Ensembl gene annotation project

Latimeria chalumnae (Coelacanth)

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

The annotation process of the high-coverage Coelacanth 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 a custom Coelacanth repeat library, Dust [2] and TRF [3]. RepeatMasker masked 41% of the genome. Adding low complexity masking (including gaps) with Dust brings the total masked to 67%.

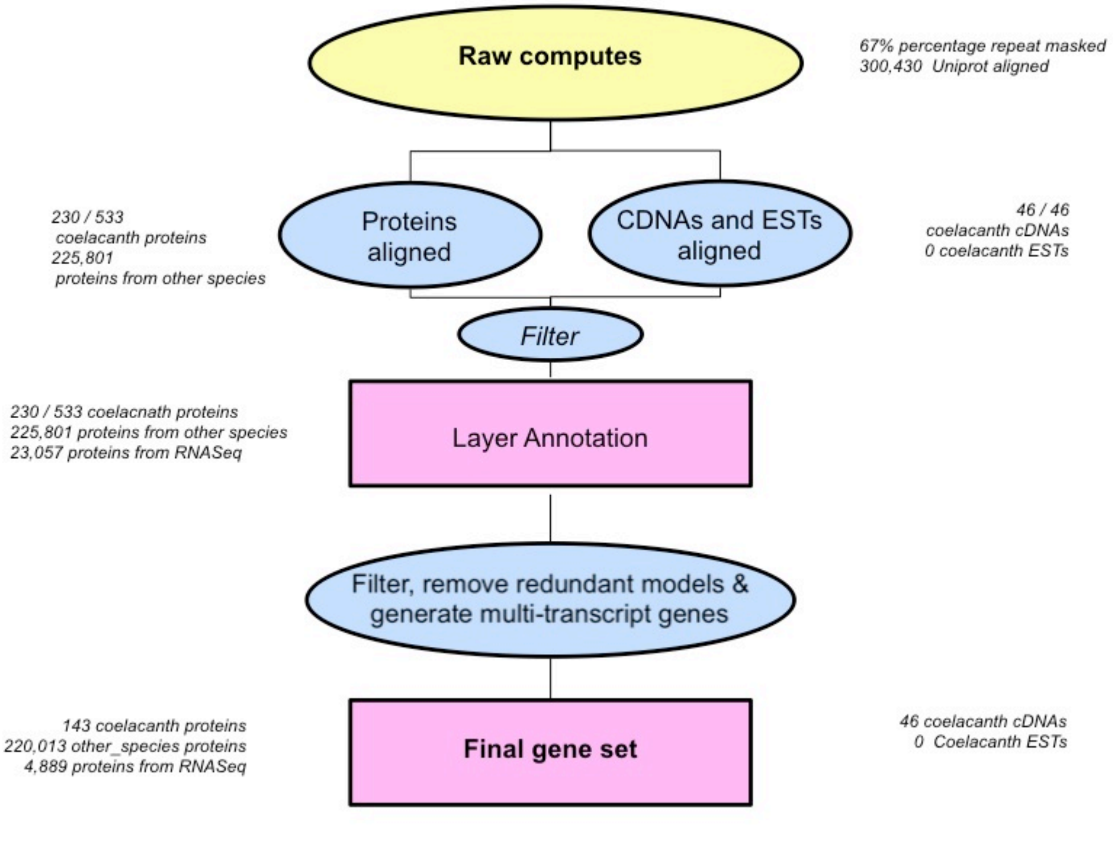
Raw computes
67% percentage repeat masked
300,430 Uniprot aligned
230 / 533
coelacanth proteins
225,801
proteins from other species
Proteins aligned
CDNAs and ESTs aligned
46 / 46
coelacanth cDNAs
0 coelacanth ESTs
Filter
230 / 533 coelacnath proteins
225,801 proteins from other species
23,057 proteins from RNASeq
Layer Annotation
Filter, remove redundant models & generate multi-transcript genes
143 coelacanth proteins
220,013 other_species proteins
4,889 proteins from RNASeq
Final gene set
46 coelacanth cDNAs
0 Coelacanth ESTs

Figure 1: Summary of Coelacanth 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 300,430 UniProt vertebrate and 98,486 Uniprot non-vertebrate sequences, 378,875 UniGene and 342,925 Vertebrate RNA sequences aligning to the genome.

Targeted Stage: Generating coding models from Coelacanth evidence

Coelacanth protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8] and Genbank) and filtered to remove sequences based on predictions. The Coelacanth sequences were mapped to the genome using Pmatch as indicated in Figure 2.

Models of the coding sequence (CDS) were produced from the proteins using Genewise [13]. 2 sets of models were produced, one with only consensus splice sites and one where non-consensus splices were allowed; where a single protein sequence had 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 Coelacanth-specific data is referred to as the "Targeted stage". This stage resulted in 143 of 533 Coelacanth proteins used to build 153 coding models.

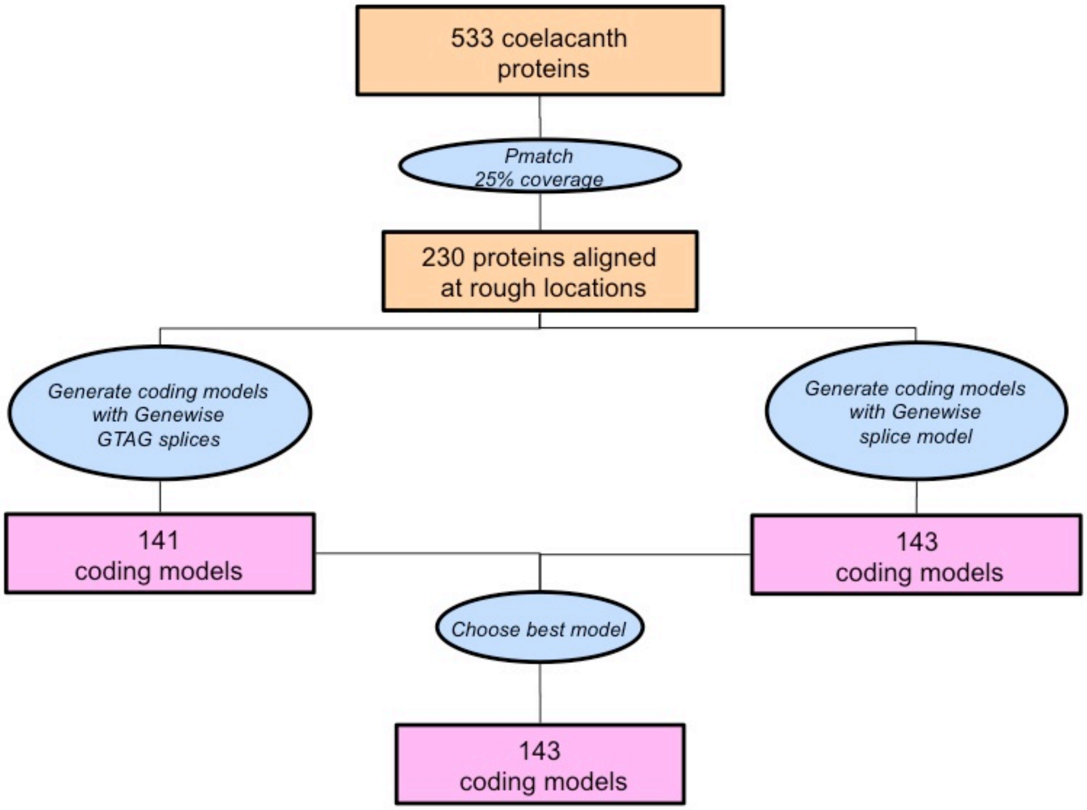
533 coelacanth proteins
Pmatch 25% coverage
230 proteins aligned at rough locations
Generate coding models with Genewise GTAG splices
Generate coding models with Genewise splice model
141 coding models
143 coding models
Choose best model
143 coding models

Figure 2: Targeted stage using Coelacanth specific proteins

cDNA and EST Alignment

Coelacanth cDNAs were downloaded from Genbank, clipped to remove polyA tails, and aligned to the genome using Exonerate. Of these, 46 of 46 Coelacanth cDNAs aligned with a cut-off of 90% coverage and 97% identity. No ESTs were found.

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

Due to the small number of Coelacanth specific protein and cDNA evidence the majority of the gene models were based on proteins from other species. UniProt alignments from the Raw Compute step were filtered to favor proteins classed by UniProt's Protein Existence (PE) classification level 1 and 2. Proteins from other PE levels were used where no other evidence was available; similarly, mammalian proteins were favored over non-mamalian. WU-BLAST was rerun for these sequences and the results were passed to Genewise 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 297,885 coding models.

Filtering Coding Models

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus, RNA-Seq spliced alignments supporting introns were used to help filter the set. 247,112 models were rejected as a result of filtering. The Apollo software [15] was used to visualise the results of filtering. [Figure 3]

Addition of RNA-Seq models

The largest set of Coelacanth specific evidence was from Illumina paired end RNASeq, this was used where appropriate to help inform our gene annotation. A set of 375 million reads that passed QC were aligned to the genome using BWA resulting in 225 million reads aligning and properly pairing. The Ensembl RNA-Seq pipeline was used to process the BWA alignments and create a further 30 million split read alignments using Exonerate. The split reads and the processed BWA alignments were combined to produce 23,057 transcript models in total; one transcript per loci. 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 discarded. The resulting models were added into the gene set where they produced a novel model or splice variant, in total 4,994 models were added.

For release e!67 an additional set of RNASeq models were added using samples from liver and testis taken from the related species Latimeria *menadoensis*, in total 150M paired end reads were aligned using BWA and a further 10M split read alignments were produced using the Ensembl RNASeq pipeline. Gene models were produced and BLAST filtered leaving 9,362 high confidence gene models, which were merged into the original coelacanth gene set where they added novelty at the gene or transcript level. In total 547 new genes were added and 1,782 transcripts.

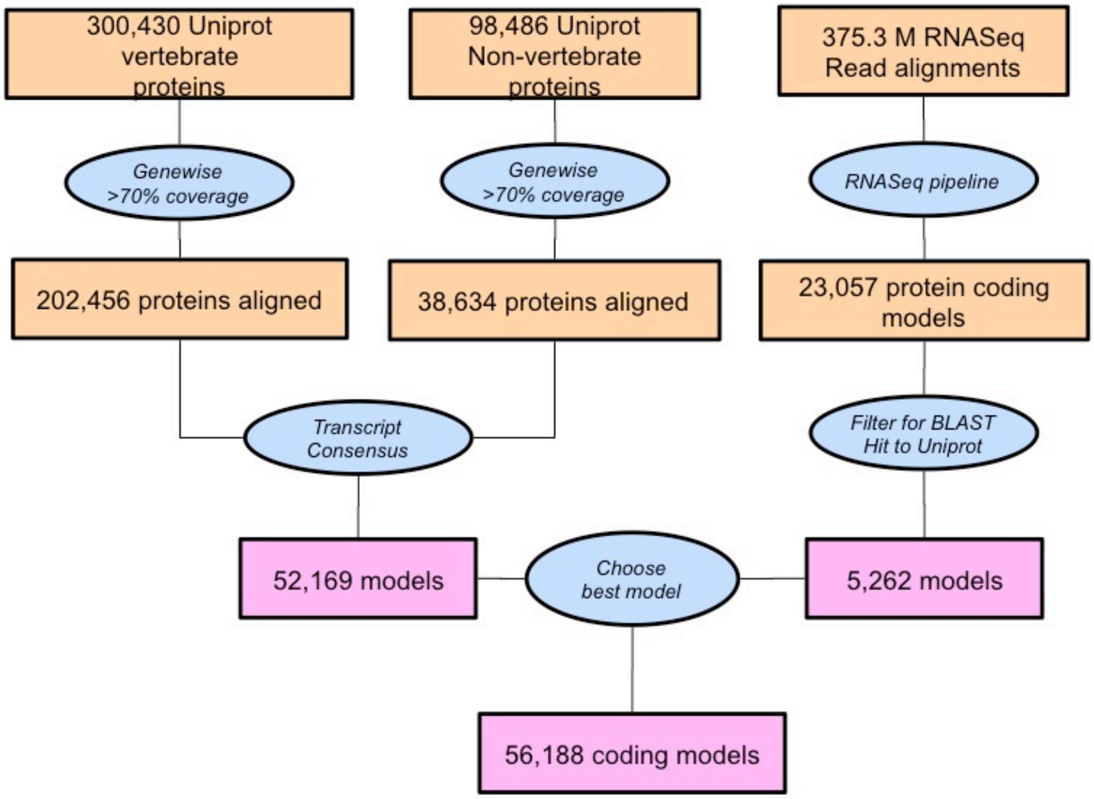

Figure 3: Alignment and filtering of other species proteins and addition of RNASeq models

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 contains 19,697 protein coding genes with 23,740 transcripts, 2,894 ncRNAs and 35 mitochondrial genes.

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 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].

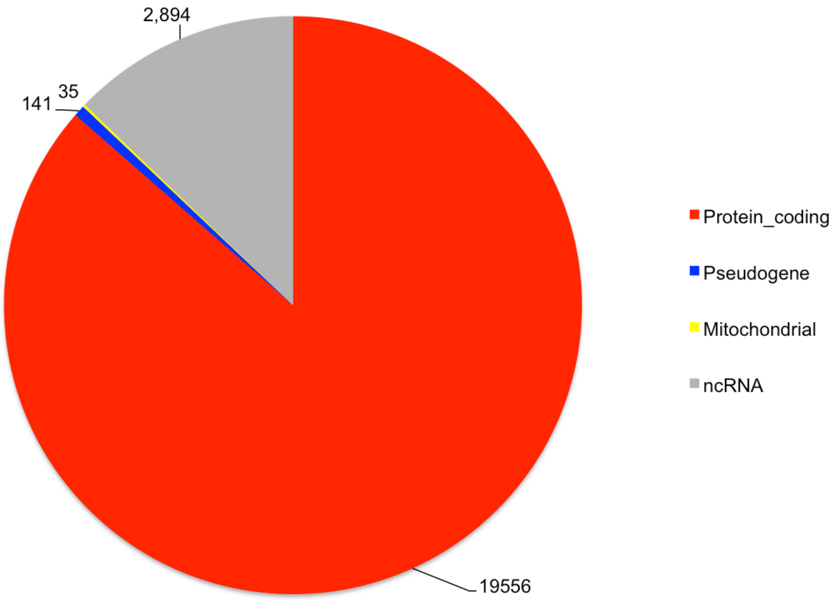

Figure 4: Composition of Coelacanth gene set.

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 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 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/ensembldoc/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 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, 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. 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