Ensembl gene annotation project

Anolis carolinensis (Anole lizard)

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

Approximate time: 1 week

The annotation process of the high-coverage anole lizard 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 -species "squamata" -s`'), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 10.7% of the anole lizard genome.
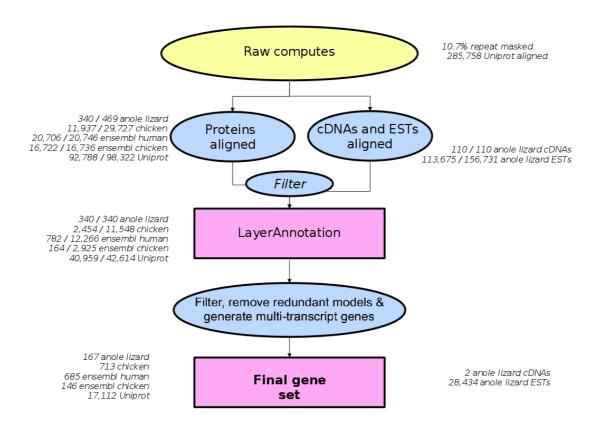
Figure 1: Summary of anole lizard gene annotation project.

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 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 285.758 UniProt, 339.179 UniGene and 316.981 Vertebrate RNA sequences aligning to the genome.

Exonerate Stage: Generating coding models from anole lizard and chicken evidence

Approximate time: 1 week

Next, anole lizard and chicken protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.]). The anole lizard and chicken protein sequences were mapped to the genome using Pmatch as indicated in [Figure 2] and [Figure 3].

Models of the coding sequence (CDS) were produced from the proteins using Genewise [13.] and Exonerate [12.]. Where one protein sequence had generated more than one coding model at a 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 species-specific (in this case anole lizard and chicken) data is referred to as the "Targetted stage". This stage resulted in 340 (of 469) anole lizard proteins and 11.548 (of 29.727) chicken proteins used to build coding models to be taken through to the genebuild stage.
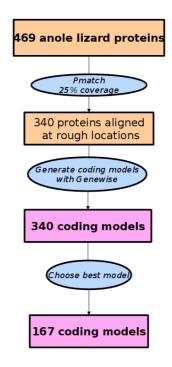
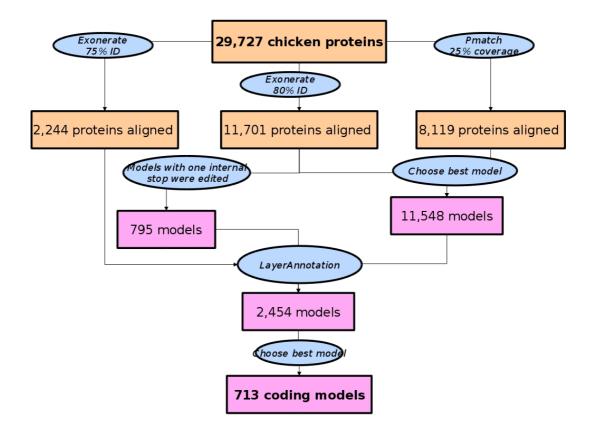**Figure 2: Targetted stage using anole lizard protein sequences.**

Figure 3: Alignment and filtering of chicken proteins.

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

Approximate time: 1 week

Following the chicken and anole lizard Targetted alignments, additional coding models were generated as follows. The UniProt alignments from the Raw Computes step were filtered and only those sequences belonging to UniProt's Protein Existence (PE) classification level 1 and 2 were kept. WU-BLAST was rerun for these sequences and the results were passed to Genewise [13.] to build coding models. The generation of transcript models using data from related species is referred to as the "Similarity stage". This stage resulted in 751 coding models classified as reptiles, 5880 as birds and 91691 as mammals.

cDNA and EST Alignment

Approximate time: 1 week

Anole lizard cDNAs and ESTs and chicken cDNAs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate [Figure 4].
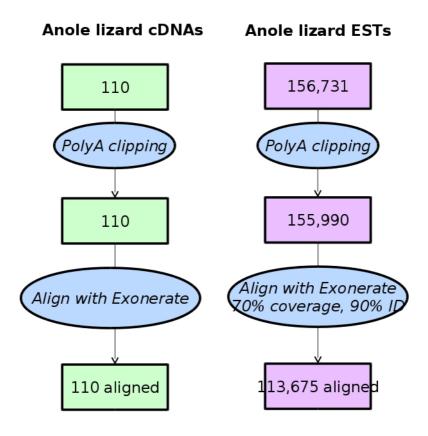
Anole lizard cDNAs   **Anole lizard ESTs**

```
┌──────────────┐          ┌──────────────┐
│     110      │          │   156,731    │
└──────┬───────┘          └──────┬───────┘
   (PolyA clipping)          (PolyA clipping)
       │                         │
┌──────▼───────┐          ┌──────▼───────┐
│     110      │          │   155,990    │
└──────┬───────┘          └──────┬───────┘
 (Align with Exonerate)   (Align with Exonerate
       │                   70% coverage, 90% ID)
┌──────▼───────┐          ┌──────▼───────┐
│ 110 aligned  │          │113,675 aligned│
└──────────────┘          └──────────────┘
```

Figure 4: Alignment of anole lizard cDNAs and ESTs to the anole lizard genome.

Of these, 110(of 110) anole lizard cDNAs aligned, and 113.675 (of 156.731) anole lizard ESTs aligned. All alignments were at a cut-off of 90% coverage and 80% identity. EST alignments were used to generate EST-based gene models similar to those for human [14.] and these are displayed on the website in a separate track from the Ensembl gene set.

Filtering Coding Models

Approximate time: 2 weeks

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus and LayerAnnotation, to result in a preliminary set of coding models. The Apollo software [15.] was used to visualise the results of filtering.

Missed orthologues : Retrieving one-to-one orthologues from other Ensembl species

Approximate time: 1 week

The preliminary set of coding models was compared to the set of Ensembl translations from human and chicken as follows : human-chicken one-to-one orthologues were retrieved using the Compara pipeline and aligned to the anole lizard models.

Any alignments from human or chicken orthologues that did not align uniquely with an anole lizard model were included in the anole set [Figure 5].

This resulted in 4.656 human orthologues and 2.740 chicken orthologues taken through to the genebuild stage.
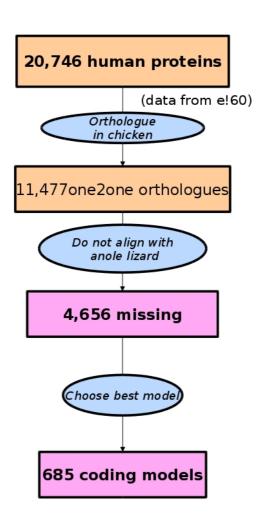

Figure 5: Retrieving missed orthologues for human and chicken.

Generating multi-transcript genes

Approximate time: 3 weeks

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 17792 genes included 170 genes with at least one transcript supported by anole lizard proteins, a further 660 genes without anole lizard evidence but with at least one transcript supported by chicken evidence. An additional 585 genes were supported by Ensembl proteins from either human or chicken. The remaining 16962 genes had transcripts supported by proteins from other sources [Figure 6].
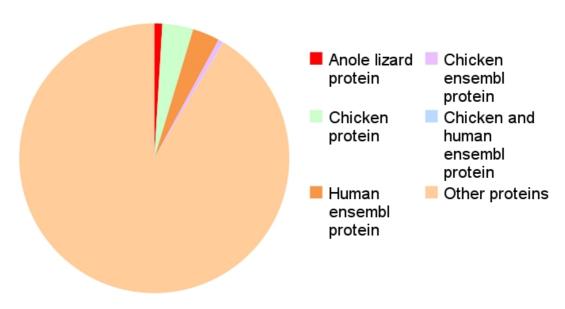
Evidence for genes

Figure 6: Supporting evidence for anole lizard final gene set.

The final transcript set of 18939 transcripts included 176 transcripts with support from anole lizard proteins, 702 transcripts with support from chicken proteins, 704 transcripts with support from human and chicken Ensembl proteins and 17357 transcripts with support from UniProt SwissProt [Figure 7].
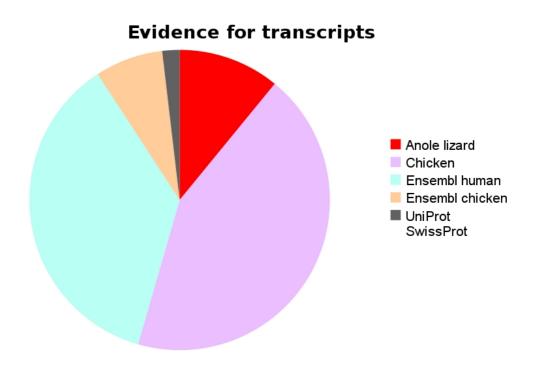
Evidence for transcripts

Legend:
- Anole lizard
- Chicken
- Ensembl human
- Ensembl chicken
- UniProt SwissProt

Figure 7: Supporting evidence for final transcript set.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 1 week

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

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 noncoding 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:

- 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, 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]

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 Bioinformatic*s 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, 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]