

Ensembl gene annotation project

Sus scrofa (Pig)

Raw Computes Stage: Searching for sequence patterns, aligning proteins and cDNAs to the genome.

The annotation process of the Sscrofa10.2 assembly began with the “raw compute” stage [Figure 1] whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.2.8 with parameters '-nolow -s -species pig'), Dust [2] and TRF [3]. RepeatMasker masked 36.3% of the genome. Adding low complexity masking (including gaps) with Dust brings the total masked to 48.2%.

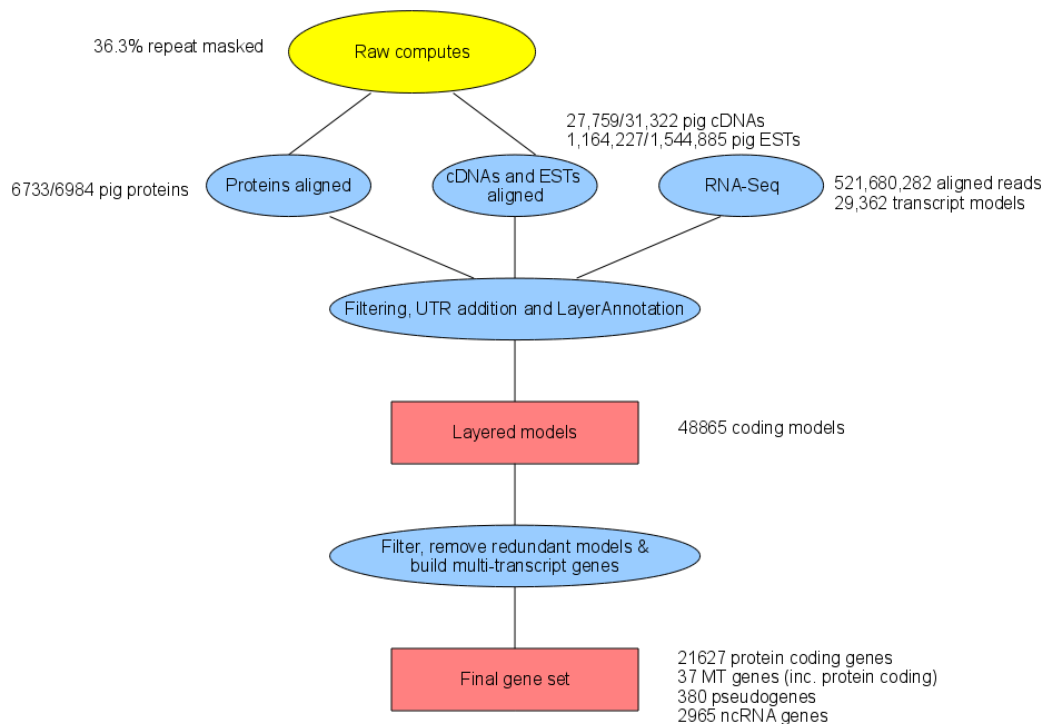


Figure 1: Summary of pig gene annotation project.

Transcription start sites were predicted using Eponine-scan [4], and FirstEF [5] CpG islands and tRNAs [6] were also predicted.

Genscan [7] was run across RepeatMasked sequence and the results were used as input for UniProt [8], UniGene [9], and Vertebrate RNA [10] alignments by WU-BLAST [11] (passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required). This resulted in 287,736 UniProt vertebrate and 72,840 UniProt non-vertebrate sequences, 340,390 UniGene and 327,557 Vertebrate RNA sequences aligning to the genome.

Targeted Stage: Generating coding models from pig evidence

Pig protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8] and Genbank) and filtered to remove sequences based on predictions. The pig sequences were mapped to the genome using Pmatch as indicated in Figure 2. For pig cDNA sequences with annotated CDS, Exonerate's [12] cdna2genome model was used.

Models of the coding sequence (CDS) were produced from the proteins using Genewise [13]. Two sets of models were produced, one with only consensus splice sites and one where non-consensus splices were allowed.

Where a single protein sequence generated two different coding models at the same locus, the BestTargetted module was used to select the coding model that most closely matched the source protein to take through to the next stage of the gene annotation process. The generation of transcript models using pig-specific data is referred to as the "Targeted stage". This stage resulted in 7,934 coding models.

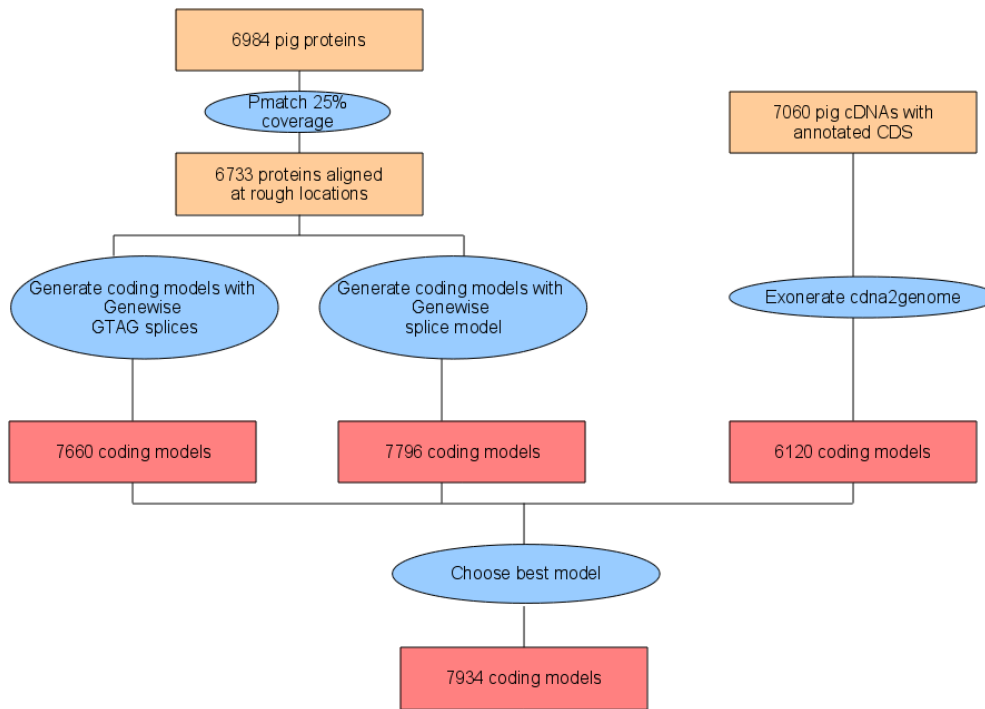


Figure 2: Targeted stage using pig specific proteins and cDNAs with annotated CDS

cDNA, EST and ENSSCP alignment

Pig cDNAs were downloaded from Genbank, clipped to remove polyA tails, and aligned to the genome using Exonerate. Of these, 27,759 of 31,322 pig cDNAs aligned with a cut-off of 50% coverage and 50% identity.

ESTs were downloaded from the same sources and processed in the same way. Of these, 1,164,227 of 1,544,885 pig ESTs aligned with a cut-off of 80% coverage and 90% identity.

ENSSSCP models from Ensembl 64 were also aligned to the genome. 17,493 canonical translations were downloaded and aligned using Exonerate. Of these 16,399 aligned with identity greater than 95% and coverage greater than 70%.

Similarity Stage: Generating additional coding models using proteins from related species

UniProt alignments from the Raw Compute step were filtered to proteins classed by UniProt's Protein Existence (PE) classification as level 1 or 2. The proteins were also divided taxonomically into the following groups: mammalian, non-mammalian vertebrates and invertebrates. WU-BLAST was rerun for these sequences and the results were passed to Genewise to build coding models. Only models from the mammalian and non-mammalian vertebrate groups were used. The generation of transcript models using data from related species is referred to as the "Similarity stage". This stage resulted in 95,917 coding models.

Filtering Coding Models

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus. RNA-Seq spliced alignments supporting introns, models from the targeted stage, RNA-Seq models, pig cDNA and pig EST alignments were used to help filter the set. 47,326 models were rejected as a result of filtering, leaving 48,550 models used in subsequent stages. The Apollo software [15] was used to visualise the results of filtering.

RNA-Seq models

RNA-Seq data provided by the Swine Genome Sequencing Consortium (SGSC) was used in the annotation. This comprised a mixture of single and paired end data from samples including: a pool of 10 tissues, alveolar macrophages, male gonad, whole blood, placenta and testis. The available reads were aligned to the genome using BWA, resulting in 521,680,282 reads aligning. Subsequently, the Ensembl RNA-Seq pipeline was used to process the BWA alignments and create a further 7,800,861 split read alignments using Exonerate. The split reads and the processed BWA alignments were combined to produce 29,362 transcript models in total. The predicted open reading frames were compared to Uniprot Protein Existence (PE) classification level 1 and 2 proteins using WU-BLAST.

Models with no BLAST alignment or poorly scoring BLAST alignments were split into a separate class.

UTR addition

UTR was added to models generated in the targeted and similarity steps using the UTR addition modules and evidence from the pig cDNA models and all RNA-Seq models.

Layering of evidence

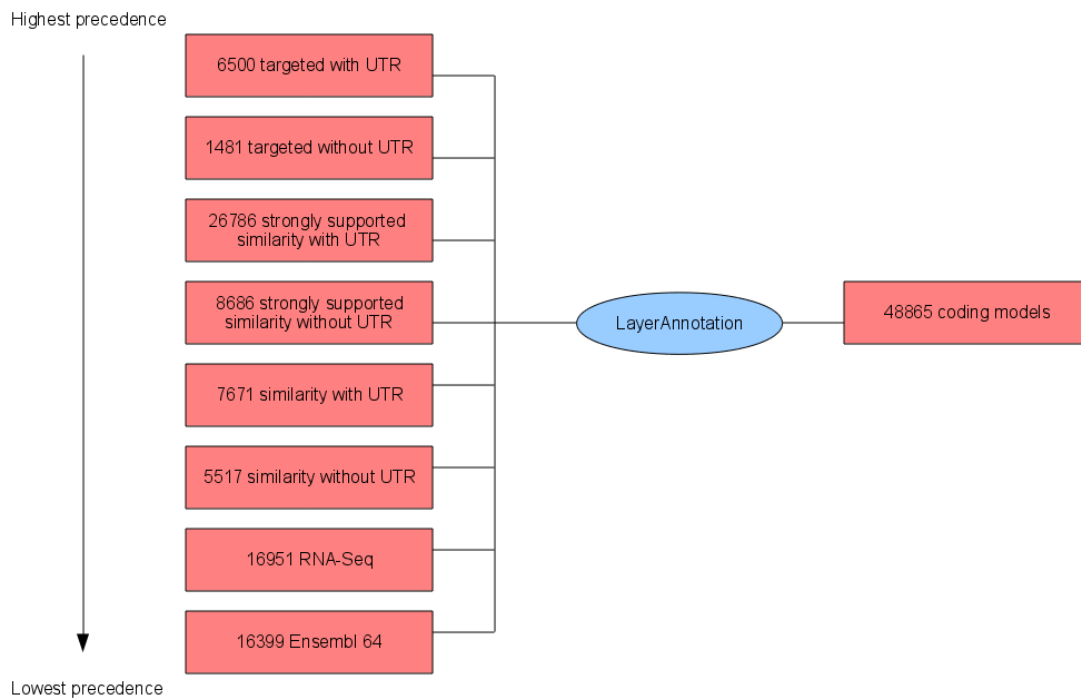


Figure 3: Combining evidence in LayerAnnotation

To combine models from different sources the LayerAnnotation module was used [Figure 3]. This takes models from lower layers only where there are no models

in a layer with higher priority. The layers, in order of precedence were: targeted with UTR, targeted without UTR, strongly supported similarity with UTR, strongly supported similarity without UTR, similarity with UTR, similarity without UTR, RNA-Seq with good BLAST scores to UniProt, and pig models from Ensembl 64, which were filtered to remove those based on projections in e64 and those with non-consensus splicing in e64.

This led to a set of transcript models containing 7,655 from the targeted step, 36,119 from the similarity step, 3,630 from RNA-Seq evidence and 480 from the e64 ENSSSCP set.

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 removed 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. The final gene set of genes included 3,807 genes built using evidence from the targeted stage, a further 14,118 genes built using proteins from other species and 3,637 genes built only from RNA-Seq evidence.

Pseudogenes, protein annotation, non-coding genes, cross referencing, stable identifiers

The gene set was screened for potential pseudogenes. Before public release the transcripts and translations were given external references, or cross references, to external databases, while translations were searched for domains/signatures of interest and labeled where appropriate. 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 the stable identifiers are propagated based on comparison of the new gene set to the previous gene set.)

Small structured non-coding genes were added using annotations taken from

RFAM [16] and miRBase [17].

The final gene set consists of 21,640 protein coding genes, including mitochondrial genes, these contain 23,118 transcripts. A total of 380 pseudogenes were identified and 2,965 ncRNAs. Of the protein coding transcripts 3,596 were made from RNASeq only, 15,072 came from proteins from other species. 3959 transcripts were pig specific, 37 transcripts were mitochondrial.



Figure 4: Composition of pig gene set.

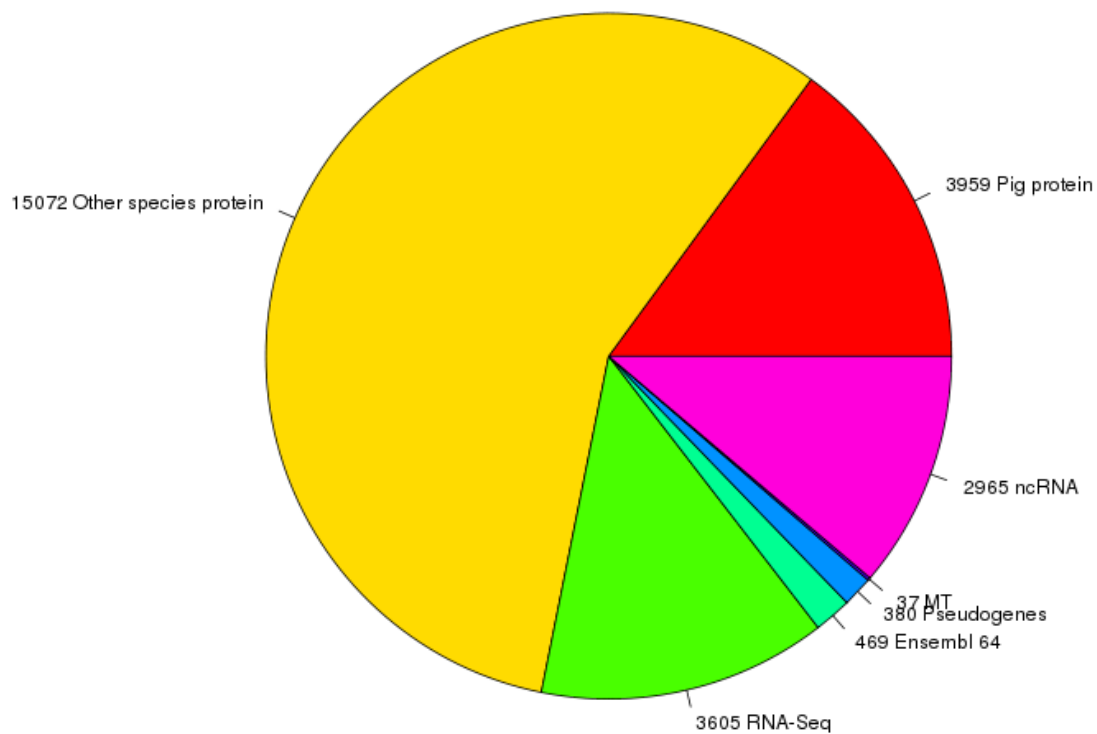


Figure 5: Composition of pig transcripts

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 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 indicate a more complete genome assembly.

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

- Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. The Ensembl automatic gene annotation system. *Genome Res.* 2004, 14(5):942-50. [PMID: 15123590]
- 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/docs/genebuild/genome_annotation.html
- * http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembl/doc/pipeline_docs/the_genebuild_process.txt?root=ensembl&view=co

References

1. Smit, AFA, Hubley, R & Green, P: RepeatMasker Open-3.0. 1996-2010.
www.repeatmasker.org
2. 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.
3. 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>
4. 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]
5. 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]
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, Mizrahi 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, Bayraktaroglu 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. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S.R. Eddy: Rfam: an RNA family database. *Nucleic Acids Research* (2003) 31(1):p439-441.
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
18. L. G. Wilming, J. G. R. Gilbert, K. Howe, S. Trevanion, T. Hubbard and J. L. Harrow: The vertebrate genome annotation (Vega) database. *Nucleic Acid Res.* 2008 Jan; Advance Access published on November 14, 2007; doi:10.1093/nar/gkm987