Ensembl Gene Annotation (*e!93*) Hagfish (Eptatretus burgeri)

Assembly: Eburgeri_3.2, GCA_900186335.2

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.

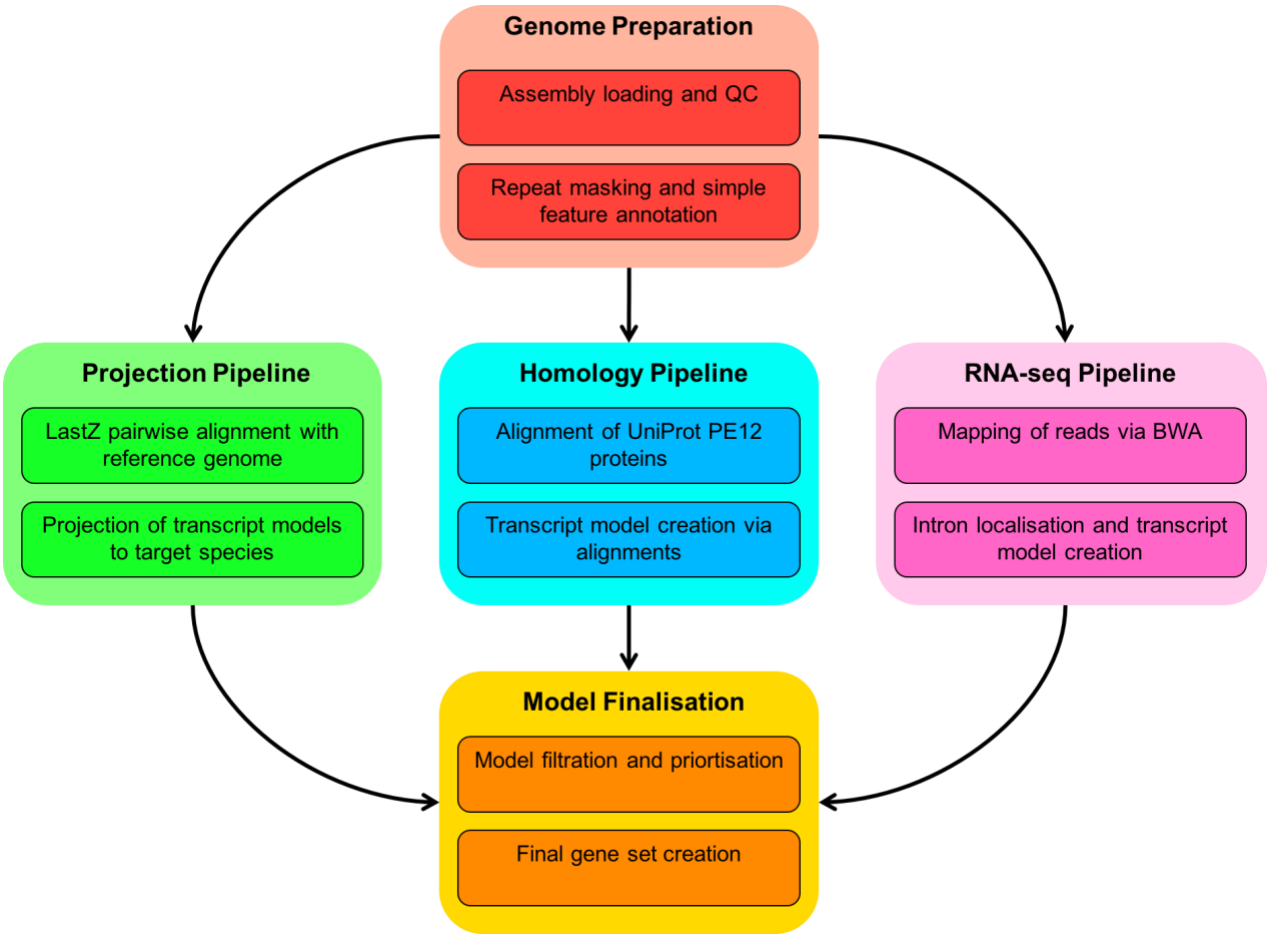

Fig. 1: Flowchart of the protein-coding annotation pipeline. Small ncRNAs, Ig genes, TR genes, and pseudogenes are computed using separate pipelines.

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 hagfish annotation, the Repbase vertebrates's library was used with RepeatMasker. This was supplemented by a custom repeat library derived via Repeatmodeler.

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.

Projection mapping pipeline

For all species a whole genome alignment is generated against a suitable reference assembly using LastZ [10]. Syntenic regions identified using this alignment are then used to map protein coding annotation from the most recent GENCODE [11] gene set. For the hagfish annotation, the human assembly, GRCh38, was used as a reference and The GENCODE 27 gene set was used to map protein coding annotation.
The mapped transcripts are 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 are passed into a realignment pipeline. Here they are re-aligned to the original sequence in the region they are mapped to. If possible, a model with canonical splicing is built otherwise the transcript model is discarded. Due to the large evolutionary distance the projection of genes from human did not yield useful results and was discarded.

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 hagfish proteome. The set consists of the following:

- Hagfish SwissProt/TrEMBL PE 1 & 2
- Human SwissProt/TrEMBL PE 1 & 2
- Mammals SwissProt/TrEMBL PE 1 & 2
- Fish SwissProt/TrEMBL PE 1 & 2

  Other vertebrates SwissProt/TrEMBL PE 1 & 2

Note: PE level = protein existence level

For the hagfish annotation, a cut-off of 30 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
Hagfish SwissProt/TrEMBL PE 1 & 2	674
Human SwissProt/TrEMBL PE 1 & 2	41612
Mammals SwissProt/TrEMBL PE1 & 2.	233498
Fish SwissProt/TrEMBL PE 1 & 2	97376
Other vertebrates SwissProt/TrEMBL PE 1 & 2	95768

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.

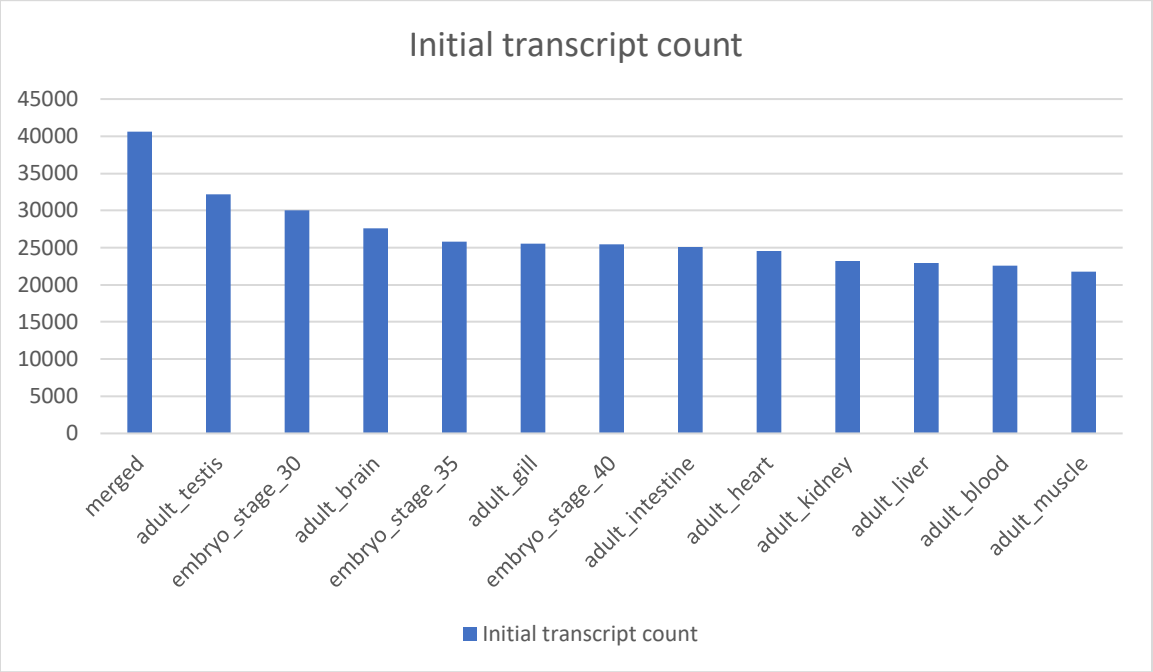

Fig 2: Counts of RNA-Seq transcript models for each sample

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 separated 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 prioritize based on the highest layer with available evidence. In general, the highest layers contain the set of evidence that is considered the most trustworthy, 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.

As hagfish is evolutionary distant to most well annotated species, it was difficult to see meaningful differences in terms of alignment values for the different groups. An alternative strategy was employed where all the protein alignments where relabeled as a single group (genblast) and then they were pre-filtered to select the model at each potential locus that contained the best balance in terms of the number of exons reported and the support among the underlying alignments for the observed introns.

This effectively split the models into two groups for layering, RNA-seq models and protein-to-genome alignments. This simplified approach helped select the most supported transcripts at each locus at the cost of missing potential alternative transcript structures. Given the relative difficulty in verifying hagfish open reading frames, this provided the best results based on manually inspecting the resulting gene set.

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.

Immunoglobulin and T-cell Receptor genes

Translations of different human IG gene segments are downloaded from the IMGT database [14] and aligned to the genome using GenBlast.

For the hagfish annotation, a cut-off of 80 percent coverage, 70 percent identity and an e-value of e-1 were used for GenBlast with the exon repair option turned on. The top 10 transcript models built by GenBlast for each protein passing the cut-offs are kept. In cases where multiple sequences aligned to the same locus, we selected the alignment with the highest combined percent identity and coverage.

No matching alignments were found, this was expected as hagfish is very distant and IG/TR gene segments are generally short and not well conserved.

Section 4: Creating the Final Gene Set

Small ncRNAs

Small structured non-coding genes are added using annotations taken from RFAM [15] and miRBase [16]. 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 [17] (Rfam hits).

lincRNAs

Candidate long intergenic non-coding RNAs (lincRNAs) should not overlap a protein-coding gene nor have a Pfam domain. The RNA-seq data sets, which were filtered against the protein-coding gene set, are used to predict lincRNAs and the Pfam analysis from InterProScan is run against the filtered gene set.
For the primate clade annotation, it was difficult to ascertain the validity of 2-exon models as lincRNA candidates so they were excluded from the set of potential lincRNAs.

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
lincRNA	446
miRNA	29
misc_RNA	7
protein_coding	16513
ribozyme	9
rRNA	89
snoRNA	15
snRNA	138

Table 2: counts of the gene classes

Transcript Class	Count
lincRNA	802
miRNA	29
misc_RNA	7
protein_coding	27960
Ribozyme	9
rRNA	89
snoRNA	15
snRNA	138

Table 3: 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
Eptatretus burgeri	Hagfish	Eburgeri_3.2	GCA_900186335.2	Contig

Table 4: Assembly info

Statistics of Interest

Species name	Common name	Total bases	Bases masked
Eptatretus burgeri	Hagfish	2608383542	836353990
%			32.06

Table 5: Number of bases repeat masked using the repbase vertebrate library combined with a custom hagfish repeatmodeler library

Species name	Common name	Eponine	CpG	tRNAscan
Eptatretus burgeri	Hagfish	126305	6332	6332

Table 6: Counts of low complexity features

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 an 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 28 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.

Note that for hagfish a specialised layering set was used that allowed for maximum sensitivity given its evolutionary distance from any well annotated reference

Layer 1

"rnaseq_merged_95_95","rnaseq_merged_95_80","rnaseq_merged_90_70","rnaseq_merged_90_60","rnaseq_tissue_90_70","rnaseq_tissue_90_60","rnaseq_merged_80_50","rnaseq_tissue_80_50","rnaseq_tissue_95_95","rnaseq_tissue_95_80"

Layer 2

"genblast_95_95","genblast_95_80","genblast_90_70","genblast_90_60"

Layer 3

"rnaseq_merged_70_40","rnaseq_tissue_70_40","genblast_80_50"

Layer 4

"genblast_70_30"

Layer 5

"genblast_60_20"

Layer 6

"rnaseq_merged_50_25","rnaseq_tissue_50_25"

Layer 7

"genblast_0_0","rnaseq_tissue_0_0","rnaseq_merged_0_0"

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, http://www.repeatmasker.org/. 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. Harris, R.S., *Improved pairwise alignment of genomic DNA*. 2007: The Pennsylvania State University.
11. Harrow, J., et al., GENCODE: the reference human genome annotation for The ENCODE Project. Genome research, 2012. **22**(9): p. 1760-1774.

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. Lefranc, M.-P., et al., *IMGT®, the international ImMunoGeneTics information system® 25 years on.* Nucleic acids research, 2014. **43**(D1): p. D413-D422.

15. Griffiths-Jones, S., et al., *Rfam: an RNA family database.* Nucleic acids research, 2003. **31**(1): p. 439-441.

16. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic acids research, 2006. **34**(suppl_1): p. D140-D144.

17. Nawrocki, E.P. and S.R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches.* Bioinformatics, 2013. **29**(22): p. 2933-2935.