Ensembl gene annotation project

Callithrix jacchus (Marmoset)

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 marmoset assembly began with the raw compute stage [Figure 1] whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.2.5 with parameters '`-nolow -species homo –s`'), Dust [2] and TRF [3]. RepeatMasker and Dust combined masked 47% of the marmoset genome.

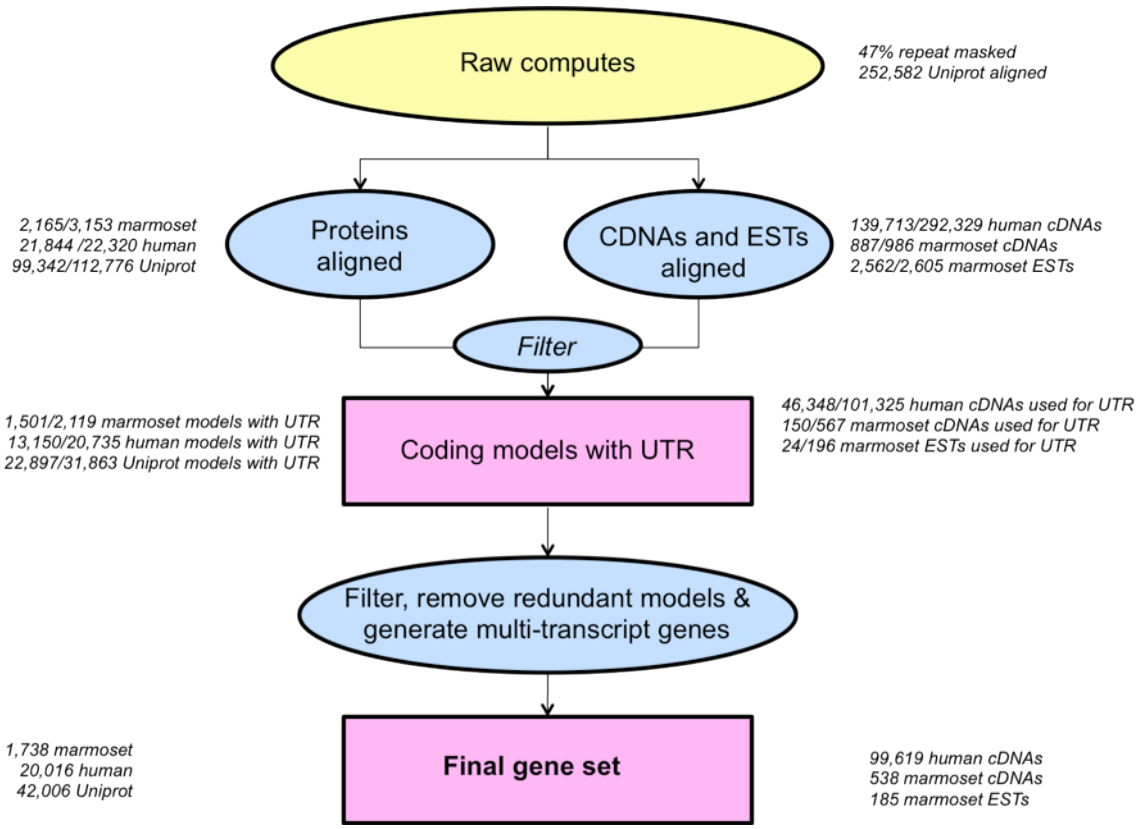

Figure 1: Summary of marmoset gene annotation project.

Transcription start sites were predicted using Eponine–scan [4] and FirstEF [5]. CpG islands [6] and tRNAs [7] were also predicted. Genscan was run across RepeatMasked 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 252,582 UniProt, 316,384 UniGene and 317,679 Vertebrate RNA sequences aligning to the genome.

Targetted Stage: Generating coding models from marmoset and human evidence

Approximate time: 1 week

Next, marmoset and human protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [13] and RefSeq [14]). The marmoset and human protein sequences were mapped to the genome using Pmatch [15] as indicated in [Figure 2] and [Figure 3].

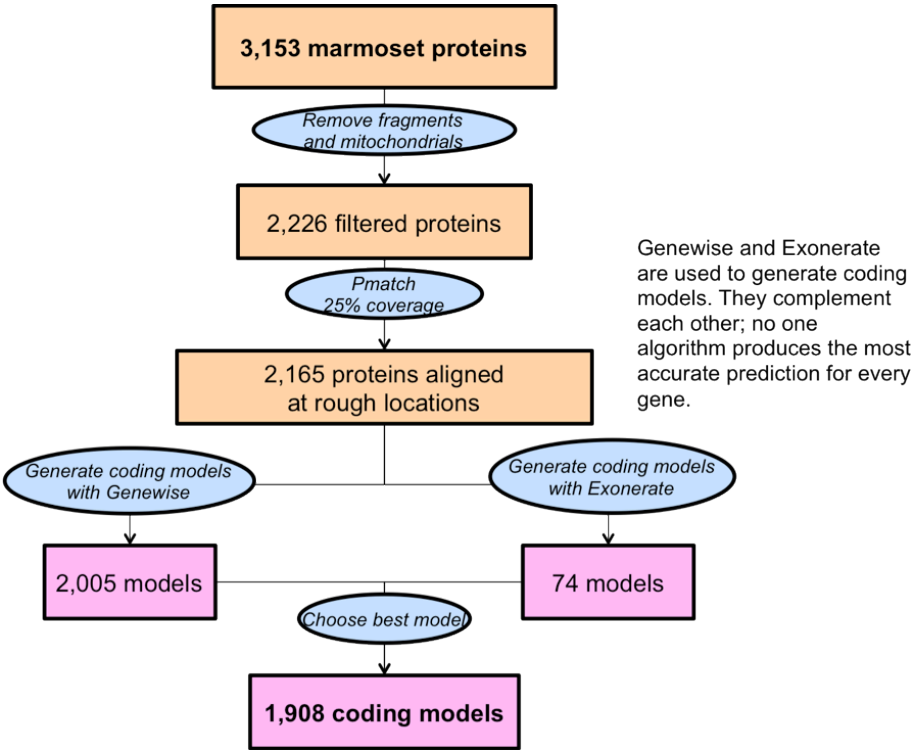

Figure 2: Targetted stage using marmoset protein sequences.

Models of the coding sequence (CDS) were produced from the proteins using Genewise [16] and Exonerate [17]. Where one protein sequence had generated more than one coding model at a locus, the BestTargetted module [18] 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 marmoset and human) data is referred to as the "Targetted stage". This stage resulted in 1,908 (of 3,153) marmoset proteins and 20,735 (of 22,320) human proteins used to build coding models to be taken through to the UTR addition stage.

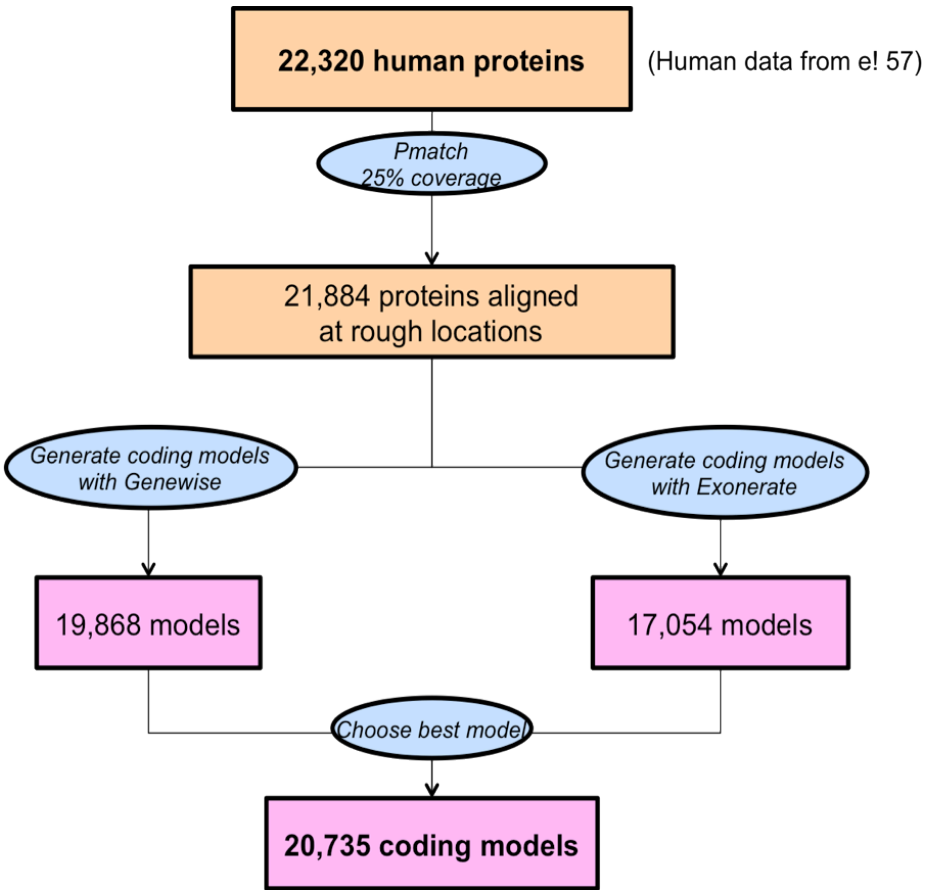

Figure 3: Targetted stage using human protein sequences.

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

Approximate time: 1 week

Following the Targetted stage, 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 to build coding models in regions not already covered by the Targetted Stage. The generation of transcript models using data from related species is referred to as the "Similarity stage". This stage resulted in 57,019 mammalian and 42,323 non-mammalian coding models.

cDNA and EST Alignment

Approximate time: 1 week

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

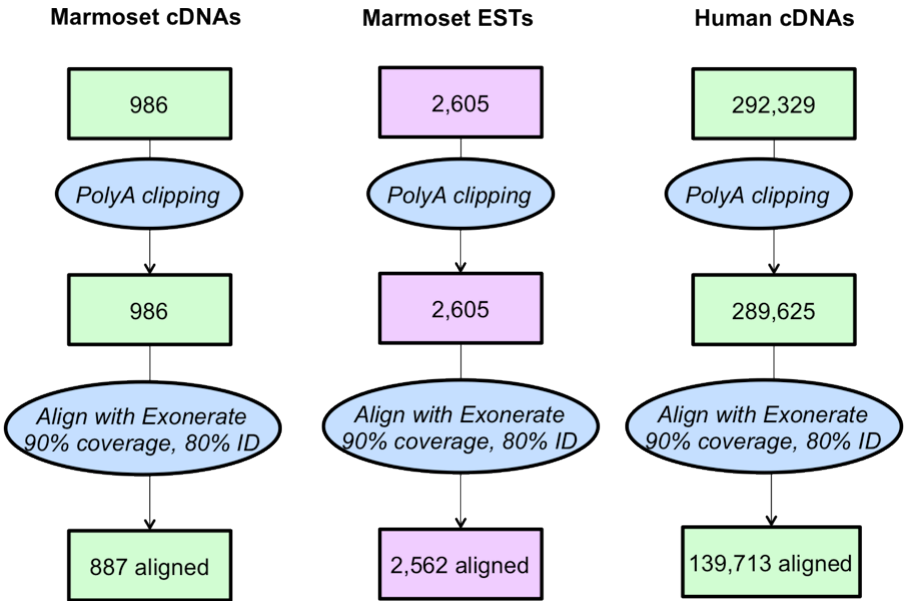

Figure 4: Alignment of marmoset cDNAs and ESTs, and human cDNAs to the marmoset genome.

Of these, 139,713 (of 292,329) human cDNAs aligned, 887 (of 986) marmoset cDNAs aligned, and 2,562 (of 2,605) marmoset 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 [19] and these are displayed on the website in a separate track from the Ensembl gene set.

Filtering Coding Models

Approximate time: 1 week

Coding models from the Similarity stage were filtered to remove models with little cDNA or EST support. Filtering modules such as TranscriptConsensus and LayerAnnotation were used; the cDNA and EST alignments were used to score the coding models in the TranscriptConsensus module. The Apollo software [20] was used to visualise the results of filtering.

Addition of UTR to coding models

Approximate time: 1 week

The set of coding models was extended into the untranslated regions (UTRs) using human cDNA, marmoset cDNA and marmoset EST sequences. This resulted in 1,501 (of 2,119) marmoset coding models with UTR, 13,150 (of 20,735) human coding models with UTR, and 22,897 (of 31,863) UniProt coding models with UTR.

Generating multi-transcript genes

Approximate time: 2-3 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 21,168 genes included 219 genes with at least one

transcript supported by marmoset protein, a further 15,706 genes without marmoset evidence but with at least one transcript supported by human evidence. The remaining 5,243 genes had transcripts supported by proteins from other sources [Figure 5].

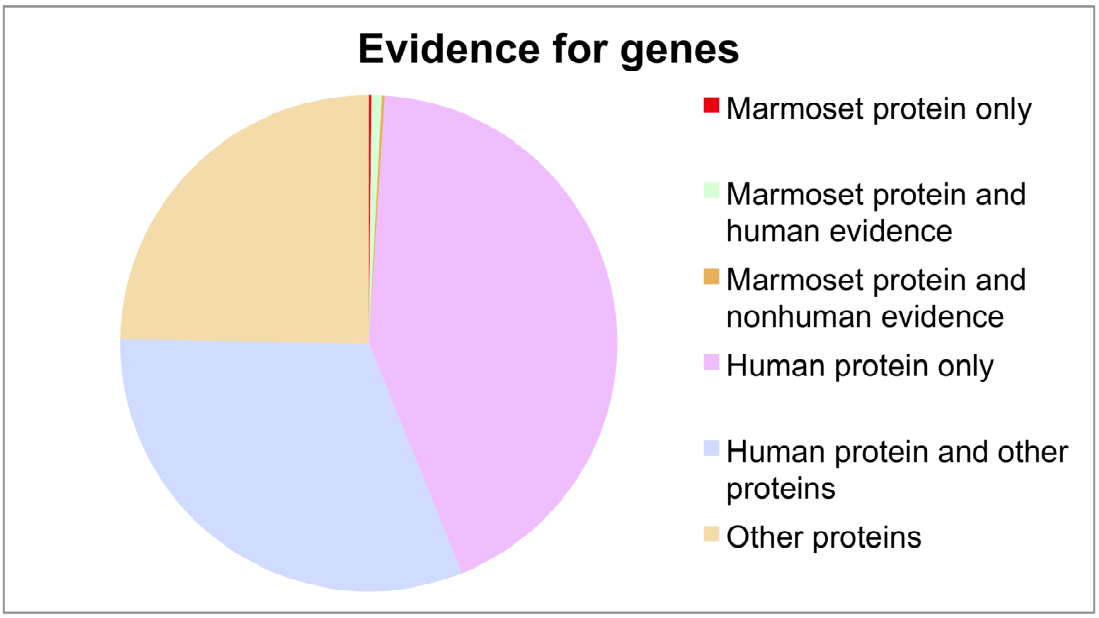

Figure 5: Supporting evidence for marmoset final gene set.

The final transcript set of 44,973 transcripts included 232 transcripts with support from marmoset proteins, 5,731 transcripts with support from human proteins, 24,718 transcripts with support from human cDNA with CDS information, 12,770 transcripts with support from UniProt SwissProt, and 1,522 transcripts with support from other protein sequences [Figure 6].

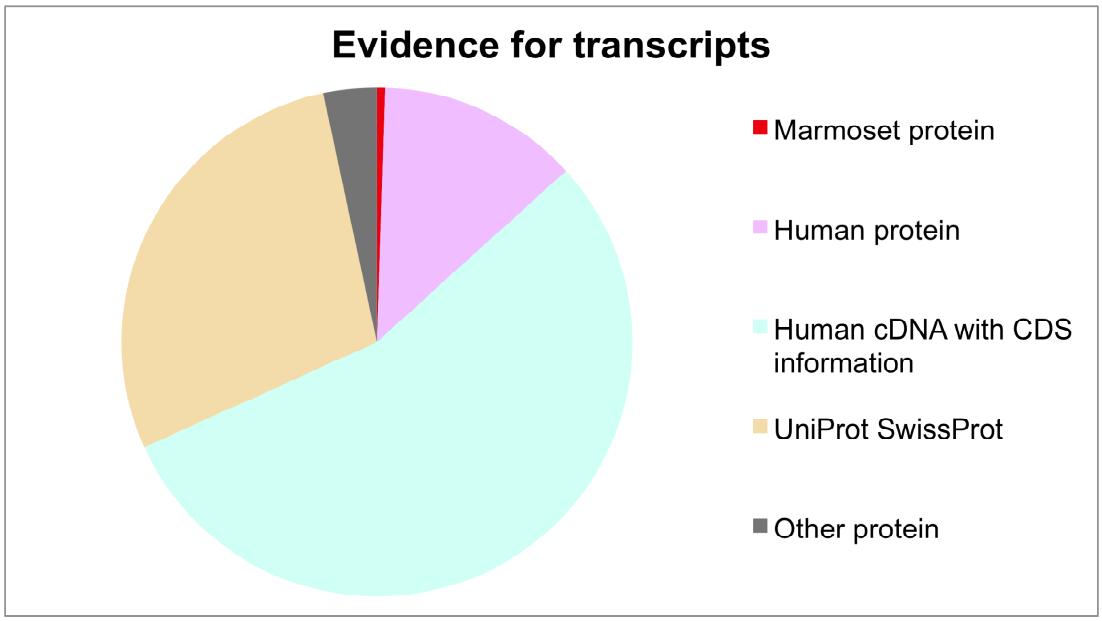

Figure 6: Supporting evidence for marmoset final transcript set.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 2 weeks

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

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. CpG. No publication found.
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. http://www.uniprot.org/
14. Pruitt KD, Tatusova T, Klimke W, Maglott DR: NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* 2009, **37(Database issue):**D32-36. [PMID: 18927115.]
15. Durbin R: Pmatch., unpublished.
16. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res.* 2004, **14(5):**988-995. [PMID: 15123596]
17. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatic*s 2005, **6:**31. [PMID: 15713233]
18. http://www.ensembl.org/info/docs/Pdoc/ensembl-analysis/index.html
19. Eyras E, Caccamo M, Curwen V, Clamp M. **ESTGenes: alternative splicing from ESTs in Ensembl.** *Genome Res.* 2004 **14(5):**976-987. [PMID: 15123595]
20. 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]