Ensembl Gene Annotation (e!109)

Horse (*Equus caballus*)

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.

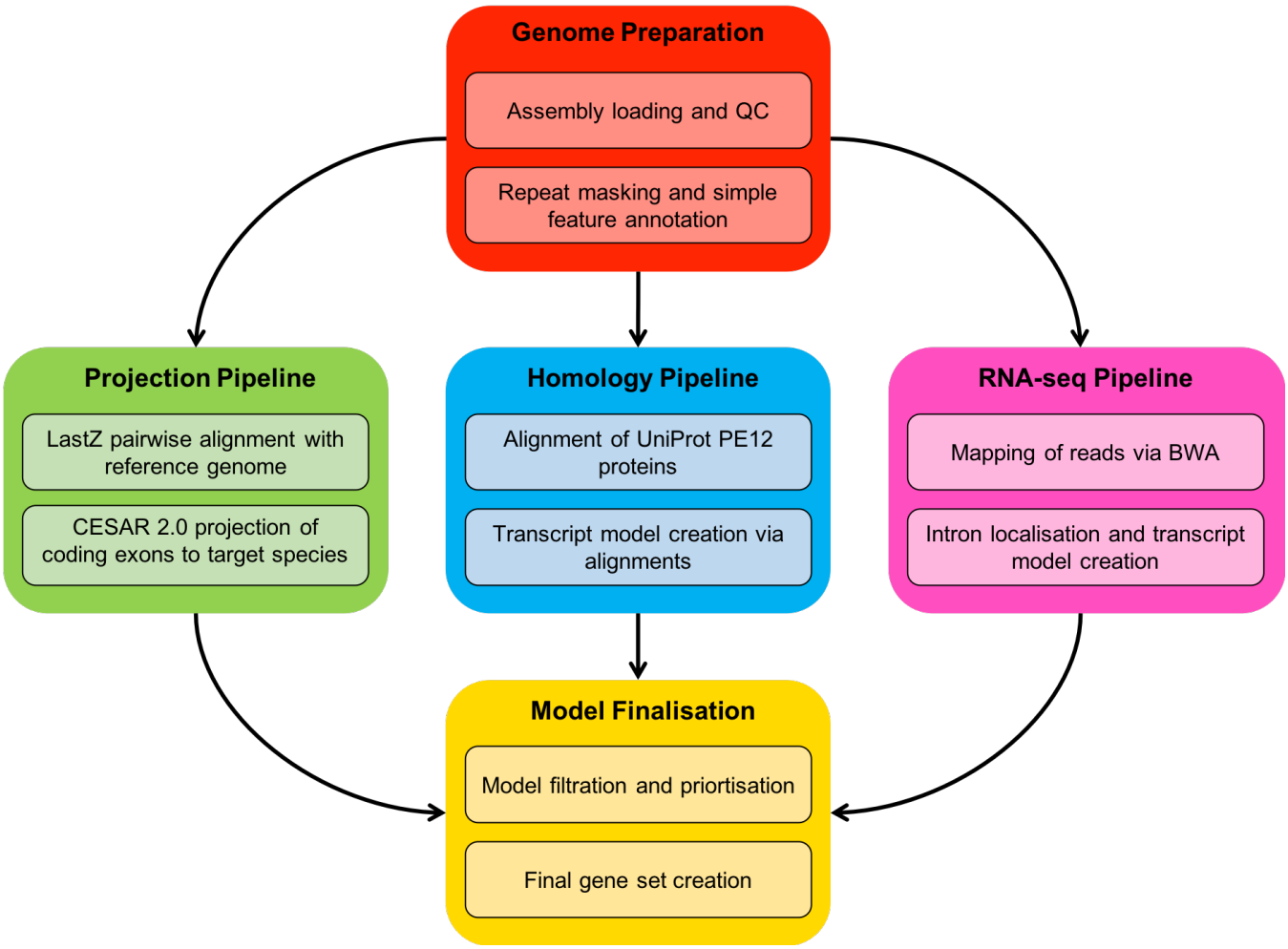

Figure 1: Flowchart of the protein-coding annotation pipeline. Small ncRNAs, Ig genes, TR genes, and pseudogenes are computed using separate pipelines.

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, *-nolow -engine "crossmatch"*), dustmasker [2] and TRF [3].

The Repbase mammals library was used with RepeatMasker. In addition to the Repbase [4] library, where available, a custom repeat library was used with RepeatMasker. This custom library was created using RepeatModeler [5].

Low complexity features, ab initio predictions

Transcription start sites are predicted using Eponine–scan [6]. CpG islands longer than 400 bases and tRNAs are also predicted. The results of Eponine-scan, CpG[7], and tRNAscan [8] are for display purposes only; they are not used in the gene annotation process.

Genscan [9] is run across repeat-masked sequence to identify ab initio gene predictions. 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.

Species specific cDNA and protein alignments

cDNAs are downloaded from ENA [12] and RefSeq [13], and aligned to the genome using Exonerate [14]. Only known mRNAs are used (NMs). The cDNAs can be used to add UTR to the protein coding transcript models if they have a matching set of introns.

Proteins are downloaded from UniProt and filtered based on protein existence (PE) at protein level and transcript level. The proteins are aligned to the genome using PMATCH to reduce the search space, then with genewise, which is a splice-aware aligner, to generate spliced models.

Projection mapping pipeline

For all species we generated a whole genome alignment against a suitable reference assembly using LastZ [15]. A suitable reference assembly would be closely related species with high quality annotation or an Ensembl/GENCODE species, i.e., human or mouse. Syntenic regions identified using this alignment are then used to map protein-coding annotation from the most recent released gene set. We have integrated CESAR 2.0 [16] into our pipeline to complement the results of our RNA-seq, cDNA and protein alignment pipelines. CESAR 2.0 is a fast method of mapping exons and genes that can deal with splice sites that have shifted significantly. We found that CESAR 2.0 performs significantly better than our previous mapping code at longer distances.

We used the human assembly GRCh38.p13 and the Ensembl release 105 gene set as reference to map protein-coding annotation.

Protein-to-genome pipeline

Protein sequences are downloaded from UniProt and aligned to the genome in a splice aware manner using GenBlast [17]. The set of proteins aligned to the genome is a subset of UniProt [10] proteins used to provide a broad, targeted coverage of the horse proteome. The set consists of the following:

- Self SwissProt/TrEMBL PE 1 & 2
- Mouse SwissProt/TrEMBL PE 1 & 2
- Human SwissProt/TrEMBL PE 1 & 2
- Other mammals SwissProt/TrEMBL PE 1 & 2

Note: PE = protein existence level

A cut-off of 50 percent coverage and 30 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.

RNA-seq pipeline

Where available, RNA-seq data is downloaded from 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 [18], 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. A second set of models to complement the first set was produced with Scallop [26] based on read alignment from STAR [25]. 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.

Long-read transcriptomic data pipeline

Where available, long-read transcriptomic data (i.e., IsoSeq or Nanopore) is downloaded from ENA (https://www.ebi.ac.uk/ena/) and used in the annotation. The long-read data is mapped to the genome using Minimap2 [24] with the recommended settings for Iso-Seq and Nanopore data. Due to the high error rate of the Nanopore data, post mapping error correction is applied to maximize the number of usable mappings. Low frequency intron/exon boundaries that are non-canonical are replaced with high frequency boundary coordinates within 50bp. In addition, low frequency potential gaps between adjoining exons are filled in based on high frequency observations of single exons with the same terminal boundary coordinates.

Immunoglobulin and T-cell receptor genes

Translations of different human IG gene segments are downloaded from the IMGT database [19] and aligned to the genome using GenBlast.

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.

Selenocysteine proteins

We aligned known selenocysteine proteins against the genome using Exonerate. Then we checked that the generated model had a selenocysteine in the same positions as the known protein. We only kept models with at least 90% coverage and 95% identity.

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 fish 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.

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 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 retro-transposed 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 [20] and miRbase [21]. NCBI-BLAST [11] was run for these sequences and models built using the Infernal [22] software suite.

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

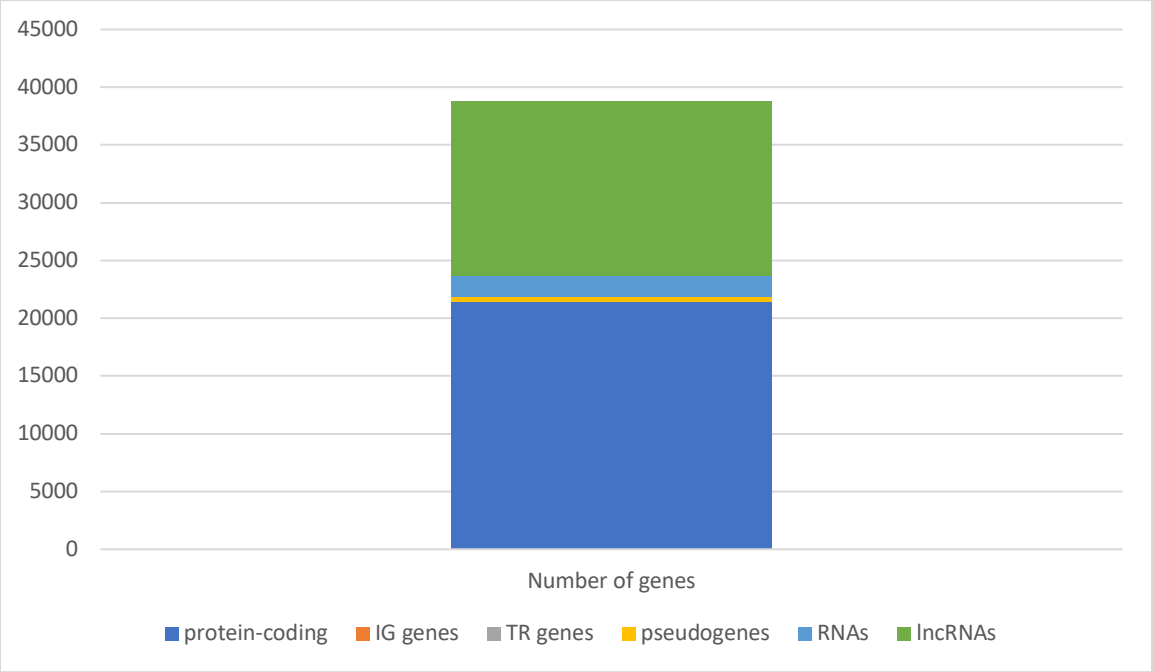

Figure 2: Counts of the major gene classes in horse.

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 of 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.

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

Aken B et al.: The Ensembl gene annotation system. Database 2016.[23]

http://www.ensembl.org/info/genome/genebuild/genome_annotation.html

Assembly information

Species	Common name	Assembly name	INSDC Accession	Release date
Equus caballus	Horse	EquCab3.0	GCA_002863925.1	10/12/2022

Table 1: Assembly information

Statistics of interest

Assembly name	Repbase	RepeatModeler
EquCab3.0	44.63%	36.37%

Table 2: Percentage of the genome masked by RepeatMasker using different repeat libraries.

References

1. Smit, A.F.A., R. Hubley, and P. Green. RepeatMasker Open-4.0. 2013-2015; Available from: http://www.repeatmasker.org.
2. Morgulis, A., et al., A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol, 2006. 13(5): p. 1028-40.
3. Benson, G., Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 1999. 27(2): p. 573-80.
4. Bao, W., K.K. Kojima, and O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA, 2015. 6: p. 11.
5. Smit, A.F.A. and R. Hubley. RepeatModeler Open-1.0. 2008-2015 Available from: http://www.repeatmasker.org.
6. Down, T.A. and T.J. Hubbard, Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res, 2002. 12(3): p. 458-61.
7. Larsen, F., et al., CpG islands as gene markers in the human genome. Genomics, 1992. 13(4): p. 1095-107.
8. Lowe, T.M. and S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997. 25(5): p. 955-64.
9. Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997. 268(1): p. 78-94.
10. UniProt Consortium, T., UniProt: the universal protein knowledgebase. Nucleic Acids Res, 2018. 46(5): p. 2699.
11. Altschul, S.F., et al., Basic local alignment search tool. J Mol Biol, 1990. 215(3): p. 403-10.
12. ENA, www.ebi.ac.uk/ena.
13. O'Leary, N.A., et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 2016. 44(D1): p. D733-45.
14. Slater, G.S. and E. Birney, Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 2005. 6: p. 31.

15. Harris, R.S., Improved pairwise alignment of genomic DNA. 2007.
16. Sharma, V., P. Schwede, and M. Hiller, CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. Bioinformatics, 2017. 33(24): p. 3985-3987.
17. She, R., et al., genBlastG: using BLAST searches to build homologous gene models. Bioinformatics, 2011. 27(15): p. 2141-3.
18. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009. 25(14): p. 1754-60.
19. Lefranc, M.P., et al., IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res, 2015. 43(Database issue): p. D413-22.
20. Nawrocki, E.P., et al., Rfam 12.0: updates to the RNA families database. Nucleic Acids Res, 2015. 43(Database issue): p. D130-7.
21. Griffiths-Jones, S., et al., miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res, 2006. 34(Database issue): p. D140-4.
22. Nawrocki, E.P. and S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013. 29(22): p. 2933-5.
23. Aken, B.L., et al., The Ensembl gene annotation system. Database (Oxford), 2016. 2016.
24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018. 34:3094-3100.
25. Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics, Volume 29, Issue 1, January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635
26. Shao, M., Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol 35, 1167–1169 (2017). https://doi.org/10.1038/nbt.4020