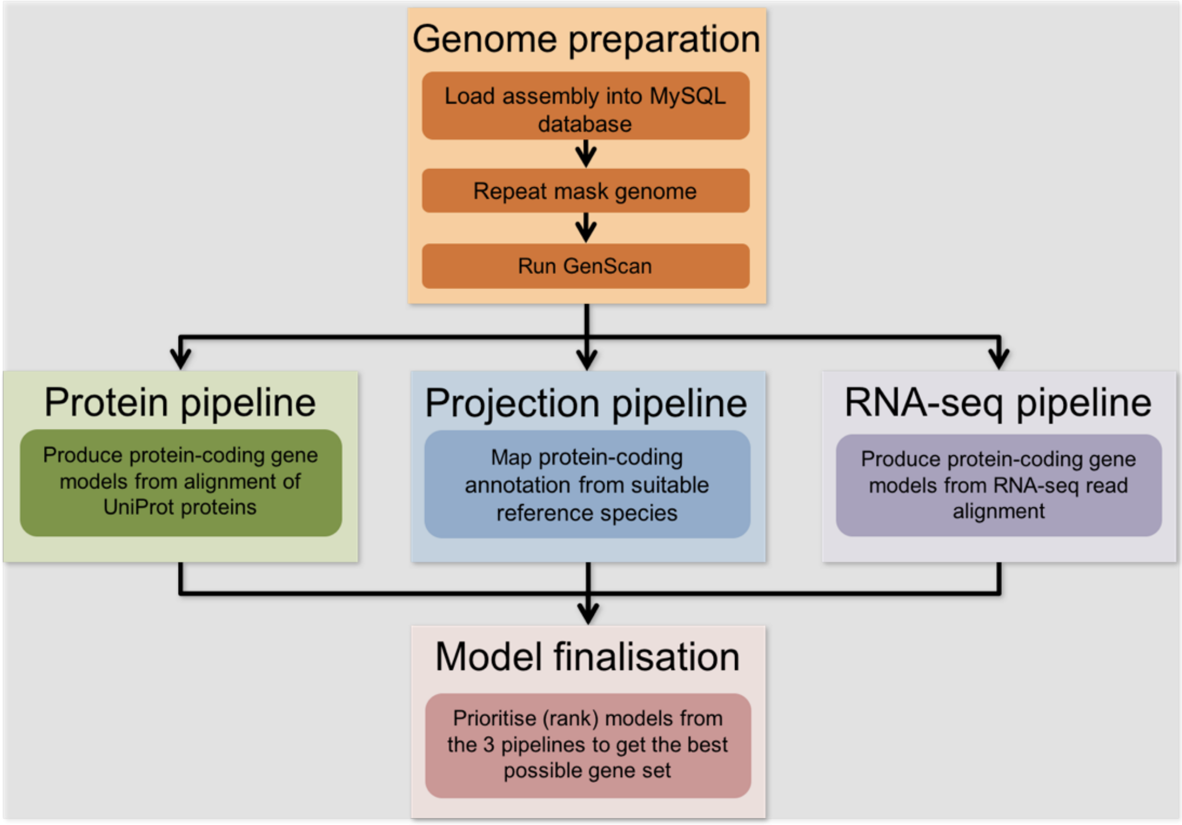
# Ensembl gene annotation

This document describes the annotation process of the assembly. The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.


Genome preparation
Load assembly into MySQL database
Repeat mask genome
Run GenScan
Protein pipeline
Produce protein-coding gene models from alignment of UniProt proteins
Projection pipeline
Map protein-coding annotation from suitable reference species
RNA-seq pipeline
Produce protein-coding gene models from RNA-seq read alignment
Model finalisation
Prioritise (rank) models from the 3 pipelines to get the best possible gene set

The above figure shows a simplified view of the standard annotation process.

# Section 1: Genome Preparation

The genome phase of the Ensembl gene annotation pipeline involves loading an assembly into the Ensembl core database schema and then running a series of analyses on the loaded assembly to identify an initial set of genomic features.

The most important aspect of this phase is identifying repeat features (primarily through RepeatMasker) as softmasking of the genome is used extensively later in the annotation process.

## Repeat Finding

After loading into a database the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version  4.0.5 with parameters, using as the search engine), Dust [3] and TRF [4]. The masked part of each assembly displayed in appendix. The Repbase rodents library was used with RepeatMasker.

## Low complexity features, ab initio predictions and BLAST analyses

Transcription start sites were predicted using Eponine–scan [5]. CpG islands [Micklem, G.] longer than 400 bases and tRNAs [7] were also predicted. The results of Eponine-scan, 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 to identify ab inito gene predictions.

The results of the Genscan analyses were also 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.

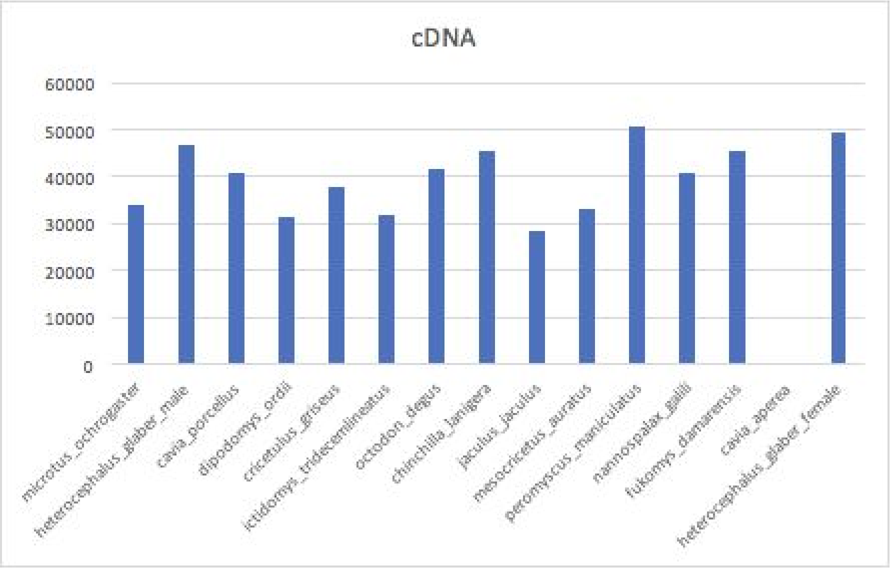
Genscan predictions are for display purposes only and are not used in the model generation phase

# Section 2: Protein-Coding 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 gene summary.

## cDNA alignment pipeline:

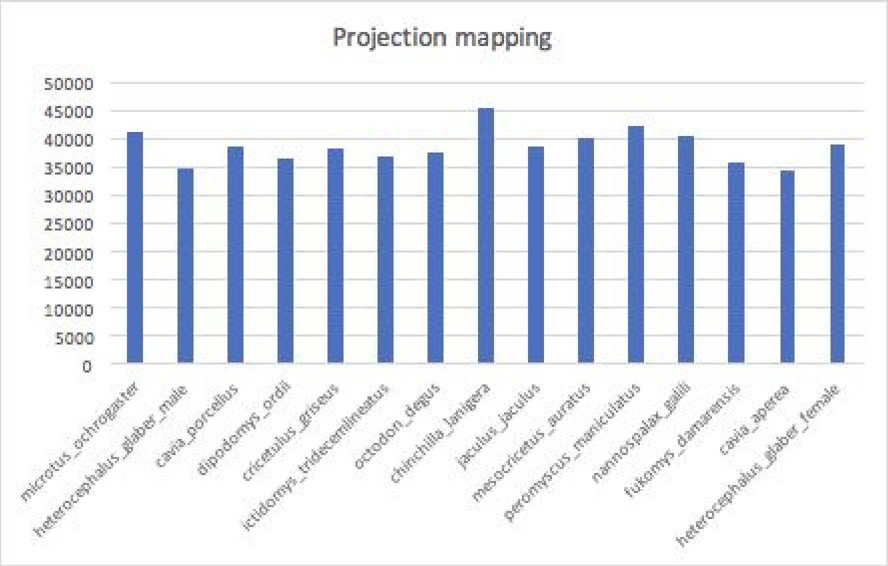
cDNAs were downloaded from RefSeq and aligned to the genome using Exonerate [13]. Only known mRNAs were used (NMs). A minimal sequence length of 60bp was and a cut-off of 97% identity and 90% coverage were required for an alignment to be kept. The cDNAs are mainly used for display purposes, but can be used to add UTR to the protein coding transcript models if they have a matching set of introns.



## Projection mapping pipeline

For all species a whole genome alignment was generated against the mouse reference assembly (GRCm38) using LastZ. Syntenic regions identified using this alignment were then used to map protein coding annotation from the GENCODE M11 gene set.

The mapped transcripts were then assessed for non-canonical splice sites and frameshifts. This can happen when mapping coordinates from one assembly to another. Mapped transcripts featuring two or more non-canonical splice sites/frameshifts were passed into a realignment pipeline, that re-aligned the original sequence in the region it was mapped to to see if a model with canonical splicing could be built. If this was not possible the transcript model was disgarded.

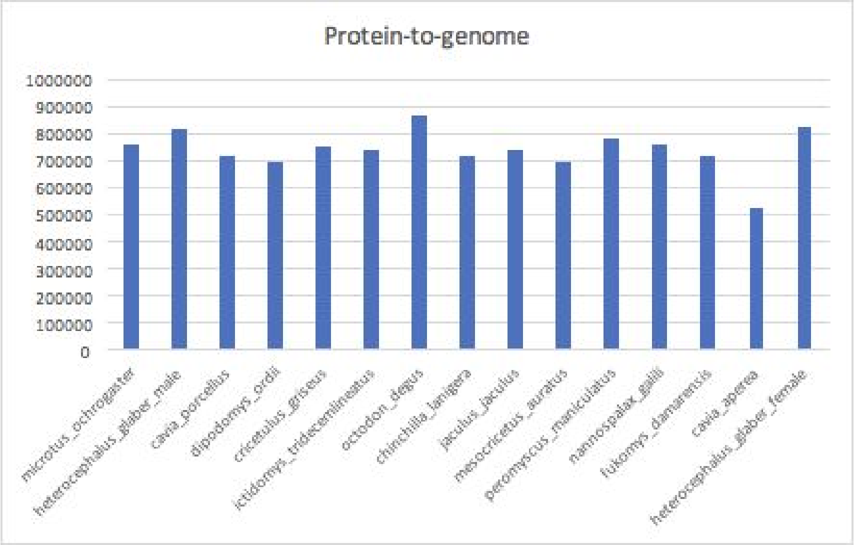


## Protein-to-genome pipeline

Protein sequences were downloaded from UniProt and aligned to the genome in a splice aware manner using GenBlast [21]. The set of proteins aligned to the genome was a subset of UniProt proteins used to provide a broad, targeted coverage of the rodent proteome. The set consists of the following:

- Mouse SwissProt/TrEMBL PE 1 & 2
- Human SwissProt/TrEMBL PE 1 & 2
- Other rodents  SwissProt/TrEMBL PE 1 & 2 & 3
- Other mammals  SwissProt/TrEMBL PE 1 & 2
- Other vertebrates  SwissProt/TrEMBL PE 1 & 2

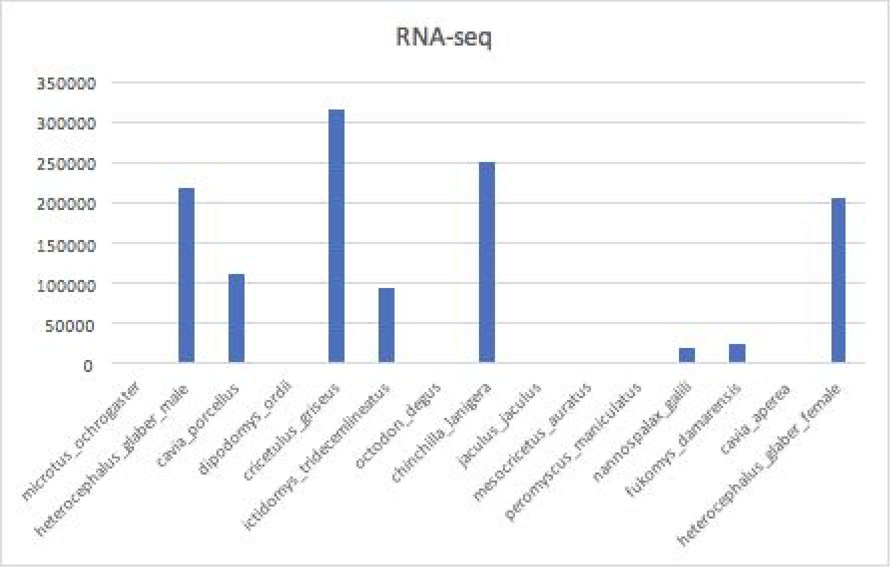
Note: PE stands for UniProt protein existence level. See here for more detail.
A cut-off of 50 percent coverage and identity and an e-value of e-20 were used for GenBlast with the exon repair option turned on. The top 5 transcript models built by GenBlast for each protein passing the cut-offs were kept.



## RNA-seq pipeline

RNA-seq data downloaded from ENA, was used in the annotation. A merged file contain reads from all tissues/samples was also created. The merged was less likely to suffer from model fragmentation due to read depth. The available reads were aligned to the genome using BWA, with a tolerance of 50 percent mismatch to allow for intron identification via split read alignment. Initial models generated from the BWA alignments were further refined via exonerate. Protein coding models were identified via a BLAST alignment of the longest ORF against the UniProt vertebrate PE 1 & 2 data set. Models with poorly scoring or no BLAST alignments were split into a separate class and considered as potential lincRNAs.

In the case where multiple tissues/samples were available we created a gene track for each such tissue/sample that can be viewed in the Ensembl browser and queried via the API.



# Section 3: Filtering The Protein-Coding 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. Model are filtered based on information such as what pipeline they were generated using, how closely related the data are to the target species and how good the alignment coverage and percent identity to the original data are.

## Prioritising models at each locus

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.

Note that models cannot exist in more than one layer. For UniProt proteins, models were also separate into clades, to help selection during the layering process. Each UniProt protein was in one clade only, for example mammal proteins were present in the mammal clade and were not present in the vertebrate clade to avoid aligning the proteins multiple times.

When selecting the model or models kept at each position, we prioritise based on the highest layer with available evidence. In general the highest layers contain the set of evidence containing the most trustworthy evidence in terms of both alignment/mapping quality, and also in terms of relevance to the species being annotated. So for example when a rodent is being annotated then well aligned evidence from either the species itself or other closely related vertebrates would be chosen over evidence from more distant species. Regardless of what species is being annotated, well aligned human proteins are usually included in the top layer as human is the current most complete vertebrate annotation. For further details on the exact layering used please refer to section 6.

## Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNA-seq data (if available) and alignments of species-specific RefSeq cDNA sequences. The criteria for adding UTR from cDNA or RNA-seq alignments to protein models lacking UTR (such as the projection models or the protein-to-genome alignment models) was that the intron coordinates from the model missing UTR exactly matched a subset of the coordinates from the UTR donor model.

## 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.

## Pseudogenes

Pseudogenes were annotated by looking for genes with evidence of frame-shifting or lying in repeat heavy regions. Single exon retrotransposed pseudogenes were identified by searching for a multi-exon equivalent elsewhere in the genome. A total number of genes that are labelled as pseudogenes or processed pseudogenes will be included in the core db, please check Final Gene set Summary.

# Section 4: Creating The Final Gene Set

## Small 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 RNAfold and the Infernal software suite [20].

## lincRNAs discovery

Using the transcriptomic data set, if available, we try to predict long intergenic non coding RNAs (lincRNAs). We used the RNA-seq data sets which were filtered against the protein-coding gene set. The candidate lincRNAs should not overlap a protein-coding gene. The Pfam analysis of InterProScan is run against the filtered gene set. A potential lincRNA should not have a Pfam domain.

## Cross-referencing

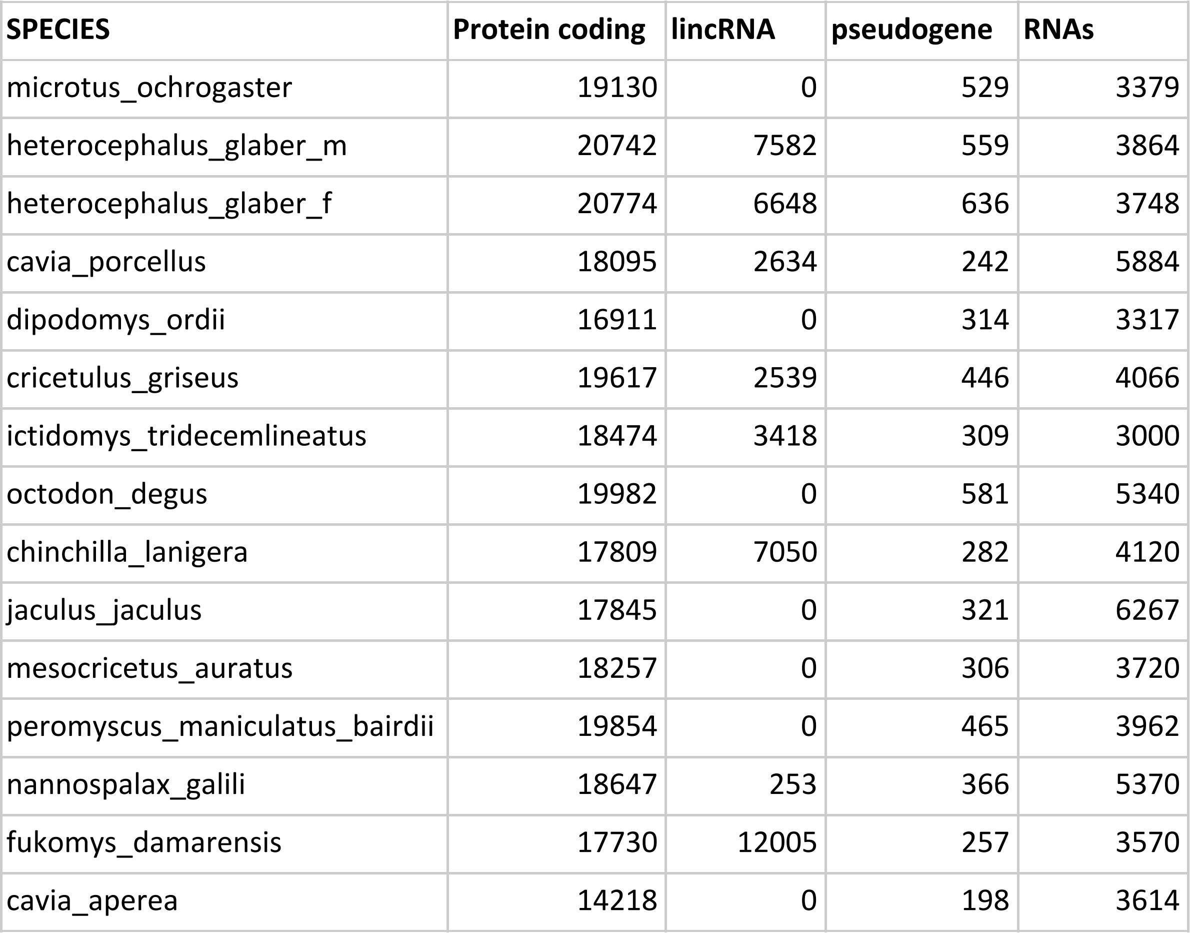
Before public release the transcripts and translations were given external references (cross-references to external databases). Translations were searched for signatures of interest and labelled where appropriate.

## Stable Identifiers

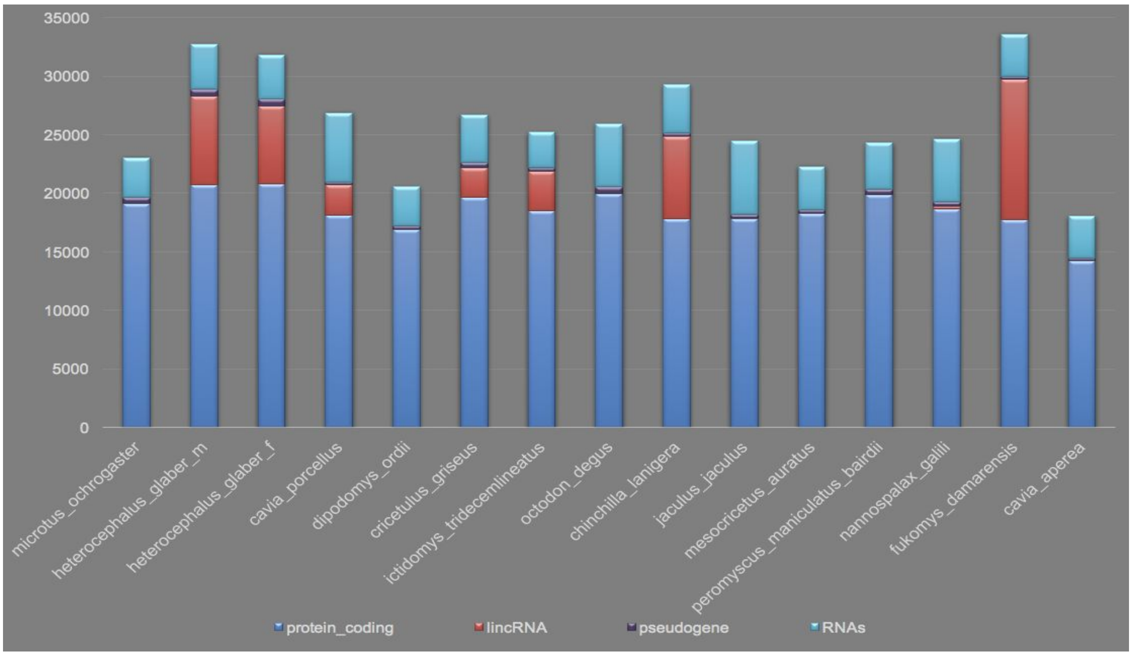
Stable identifiers were assigned to each gene, transcript, exon and translation. When annotating a species for the first time, these identifiers are auto-generated. In all subsequent annotations for a species, the stable identifiers are propagated based on comparison of the new gene set to the previous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

# Section 5: Final Gene Set Summary

The final gene set consists of 15 that annotated by Ensembl:



| SPECIES | Protein coding | lincRNA | pseudogene | RNAs |
|---|---:|---:|---:|---:|
| microtus_ochrogaster | 19130 | 0 | 529 | 3379 |
| heterocephalus_glaber_m | 20742 | 7582 | 559 | 3864 |
| heterocephalus_glaber_f | 20774 | 6648 | 636 | 3748 |
| cavia_porcellus | 18095 | 2634 | 242 | 5884 |
| dipodomys_ordii | 16911 | 0 | 314 | 3317 |
| cricetulus_griseus | 19617 | 2539 | 446 | 4066 |
| ictidomys_tridecemlineatus | 18474 | 3418 | 309 | 3000 |
| octodon_degus | 19982 | 0 | 581 | 5340 |
| chinchilla_lanigera | 17809 | 7050 | 282 | 4120 |
| jaculus_jaculus | 17845 | 0 | 321 | 6267 |
| mesocricetus_auratus | 18257 | 0 | 306 | 3720 |
| peromyscus_maniculatus_bairdii | 19854 | 0 | 465 | 3962 |
| nannospalax_galili | 18647 | 253 | 366 | 5370 |
| fukomys_damarensis | 17730 | 12005 | 257 | 3570 |
| cavia_aperea | 14218 | 0 | 198 | 3614 |



Counts of the major gene classes in each species

# Section 6: Appendix - 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

o A higher coverage usually indicates a more complete assembly.

o Using Sanger sequencing only, a coverage of at least 2x is preferred.

2. N50 of contigs and scaffolds

o A longer N50 usually indicates a more complete genome assembly.

o 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

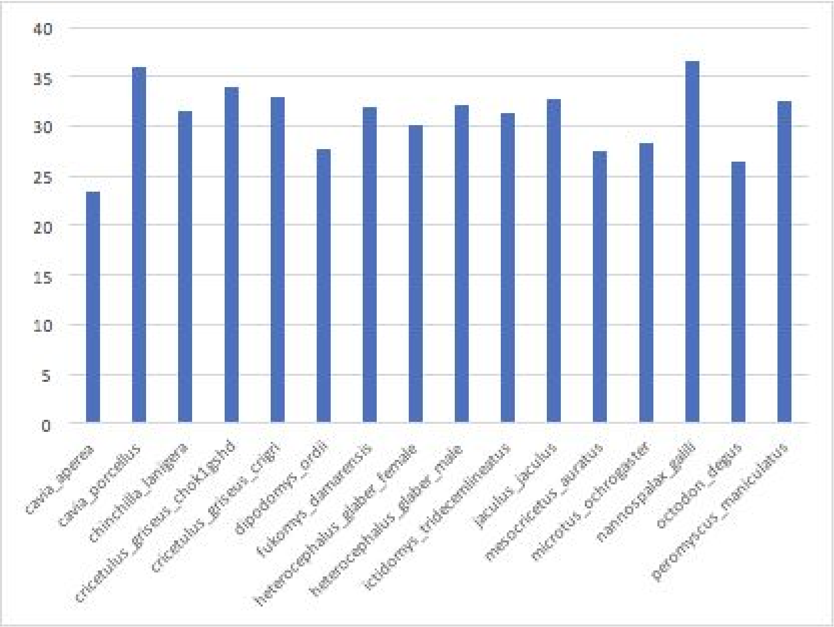
o A lower number toplevel sequences usually indicates a more complete genome assembly.

4. Alignment of cDNAs and ESTs to the genome

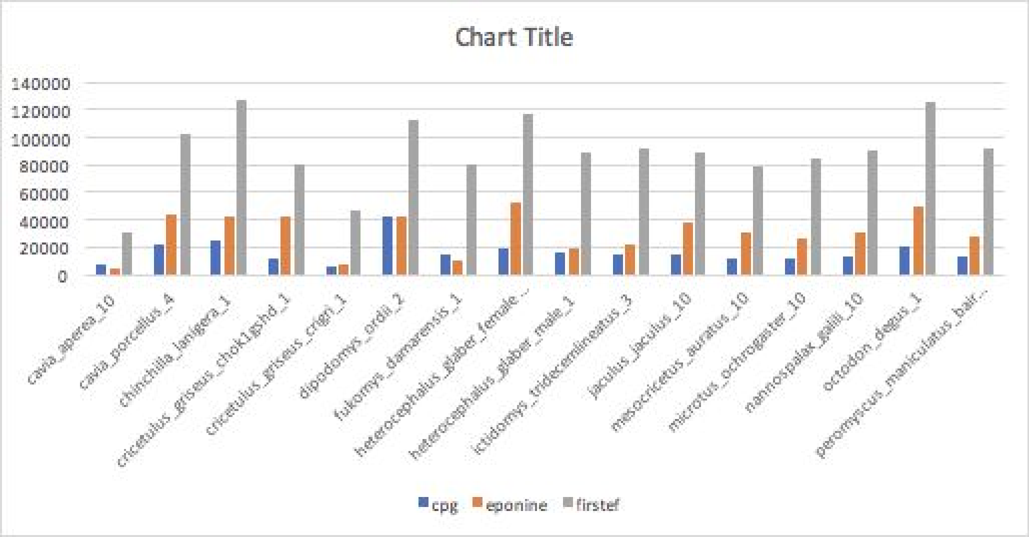
o A higher number of alignments, using stringent thresholds, usually indicates a more complete genome assembly.

More info for the assemblies:

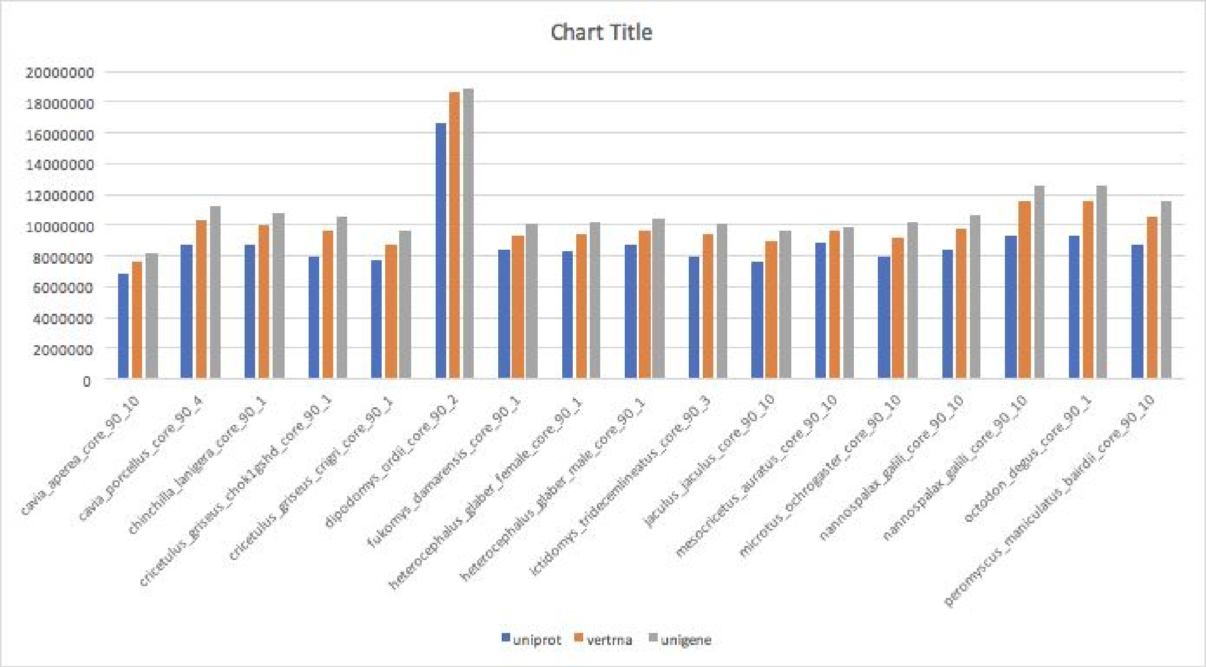
| Species name | Common name | Genbank accession ID | Assembly level |
|---|---|---|---|
| apodemus_sylvaticus | European woodmouse | GCA_001305905.1 | Scaffold |
| cavia_aperea | Brazilian guinea pig | GCA_000688575.1 | Scaffold |
| cavia_porcellus | Domestic guinea pig | GCA_000151735.1 | Scaffold |
| chinchilla_lanigera | Long-tailed chinchilla | GCA_000276665.1 | Scaffold |
| Cricetulus_griseus chok1gshd | Chinese hamster cell | | |
| cricetulus_griseus | Chinese hamster | GCA_000223135.1 | Scaffold |
| Dipodomys_ordii | Ord's kangaroo rat | GCA_000151885.2 | Scaffold |
| fukomys_damarensis | Damara mole rat | GCA_000743615.1 | Scaffold |
| heterocephalus glaber | Naked mole rat male | GCA_000230445.1 | Scaffold |
| heterocephalus glaber | Naked mole rat female | GCA_000247695.1 | Scaffold |
| ictidomys_tridecemlineatus | Thirteen-lined ground squirrel | GCA_000236235.1 | Scaffold |
| jaculus_jaculus | Lesser Egyptian jerboa | GCA_000280705.1 | Scaffold |
| mesocricetus_auratus | Golden hamster | GCA_000349665.1 | Scaffold |
| microtus_ochrogaster | Prairie vole | GCA_000317375.1 | Chromosome |
| mus_caroli | Ryukyu mouse | | Chromosome |
| mus_pahari | Gairdner's shrewmouse | | Chromosome |
| nannospalax_galili | Upper Galilee mountains blind<br>molde rat | GCA_000622305.1 | Scaffold |
| octodon_degus | Brush-tailed rat or Common degu | GCA_000260255.1 | Scaffold |
| peromyscus_maniculatus | Northern American deer mouse | GCA_000500345.1 | Scaffold |



Repeat masking percentages per genome



Count of low complexity features per genome



Counts of UniProt, VertRNA and UniGene sequences aligned per genome

Layers used in detail:
'LAYER1':['realign_95','rnaseq_95','self_pe12_sp_95','mouse_pe12_sp_95','rodents_pe12_sp_95','human_pe12_sp_95','realign_80']
'LAYER2':['self_pe12_tr_95','mouse_pe12_tr_95','rodents_pe12_tr_95','human_pe12_tr_95','self_pe12_sp_80']
'LAYER3':['mouse_pe12_sp_80','rodents_pe12_sp_80','human_pe12_sp_80','mammals_pe12_sp_95','vert_pe12_sp_95','rnaseq_80']
'LAYER4':['self_pe12_tr_80','mouse_pe12_tr_80','rodents_pe12_tr_80','human_pe12_tr_80','mammals_pe12_tr_95','vert_pe12_tr_95']
'LAYER5':['rodents_pe3_sp_95','rodents_pe3_tr_95','mammals_pe12_sp_80','vert_pe12_sp_80']
'LAYER6':['realign_50']

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

Aken B et al.: The Ensembl gene annotation system. Database 2016. [PMCID: PMC4919035]

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/index.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 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 Bioinformatics 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 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]

19 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]

20 She R, Chu JS, Uyar B, Wang J, Wang K, and Chen N: genBlastG: using BLAST searches to build homologous gene models. Bioinformatics, 2011, [PMID: 21653517]