot.org/downloads [PMID: 20439314]

9 Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2010, **38(Database issue):D5-16.** [PMID: 19910364]

10 http://www.ebi.ac.uk/ena/

11 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol.* 1990, **215(3):**403-410. [PMID: 2231712]

12 Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatic*s 2005, **6:**31. [PMID: 15713233]

13 Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res.* 2004, **14(5):**988-995. [PMID: 15123596]

14 Eyras E, Caccamo M, Curwen V, Clamp M: **ESTGenes: alternative splicing from ESTs in Ensembl.** *Genome Res.* 2004 **14(5):**976-987. [PMID: 15123595]

15 Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol.* 2002, **3(12):**RESEARCH0082. [PMID: 12537571]

16 Griffiths-Jones S., Bateman A., Marshall M., Khanna A., Eddy S.R: **Rfam: an RNA family database.** Nucleic Acids Research (2003) **31(1):**p439-441. [PMID: 12520045]

17 Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** NAR 2006 **34(Database Issue):**D140-D144 [PMID: 16381832]

18 http://vega.sanger.ac.uk/Homo_sapiens/Info/Index

19 Wilming L. G., Gilbert J. G. R., Howe K., Trevanion S., Hubbard T. and Harrow J. L: **The vertebrate genome annotation (Vega) database.** Nucleic Acid Res. 2008 Jan; Advance Access published on November 14, 2007; doi:10.1093/nar/gkm987 [PMID: 18003653]

20 http://www.ensembl.org/info/docs/genebuild/ncrna.html

21 http://www.lrg-sequence.org/

22 Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. et al. **GENCODE: producing a reference**

annotation for ENCODE. *Genome Biol*., 2006 **7**(Suppl. 1), S4.1–S4.9.

23 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. **The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res.* 2009, **19**, 1316–1323.

24 Eddy, SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure**. BMC Bioinformatics 2002, 3:18. [PMID:12095421]