Ensembl Gene Annotation (*e!91*)

Cat (felis catus)

Assembly: Felis_catus_8.0, GCA_000181335.3

Table of Contents

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

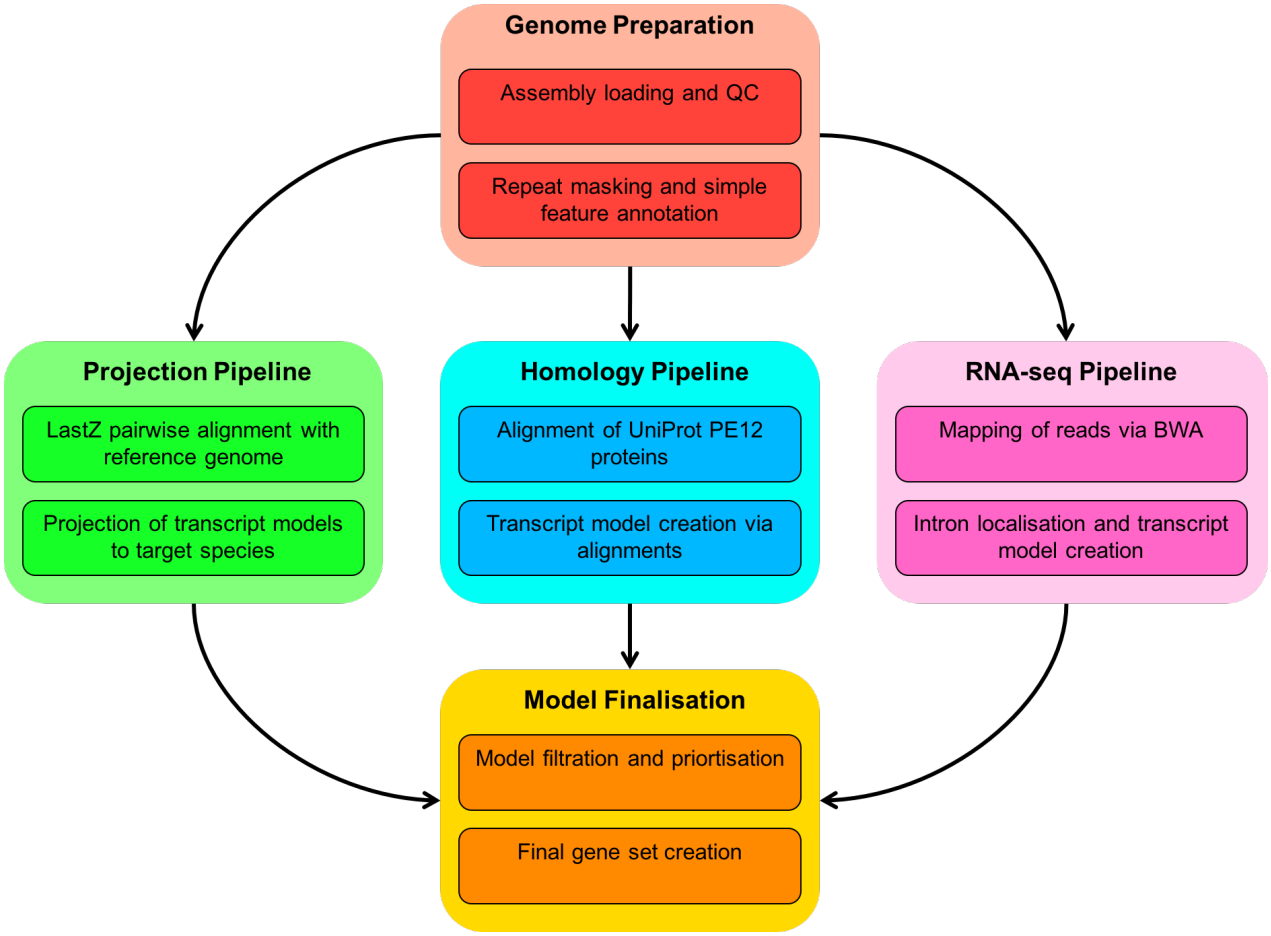

Fig. 1: Flowchart of the protein-coding annotation pipeline. Small ncRNAs, Ig genes, TR genes, and pseudogenes are computed using separate pipelines.

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 soft masking of the genome is used extensively later in the annotation process.

Repeat Finding

After the genomic sequence has been loaded into a database, it is screened for sequence patterns including repeats using RepeatMasker [1] (version 4.0.5 with parameters, using as the search engine), Dust [2] and TRF [3].

For the cat annotation, the Repbase mammal library was used with RepeatMasker.

Low complexity features, ab initio predictions and BLAST analyses

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

Genscan [6] is run across repeat-masked sequence to identify ab initio gene predictions. The results of the Genscan analyses are also used as input for UniProt [7], UniGene [8] and Vertebrate RNA alignments by NCBI-BLAST [9]. 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 are 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 are downloaded from RefSeq [10] and aligned to the genome using Exonerate [11]. 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.

For the cat annotation, a minimal sequence length of 60bp and a cut-off of 95% identity and 50% coverage were required for an alignment to be kept.

Total cDNA alignment count: 57,835

Protein-to-genome pipeline

Protein sequences are downloaded from UniProt and aligned to the genome in a splice aware manner using GenBlast [12]. The set of proteins aligned to the genome is a subset of UniProt proteins used to provide a broad, targeted coverage of the cat proteome. The set consists of the following:

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

Note: PE level = protein existence level

For the cat annotation, a cut-off of 50 percent coverage and identity and an e-value of e-1 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 are kept.

Protein group	Count
Cat SwissProt/TrEMBL PE 1, 2 & 3	14100
Human SwissProt/TrEMBL PE 1 & 2	48170
Mouse SwissProt/TrEMBL PE 1 & 2	67203
Other mammals SwissProt/TrEMBL PE 1 & 2	213549
Other vertebrates SwissProt/TrEMBL PE 1 & 2	155608

Table 1: Counts of transcript models built by GenBlast for each protein group

RNA-seq pipeline

RNA-seq data is downloaded from ENA (https://www.ebi.ac.uk/ena/) and used in the annotation. A merged file containing reads from all tissues/samples is created. The merged data is less likely to suffer from model fragmentation due to read depth. The available reads are aligned to the genome using BWA [13], with a tolerance of 50 percent mismatch to allow for intron identification via split read alignment. Initial models generated from the BWA alignments are further refined via exonerate. Protein coding models are identified via a BLAST alignment of the longest ORF against the UniProt vertebrate PE 1 & 2 data set.

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

Sample name	Count
Merged	30425
Testes	23507
Fetus (embryo)	20392
Retina	18763
Spleen	18718
Occipital lobe	18322
Kidney	17829
Cerebellum	17776
Spinal cord	17653
Ear tip	17632
Skin (orange colour)	17512
Ear cartilage	17505
Body (embryo)	17487
Skin	17260
Hippocampus	17214
Temporal lobe	17111
Skin (white colour)	16982
Lung	16896
Parietal lobe	16831
Head (embryo)	16563
Uterus	16074
Salivary gland	15787
Heart	15485
Liver	14776
Bone marrow	13952
Thymus	13934
Muscle	13296
Pancreas	12979

Table 2: Counts of RNA-Seq transcript models for each sample

Section 3: Filtering the Protein-Coding Models

The filtering phase decides the subset of protein-coding transcript models, generated from the model-building pipelines, that comprise the final protein-coding gene set. Models are filtered based on information such as what pipeline was used to generate them, 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 is used to define a hierarchy of input data sets, from most preferred to least preferred. The output of this pipeline includes 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 are also separate into clades, to help selection during the layering process. Each UniProt protein is in one clade only, for example mammal proteins are present in the mammal clade and are 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 primate is being annotated, 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 is 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) is that the intron coordinates from the model missing UTR exactly match a subset of the coordinates from the UTR donor model.

Generating multi-transcript genes

The above steps generate a large set of potential transcript models, many of which overlap one another. Redundant transcript models are collapsed and the remaining unique set of transcript models are 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 are annotated by looking for genes with evidence of frame-shifting or lying in repeat heavy regions. Single exon retrotransposed pseudogenes are 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 are added using annotations taken from RFAM [14] and miRBase [15]. Rfam and miRBase annotations were searched against the genomic sequence using NCBI-BLAST. The resulting alignments were then filtered using RNA-fold (miRBase hits) and Infernal [16] (Rfam hits).

Cross-referencing

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

Stable Identifiers

Stable identifiers are 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

Gene class	Count
Protein coding	20374
snRNA	1520
rRNA	846
snoRNA	652
Miscellaneous RNA	623
miRNA	482
Processed pseudogene	95
scaRNA	33
Pseudogene	26
Mitochondrial tRNA	22
Riobzyme	10
sRNA	3
Mitochondrial rRNA	2

Table 3: counts of the gene classes

Transcript class	Count
Protein coding	30565
snRNA	1520
rRNA	846
snoRNA	652
Miscellaneous RNA	623
miRNA	482
Processed pseudogene	95
scaRNA	33
Pseudogene	26
Mitochondrial tRNA	22
Riobzyme	10
sRNA	3
Mitochondrial rRNA	2

Table 4: counts of the transcript classes

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 be 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 estimates

    - 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 top level 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.

Assembly Information

Species name	Common name	Assembly name	Genbank accession ID	Assembly level
Felis catus	Cat	Felis_catus_8.0	GCA_000181335.3	Chromosome

Table 5: Assembly info

Layers in detail

Each group within a layer has two numbers attached to the end of it. The first number is the alignment coverage of the original evidence that was aligned. The second number is the percent identity between the original evidence that was aligned and the translation that was created. So for example, human_pe12_sp_95_80 can be read as: “A human protein with protein existence level one or two, that was present in SwissProt and that had a alignment coverage of greater than or equal to 95 percent when aligned to the Ensembl transcript and had a percent identity of greater than or equal to 80 percent when aligned to the Ensembl transcript”. Note that the alignment percent identity and coverage values are in bins, so if you see two groups such as human_pe12_sp_95_95 and human_pe12_sp_95_80, they are mutually exclusive (so the second group covers >= 80, but < 95 percent identity.

Most of the names refer to different alignments of UniProt proteins via GenBlast. The realign names refer to the projection models (if projection was carried out), while the

rnaseq_merged and rnaseq_tissue refer to models built from the merged RNA-seq track or the tissue RNA-seq tracks respectively (if RNA-seq data were aligned).

Below you can view the exact layering rules used to select the transcripts for each protein-coding gene. Note that regardless of whether certain evidence types were not used in the annotation (e.g. if a species does not have RNA-seq data or if no projection was carried out), they will still be listed in the layers below (as we use different layering templates based on which clade the species belongs to), but would be ignored by the layering code.

Layer 1

"realign_95_95","realign_95_80","rnaseq_merged_95_95","rnaseq_merged_95_80","rnaseq_merged_90_80","rnaseq_tissue_95_95","self_pe12_sp_95_95","self_pe12_tr_95_95","self_pe12_sp_95_80","self_pe12_tr_95_80"

Layer 2

"rnaseq_tissue_95_80","rnaseq_tissue_90_80","human_pe12_sp_95_95","human_pe12_tr_95_95","mouse_pe12_sp_95_95","mouse_pe12_tr_95_95","mammals_pe12_sp_95_95","mammals_pe12_tr_95_95"

Layer 3

"self_pe3_sp_95_95","self_pe3_tr_95_95","human_pe12_tr_95_80","human_pe12_sp_95_80","mouse_pe12_sp_95_80","mouse_pe12_tr_95_80"

Layer 4

"mammals_pe12_sp_95_80","mammals_pe12_tr_95_80","human_pe12_sp_90_80","human_pe12_tr_90_80","mouse_pe12_sp_90_80","mouse_pe12_tr_90_80","vert_pe12_sp_95_95","vert_pe12_tr_95_95"

Layer 5

"rnaseq_merged_80_60","mammals_pe12_sp_90_80","mammals_pe12_tr_90_80","vert_pe12_sp_90_80","vert_pe12_tr_90_80"

Layer 6

"human_pe12_sp_80_60","human_pe12_tr_80_60","mouse_pe12_sp_80_60","mouse_pe12_tr_80_60","rnaseq_merged_70_60","rnaseq_tissue_70_60","human_pe12_sp_70_60","mouse_pe12_sp_70_60"

Layer 7

"human_pe12_tr_70_60","mouse_pe12_tr_70_60","mammals_pe12_sp_80_60","mammals_pe12_tr_80_60","vert_pe12_sp_80_60","vert_pe12_tr_80_60","realign_70_60"

More information

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

- Publication

Aken B et al.: The Ensembl gene annotation system. Database 2016.

- Web

Link to Ensembl gene annotation documentation

References

1. Smit, A., R. Hubley, and P. Green,. RepeatMasker Open, 1996. **3**: p. 1996-2004.
2. Kuzio, J., R. Tatusov, and D. Lipman, *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): p. 1028-1040.
3. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic acids research, 1999. **27**(2): p. 573.
4. Down, T.A. and T.J. Hubbard, *Computational detection and location of transcription start sites in mammalian genomic DNA.* Genome research, 2002. **12**(3): p. 458-461.
5. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.* Nucleic acids research, 1997. **25**(5): p. 955-964.
6. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* Journal of molecular biology, 1997. **268**(1): p. 78-94.
7. Consortium, U., *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
8. Pontius, J.U., L. Wagner, and G.D. Schuler, *21. UniGene: A unified view of the transcriptome.* The NCBI Handbook. Bethesda, MD: National Library of Medicine (US), NCBI, 2003.
9. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
10. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic acids research, 2015. **44**(D1): p. D733-D745.
11. Slater, G.S.C. and E. Birney, *Automated generation of heuristics for biological sequence comparison.* BMC bioinformatics, 2005. **6**(1): p. 31.
12. She, R., et al., *genBlastG: using BLAST searches to build homologous gene models.* Bioinformatics, 2011. **27**(15): p. 2141-2143.
13. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
14. Griffiths-Jones, S., et al., *Rfam: an RNA family database.* Nucleic acids research, 2003. **31**(1): p. 439-441.
15. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic acids research, 2006. **34**(suppl_1): p. D140-D144.
16. Nawrocki, E.P. and S.R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches.* Bioinformatics, 2013. **29**(22): p. 2933-2935.