Ensembl gene annotation project (*e76*)
Papio anubis

(Olive baboon)

This document describes the annotation process of the high-coverage olive baboon assembly, described in Figure 1. The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.
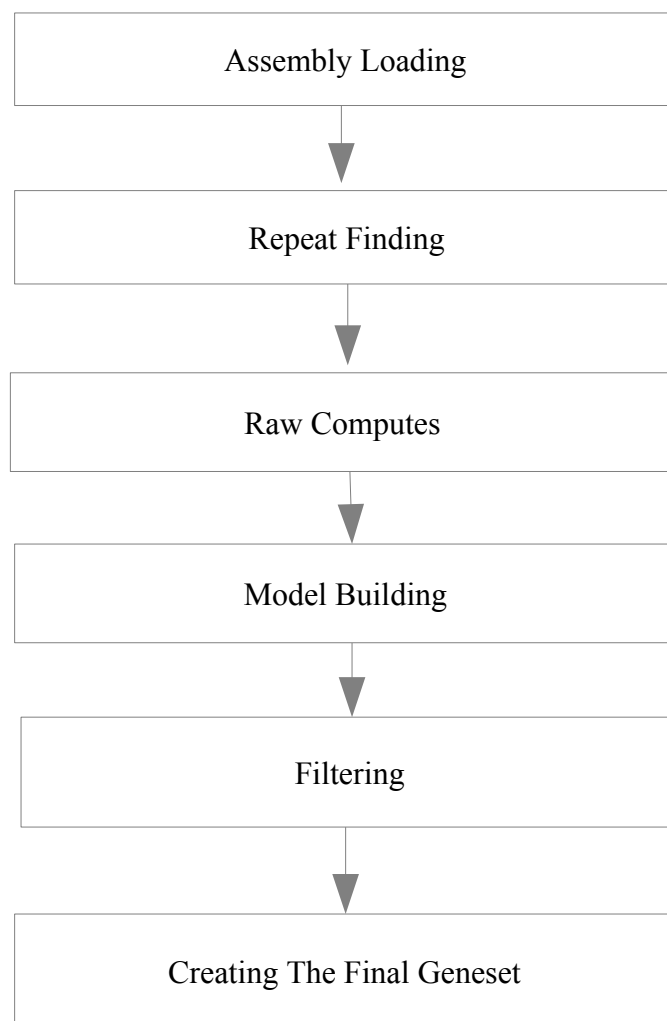
```
┌─────────────────────────────────┐
│        Assembly Loading         │
└─────────────────────────────────┘
                 │
                 ▼
┌─────────────────────────────────┐
│         Repeat Finding          │
└─────────────────────────────────┘
                 │
                 ▼
┌─────────────────────────────────┐
│          Raw Computes           │
└─────────────────────────────────┘
                 │
                 ▼
┌─────────────────────────────────┐
│         Model Building          │
└─────────────────────────────────┘
                 │
                 ▼
┌─────────────────────────────────┐
│            Filtering            │
└─────────────────────────────────┘
                 │
                 ▼
┌─────────────────────────────────┐
│    Creating The Final Geneset   │
└─────────────────────────────────┘
```

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 3.2.8 with parameters '-nolow -species "papio_anubis" -s'), Dust [3] and TRF [4]. Both executions of RepeatMasker and Dust combined masked 50.09% of the species genome.

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 438658 UniProt, 269266 UniGene and 345033 Vertebrate RNA sequences aligning to the genome.

cDNA and EST Alignments

Olive baboon cDNAs and ESTs were downloaded from ENA/Genbank, clipped to remove polyA tails, and aligned to the genome using Exonerate. These alignments provide supporting evidence for models.

Species	cDNA/EST	Sequences Downloaded	Sequences Aligned
Olive baboon	cDNA	456	4043
	EST	145582	424016

Table 1: cDNA/EST alignments

EST alignments were at a cut-off of 80% coverage and 80% identity. All cDNA alignments were used.

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 2.

Pipeline	Source	Number of Models
Targeted	96 UniProt olive baboon proteins 462 RefSeq olive baboon proteins	642
Similarity	438658 UniProt proteins	359234
RNASeq	Baylor College of Medicine, Nonhuman Primate Reference Transcriptome Resource	169644
Ensembl Longest Translations	20826 Ensembl Release 70 proteins for human	20095

Table 2: Gene Model Generation Overview

Targeted Pipeline: Generating coding models using species specific proteins

Protein sequences for olive baboon were downloaded from public databases (UniProt SwissProt/TrEMBL [9] and RefSeq [10]). The olive baboon protein sequences were mapped to the genome using Pmatch set at a low threshold (-T 14). Two sets of coding models were then produced from the proteins using Exonerate [13] and Genewise [14].

Where one protein sequence had generated more than one coding model at a locus, the BestTargeted 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. This pipeline is shown in Figure 2.

Similarity Pipeline: Generating coding models using proteins from related species

Coding models were generated using data from related species. WU-BLAST was rerun for the UniProt alignments from the Raw Computes step and the results were passed to Genewise [14] to build coding models.
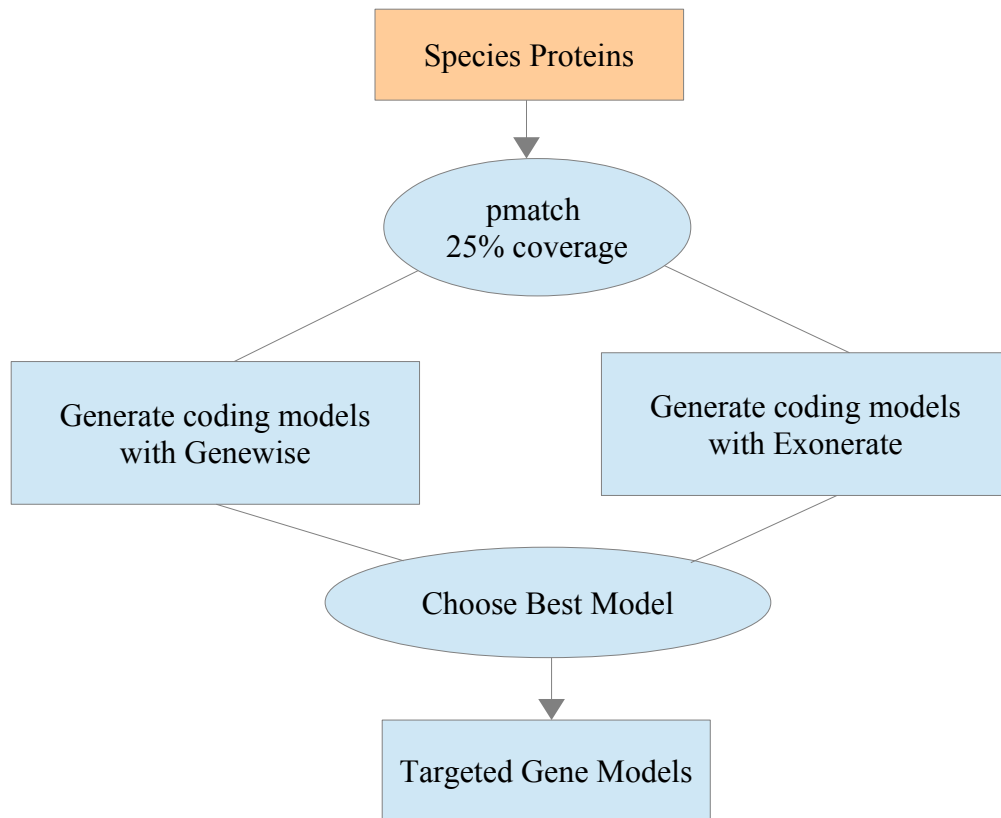
Figure 2: Targeted Pipeline

RNASeq Pipeline

The tissue-specific RNASeq data were provided through a collaboration between the Nonhuman Primate Reference Transcriptome Resource and the Human Genome Sequencing Center, Baylor College of Medicine. These data comprised a mixture of single and paired end data from samples including: a pool of 14 tissues, bone marrow, cerebellum, colon, frontal cortex, heart, kidney, liver, lung, lymph node, pituitary, skeletal muscle, spleen, temporal lobe and thymus,. The available reads were aligned to the genome using BWA. The Ensembl RNASeq 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 169644 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.

Ensembl Longest Translations

The longest translation for each protein coding gene in Ensembl release 76

for human were downloaded. These proteins were aligned against the olive baboon genome using Exonerate [13] to produce a set of coding models.

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.

Models were filtered using the TranscriptConsensus, LayerAnnotation and GeneBuilder modules.

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. Highly supported RNASeq models were ones where a Uniprot PE12 sequence could be aligned using BLAST with a hit coverage and percentage identity cut off of at 80 percent. Partially supported RNASeq models had a cut off of 50-80 percent coverage and identity. 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. A basic overview of the hierarchy used is as follows:

- Baboon-specific proteins models, highly supported RNASeq models and SwissProt PE12 primate models were given the highest priority.

- Next highest were mammal SwissProt PE12 models.

- After these vertebrate SwissProt PE12 models and partially supported RNASeq models were used.

- Below these primate/mammal/vertebrate TrEMBL PE12 similarity models were selected.

- Next were fragmented primate/mammal/vertebrate PE12 models from both SwissProt and TrEMBL.

- Last and with lowest priority were human longest translation models.

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNASeq, cDNA and EST sequences. At the UTR addition stage 48193 gene models out of 66732 non-RNASeq pipeline generated gene models had UTR added.

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 19922 genes with 21693 transcripts.

Pseudogenes

The Pseudogene module was run to identify processed pseudogenes from within the set of gene models – these were labelled as pseudogenes. A total of 720 genes were labelled as pseudogenes.

Creating The Final Gene Set

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].

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.

Final Gene Set Summary

The final gene set consists of 19210 protein coding genes, including 13 mitochondrial protein coding genes. These contain 20977 transcripts. A total of 720 pseudogenes were identified. 7579 ncRNAs were added by the ncRNA pipeline.
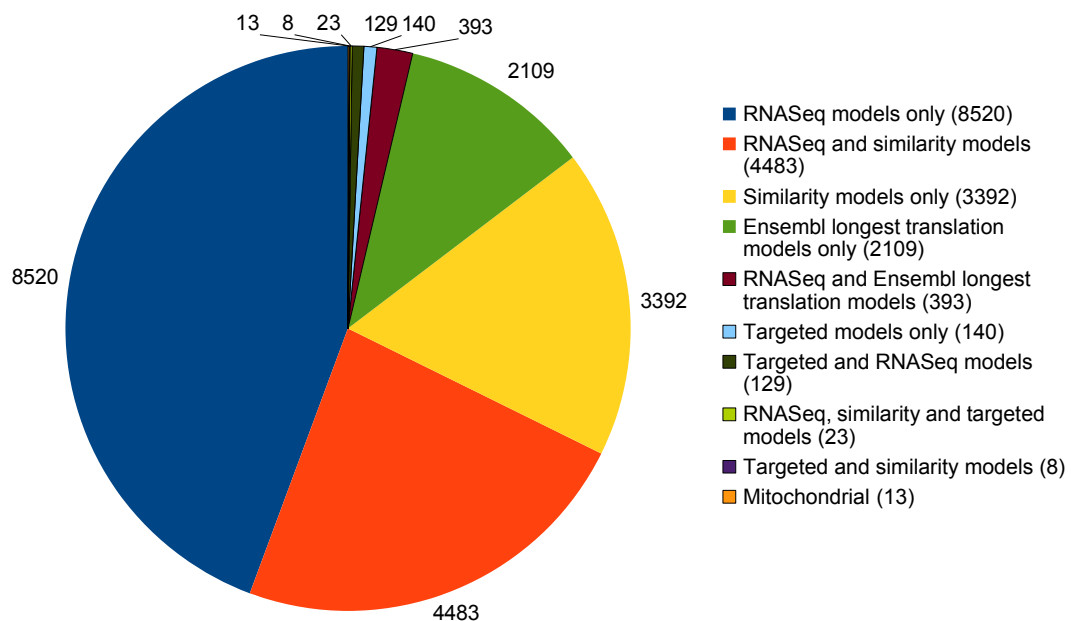

Figure 3: Supporting evidence for the protein coding gene models

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 estimate

    o A higher coverage usually indicates a more complete assembly.

    o Using Sanger sequencing only, a coverage of at least 2x is preferred.

2. N50 of contigs and scaffolds

    o A longer N50 usually indicates a more complete genome assembly.

    o 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

    o A lower number toplevel sequences usually indicates a more complete genome assembly.

4. Alignment of cDNAs and ESTs to the genome

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 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]