Ensembl gene annotation project (*e!80*)

Rattus norvegicus (rat, Rnor_6.0)

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

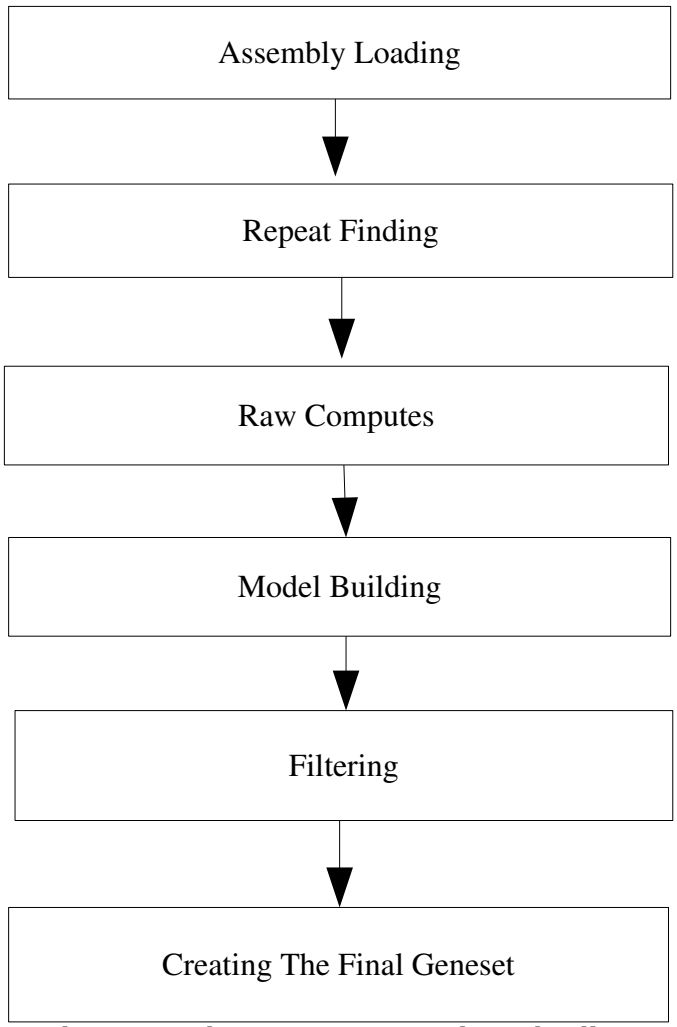


Figure 1: The Gene Annotation Pipeline

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 "rattus norvegicus" -s***'),

RepeatModeler [2] (version open-1.0.5, to obtain a repeats library, then filtered for an additional RepeatMasker run), Dust [3] and TRF [4]. Both executions of RepeatMasker and Dust combined masked 46.57% of the species genome.

Raw Computes

Transcription start sites were predicted using Eponine–scan [5] and FirstEF [6]. CpG islands [Micklem, G.] longer than 400 bases and tRNAs [7] 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 [8] was run across repeat-masked sequence and the results were used as input for UniProt [9], UniGene [10] and Vertebrate RNA [11] alignments by WU-BLAST [12]. Passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required. This resulted in 10,056,246 UniProt, 7,361,390 UniGene and 12,247,902 Vertebrate RNA sequences aligning to the genome.

cDNA and EST Alignments

rat 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
rat	cDNA	109,095	65,192
	EST	1,103,339	994,706

Table 1: cDNA/EST alignments

All alignments were at a cut-off of 90% coverage and 95% identity.

Model Generation

Various sources of transcript and 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	13,195 UniProt rat proteins	20,041
	17,526 RefSeq rat proteins	
	40,053 UniProt mouse proteins	75,459
	77,908 RefSeq mouse proteins	
	65,192 rat cDNAs	29,205
Similarity	90,992 UniProt proteins	115,762
RNASeq	Hubrecht Institute KNAW	22,313

Table 2: Gene Model Generation Overview

Targeted Pipeline: Generating coding models using species specific proteins

Protein sequences for rat and mouse were downloaded from public databases (UniProt SwissProt/TrEMBL [9] and RefSeq [10]). The rat and mouse protein sequences were mapped to the genome using Pmatch set at a low threshold (-T 14). Two sets of coding models were then produced from the proteins using Exonerate [13] and Genewise [14].

In parallel, rat 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 one protein sequence 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.

Similarity Pipeline: Generating coding models using proteins from related species

Coding models were generated using data from related species. WU-BLAST was rerun for the UniProt alignments from the Raw Computes step and the results were passed to Genewise [14] to build coding models.

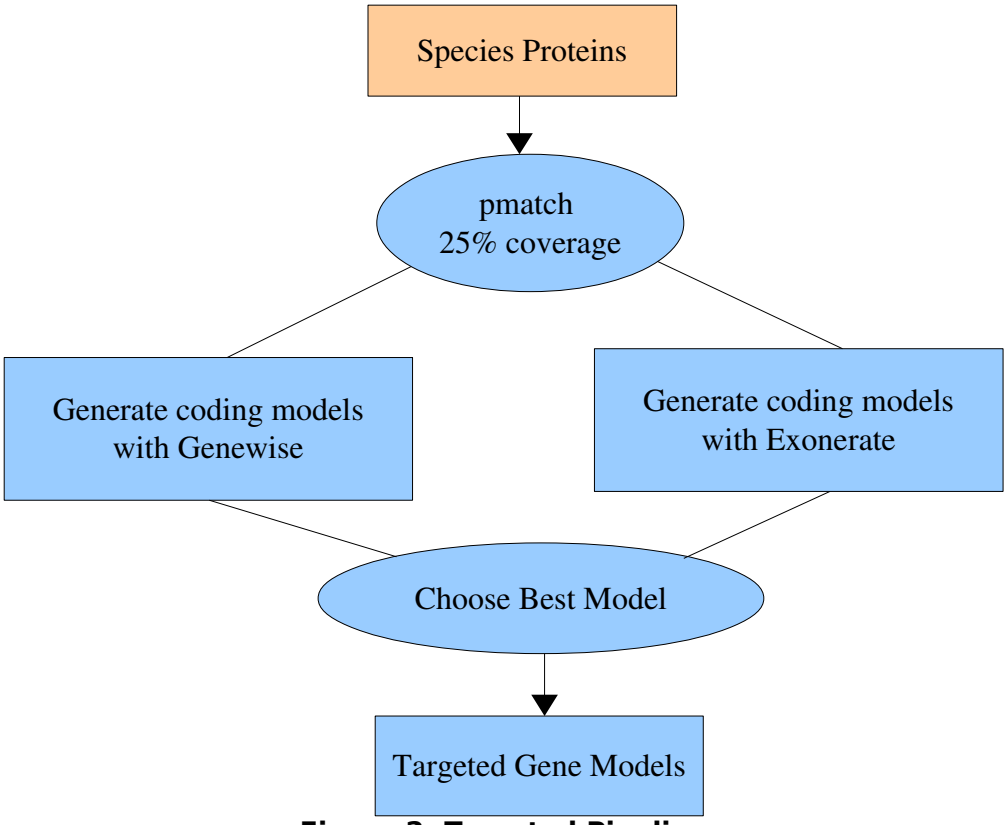


Figure 2: Targeted Pipeline

RNASeq Pipeline

RNASeq data provided by the Hubrecht Institute KNAW was used in the annotation. This comprised a mixture of single and paired end data from samples including: a pool of 12 tissues, liver, muscle, brain, heart, thymus, blood, testis, ovary, kidney, skin, spleen and lung. The available reads were aligned to the genome using BWA. The Ensembl RNASeq pipeline was used to process the BWA alignments and create further split read alignments using Exonerate.

The split reads and the processed BWA alignments were combined to produce 192,388 transcript models in total. The predicted open reading

frames were compared to UniProt proteins using WU-BLAST. Models with poorly scoring or no BLAST alignments were split into a separate class.

Filtering the Models

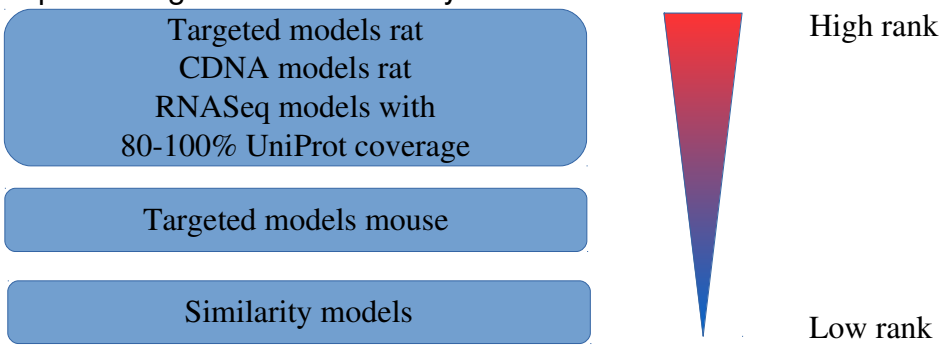
The filtering phase decided the subset of protein-coding transcript models, generated from the model-building pipelines, that comprise the final protein-coding gene set.

Models were filtered using the TranscriptConsensus, LayerAnnotation and GeneBuilder modules.

Apollo software [16] was used to visualise the results of filtering.

LayerAnnotation

The LayerAnnotation module was used to define a hierarchy of input data sets, from most preferred to least preferred. The output of this pipeline included all transcript models from the highest ranked input set. Models from lower ranked input sets are included only if their exons do not overlap a model from an input set higher in the hierarchy.



Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNASeq, cDNA and EST sequences. At the UTR addition stage 122,091 gene models out of 153,340 non-RNASeq pipeline generated gene models 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 23,718 genes with 29,327 transcripts.

Pseudogenes

The Pseudogene module was run to identify pseudogenes from within the set of gene models. A total of 1,039 genes were labelled as pseudogenes or processed pseudogenes.

Creating The Final Gene Set

ncRNAs

Small structured non-coding genes were added using annotations taken from RFAM [17] and miRBase [18]. WU-BLAST was run for these sequences and models built using the Infernal software suite [20].

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

Final Gene Set Summary

The final gene set consists of 22,379 protein coding genes, including 13 mitochondrial genes. These contain 26,494 transcripts. A total of 1,039 pseudogenes were identified. 5,843 ncRNAs were added by the ncRNA pipeline.

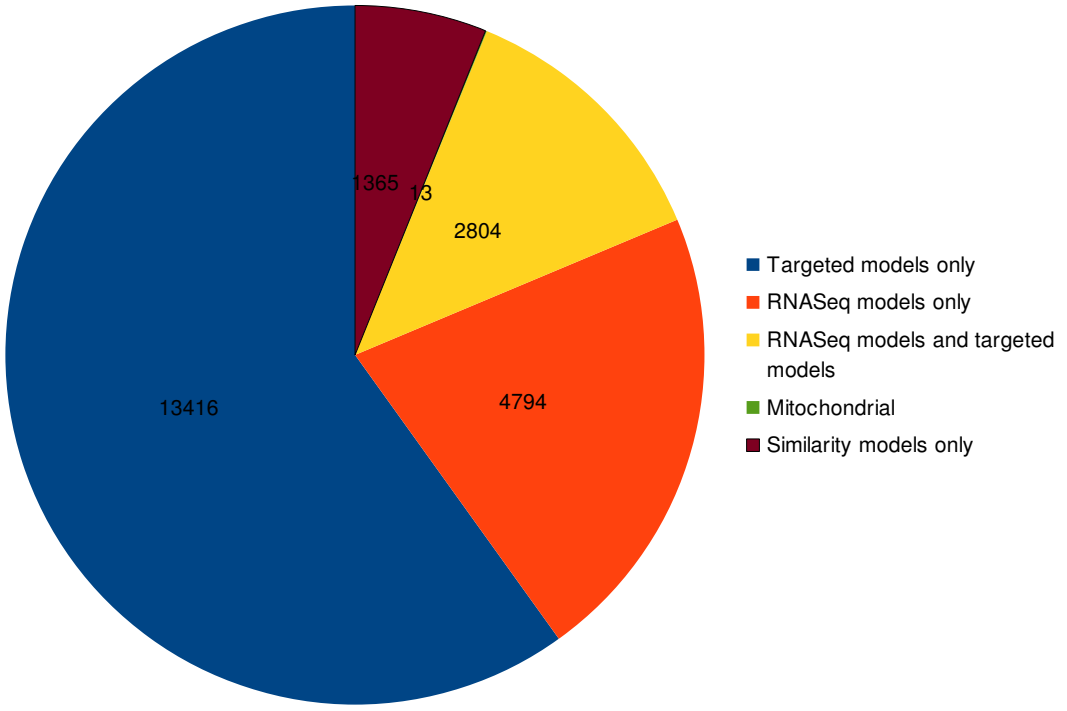


Figure 3: Supporting evidence for the protein coding gene models

Merging Ensembl and HAVANA gene sets

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 [12] were merged at the transcript level to create the final gene set. The Vega database (as of 2 February 2015) contained 1,615 genes and 2,709 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.

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, 25.06% of them were merged with Ensembl transcripts. The vast majority of merged transcripts (92.64%) are of protein-coding biotype. HAVANA transcripts which were not merged (74.94% 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 32,545 genes and 39,595 transcripts. Of these transcripts, 1.71% (679) were the result of merging Ensembl and HAVANA annotations, 93.06% (36,849) originated from Ensembl, 5.13% (2,030) originated from HAVANA, and the remaining 0.09% were incorporated from other sources (e.g. mithochondrial genes).

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/genome/genebuild/genome_annotation.html
- https://github.com/Ensembl/ensembl-doc/blob/master/pipeline_docs/the_genebuild_process.txt

References

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

2 Smit, AFA, Hubley, R. **RepeatModeler Open-1.0.** 2008-2010. www.repeatmasker.org

3 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.

4 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

5 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]

6 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]

7 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]

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

9 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]

10 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]

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

12 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]

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

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

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

16 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]

17 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]

18 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]

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 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]