Ensembl gene annotation project (*e!86*)

Gallus gallus, Gallus_gallus-5.0

This document describes the annotation process of the high-coverage chicken Gallus_gallus-5.0 assembly, described in Figure 1. The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.

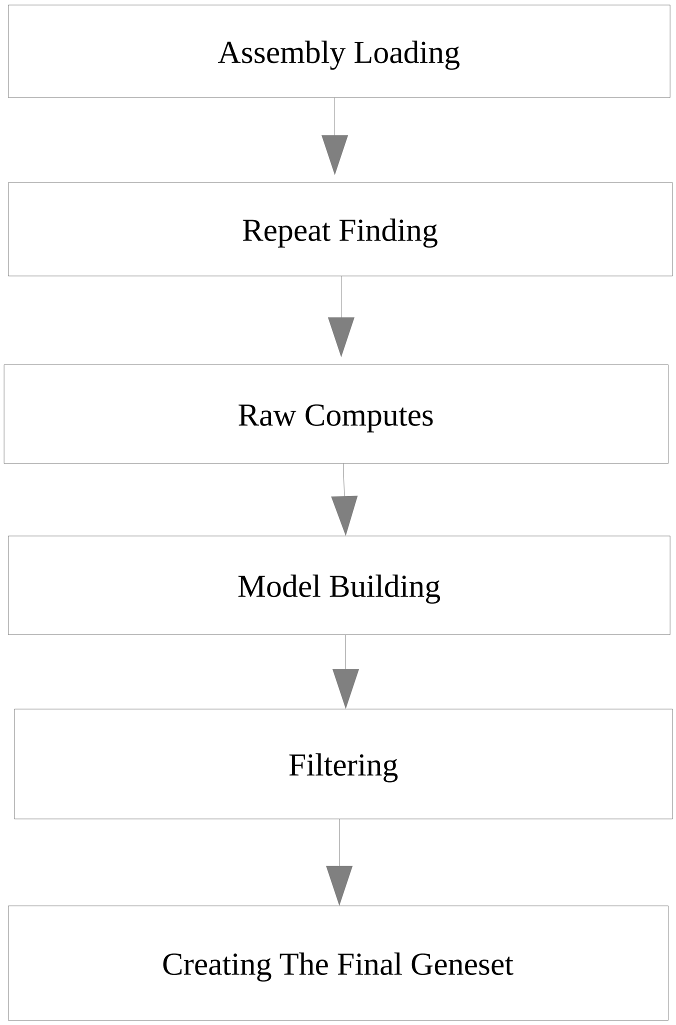

Figure 1: The Gene Annotation Pipeline

Repeat Finding

After loading into a database the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 4.0.5 with parameters '-nolow -species "aves"', using 'wublast' as the search engine), Dust [3] and TRF [4]. Both executions of RepeatMasker and Dust combined masked 17.9% of the assembly.

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 8988657 UniProt, 12337932 UniGene and 11555946 Vertebrate RNA sequences aligning to the genome.

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 1.

Pipeline	Source	Number of Models
Species specific cDNAs	RefSeq	
PacBio IsoSeq	Roslin Institute	
RNA-seq	Roslin Institute	147065
Protein-to-genome	Subset of UniProt vertebrate proteins	916778

Table 1: Gene Model Generation Overview

cDNA Alignments

Chicken 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.

Species	Initial mRNA sequences	Sequences aligned
Chicken	6322	6267

Table 2: Species specific cDNAs aligned against Gallus_gallus-5.0

PacBio IsoSeqs

PacBio IsoSeqs are transcriptomic long reads sequenced at a high coverage to allow correction of the technology. We downloaded the consensus sequences from ENA representing two tissue types, embryo (PRJEB13248) and brain (PRJEB13246). The sequences were aligned to the genome using Exonerate using a cut-off of 95% identity and 90% coverage.

Both sets had 3' capping and were used for adding UTRs to homology-based protein-coding models. Both sets were used as lincRNA candidate for our lincRNA prediction pipeline. The embryo set also had 5' capping and was used for generating protein-coding models.

Tissue sample	Initial IsoSeq sequences	Sequences aligned
Embryo	14909	12928
Brain	211289	189172

Table 3: PacBio Isoseq sequences aligned against Gallus_gallus-5.0

Protein-to-genome Pipeline: Generating coding models using UniProt proteins

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 chicken genome. The set consists of the following:

- Chicken PE level 1, 2, 3, 4
- Other birds PE level 1, 2, 3
- Human PE level 1, 2, 3
- Other mammals PE level 1, 2, 3
- Other vertebrates PE level 1, 2, 3

Note: PE level = protein existence level

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. This process produced 916778 transcript models in total.

RNA-seq Pipeline

RNA-seq data downloaded from ENA, PRJEB12891, was used in the annotation. This consisted of paired end data from twenty-one tissue samples: breast muscle, bursa, caecal tonsil, cerebellum, duodenu, gizzard fat, harderian gland, heart muscle, ileum, kidney, left optic lobe, liver, lung, ovary, pancreas, proventriculus, skin, spleen, thymus, thyroid, trachea. A merged file contain reads from all tissues 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. The Ensembl RNA-seq 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 186126 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 and considered as potential lincRNAs.

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. 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 precent identity to the original data are.

Models were filtered using the LayerAnnotation and GeneBuilder modules. The 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.

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.

Layer 1:

- Chicken cDNAs models with >= 90 percent coverage and 97% identity
- Chicken embryo IsoSeq models with >= 90% coverage and 95% identity
- RNA-seq models with >= 80 percent coverage and identity
- Chicken UniProt proteins from PE levels 1 & 2 with >= 80 percent coverage and identity

Layer 2:

- Chicken UniProt proteins from PE levels 3 with >= 80 percent coverage and identity

- RNA-seq models with >= 50 percent coverage and identity
- Human UniProt proteins from PE levels 1 & 2 with >= 80 percent coverage and identity

Layer 3:

- RNA-seq models with >= 0 percent coverage and identity
- Birds UniProt proteins from PE levels 1 & 2 with >= 95 percent coverage and identity
- Human UniProt proteins from PE levels 3 with >= 80 percent coverage and identity
- Vertebrates UniProt proteins from PE levels 1 & 2 with >= 80 percent coverage and identity

Layer 4:

- Vertebrate UniProt proteins from PE level 3 with >= 80 percent coverage and identity

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNA-seq and cDNA and IsoSeqs sequences. The source of the UTRs was prioritised with UTR coming from cDNAs and embryo IsoSeqs, then brain IsoSeqs and finally RNA-seq.

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 16399 genes with 25734 transcripts.

Pseudogenes

The Pseudogene module was run to identify pseudogenes from within the set of gene models. A total of 50 genes were labelled as pseudogenes or processed pseudogenes.

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 the Infernal software suite [20].

lincRNAs discovery

Using the transcriptomic data set, we try to predict long intergenic non coding RNAs (lincRNAs). We used the RNA-seq data and the two IsoSeq 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.

As chicken has been previously released in Ensembl a comparison was made to the previous gene set and as many stable identifiers as possible were mapped between the two annotations.

Final Gene Set Summary

The final gene set consists of 16362 protein coding genes, including 13 mitochondrial genes. These contain 25697 transcripts. A total of 50 pseudogenes were identified. 1849 small ncRNAs were added by the small ncRNA pipeline and 7841 lincRNA were added by the lincRNA pipeline.

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:

- 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 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 Bioinformatic*s 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]

21 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]