Ensembl gene annotation project (*e!74*)
Dasypus novemcinctus (Armadillo)

Konstantinos Billis

This document describes the annotation process of the Armadillo draft assembly (GCA_000208655.2), described in Figure 1.

Assembly Loading

The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.

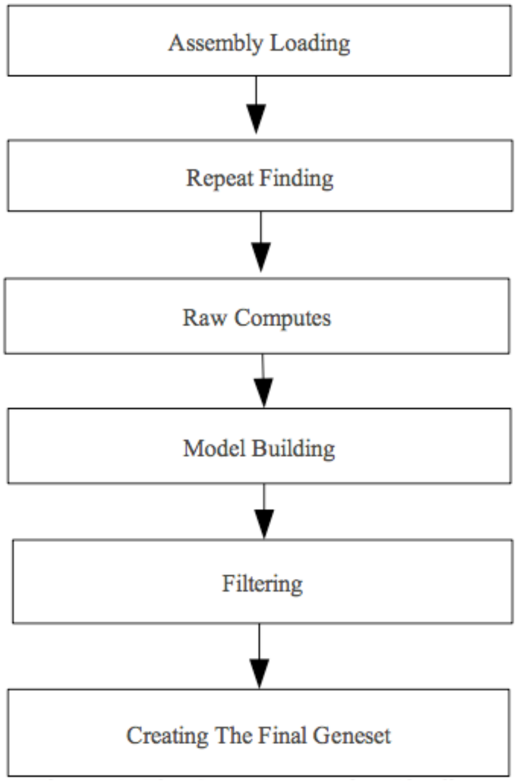

Figure 1: The gene annotation pipeline

Repeat Finding

After loading into a database the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.3.0), Dust [2] and TRF [3]. RepeatMasker was run twice; the first run used a repbase library with parameters -nolow -s -species 'mammal'), and the second run used a custom library generated with RECON. The RepeatMasker library and custom RECON [4] library combined to mask 49.35% of the armadillo 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 468955 UniProt, 360540 UniGene and 356842 Vertebrate RNA sequences aligning to the genome.

Model Generation

Our gene annotation system is evidence-based; all protein coding and non-coding RNA gene models are supported by biological sequences from public databases. Input data for the protein coding gene models came from UniProt, Broad Institute and the Ensembl release 71 for human. Data from each source were aligned to the genome and filtered in order to generate gene models. The number of preliminary gene models (before filtering) generated from each data source/pipeline are outlined in Table 1.

Pipeline	Source	Number of Models
Similarity	Uniprot proteins	447160
RNAseq	Broad Institute	281961
Ensembl Longest Translations	Ensembl Release 71 proteins for human	17025

Table 1: Gene Model Generation Overview

Similarity Pipeline: Generating coding models using proteins from related species

Coding models were generated using data from related species. The UniProt alignments from the Raw Computes step were filtered using only sequences from vertebrate species. WU-BLAST was rerun for these sequences and the results were passed to Genewise [14] to build coding models.

RNAseq Pipeline

RNAseq data provided by the Broad Institute were used in the annotation. These comprised paired end reads from samples including: ascending colon, Cerebellum with brainstem, Heart, Kidney, Liver, Lung, Quadriceps and Spleen. 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 RNAseq pipeline produced 281961 transcript models in total. The predicted open reading frames were compared to Uniprot Protein Existence (PE) classification level 1, 2, 3, 4 and 5 proteins using WU-BLAST. Models with poorly scoring or no BLAST alignments were split into a separate class and not used in the final protein coding gene set.

Ensembl Longest Translations

The longest translation for each protein coding gene in Ensembl release e71 for human were downloaded. These proteins were aligned against the armadillo DasNov3.0 assembly using Exonerate [13] to produce a set of 17025 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 [15] 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.

Both similarity and RNAseq models were ordered high in the layering process. A basic overview of the final layering is as follows:

Mammal 1,2 similarity models

Only vertebrate, non-mammal 1,2 similarity models

Strong RNAseq models

Mammal 3,4 and 5 similarity models

Only vertebrate, non-mammal 3,4 and 5 similarity models

Weaker RNAseq models

Human Ensembl translations (Exonerate)

Small mammals 1,2,3,4 and 5 similarity models

In the above ordering strong RNAseq models were ones where we had a matching BLAST alignment to a Uniprot protein that had both a hit coverage and percent identity of greater than or equal to 80 percent. For the weak RNAseq models the hit coverage and percent identity of the BLAST alignments were between 50 to 80 percent.

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNAseq sequences. At the UTR addition stage 47046 gene models out of total of 87610 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 of 24310 genes with 28189 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 1500 genes were labelled as pseudogenes.

Creating The Final Gene Set

ncRNAs

Small structured non-coding genes were added using annotations taken from RFAM [16] and miRBase [17]. WU-BLAST was run for these sequences and models built using the Infernal software suite [18].

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. Databases searched include: Seg, SignalP, Ncoils, Tmhmm, Prints, Pfscan, Pfam, Tigrfam, Superfamily, Smart and Pirsf.

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 22711 protein-coding genes, of which 13 are protein-coding mitochondrial models. These represent 26551 transcripts. A total of 24 non-coding mitochondrial genes were imported. Pseudogene analysis identified 1500 pseudogenes. The ncRNA pipeline added a total of 5960 ncRNAs.

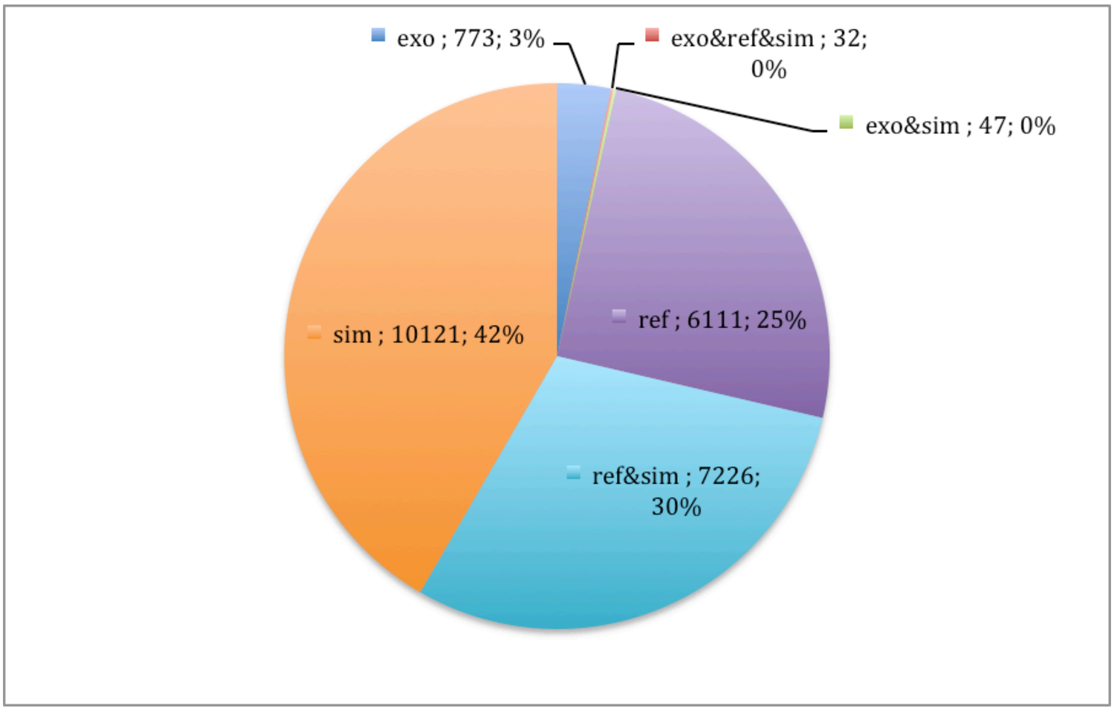

Figure 2: Distribution of gene predictions. sim: similarity, ref: RNAseq data, exo: exonerate.

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.

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:

Coverage estimate

A higher coverage usually indicates a more complete assembly.

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

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.

Number of contigs and scaffolds

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

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:

- 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/-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 Bao Z. and Eddy S.R: **Automated de novo Identification of Repeat Sequence Families in Sequenced Genomes.** Genome Research 2002, **12**:1269-1276.

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 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]

16 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]

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 [PMID: 16381832]

18 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]