Ensembl gene annotation project

Nomascus leucogenys (Northern White-Cheeked Gibbon)

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

Approximate time: one week

The annotation process of the high-coverage Gibbon 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, run twice, with parameters '`-nolow -Gibbon "Nomascus leucogenys" –s`' and '`-nolow -mammal –s`'), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 54% of the Gibbon genome.

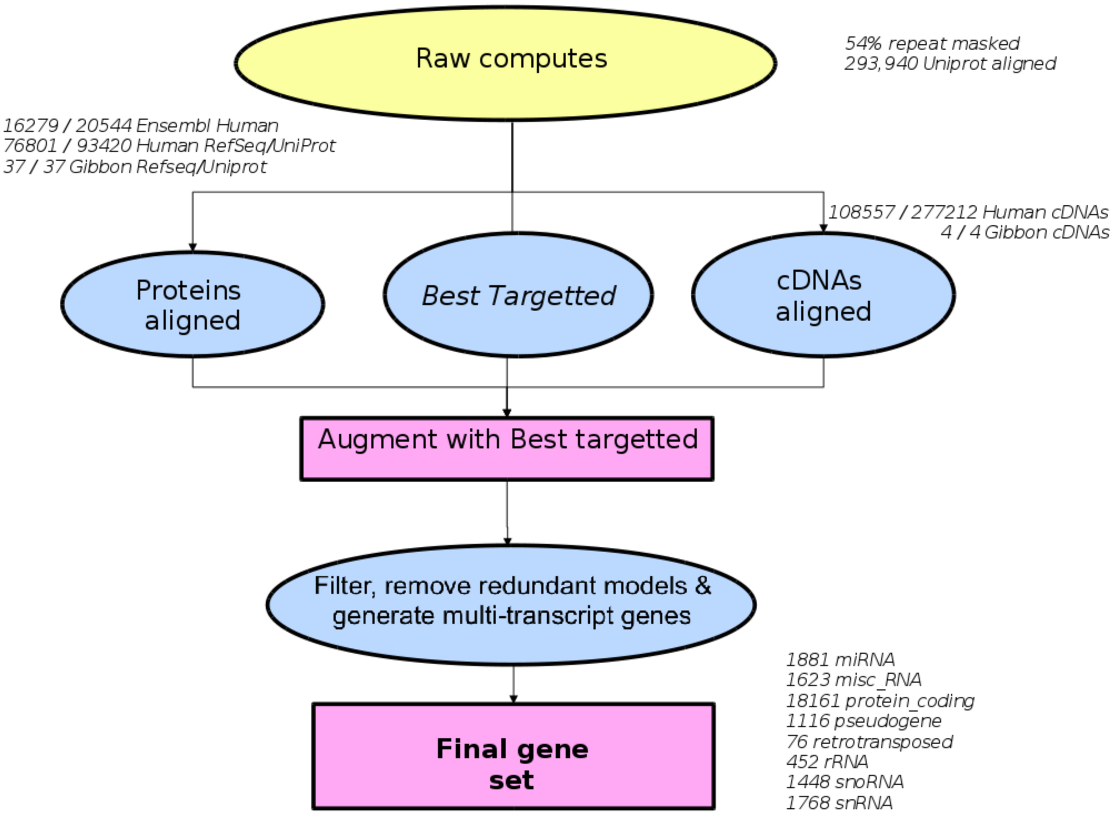

Figure 1: Summary of Gibbon gene annotation project.

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 293,940 UniProt, 343,641 UniGene and 336,483 Vertebrate RNA sequences aligning to the genome.

Generating coding models from Human Ensembl Translations

Approximate time: two weeks

First, Human Ensembl data from e!61 was taken and aligned to the genome using Exonerate [12.]. This resulted in 16279 models after cut offs were set at 85% coverage and 80% identity. Additionally, 'mid-ranged' models as low as 50% coverage and identity were taken where they matched a best targetted (see below) entry by intronic regions and the best targetted model had a translation of >=50 amino acids.

Generating a supportive evidence coding model set from Human and Gibbon proteins

Approximate time: three weeks

Next, Gibbon and Human protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [9] and RefSeq [9.]). The Gibbon and Human protein 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.] 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 Gibbon-specific (in this case Gibbon and Human) data is referred to as the "Targetted stage". This stage resulted in 99069 coding models and was used as evidence for alignments of "mid-ranged" matches from the Human Ensembl exonerate alignments mentioned in the previous stage.

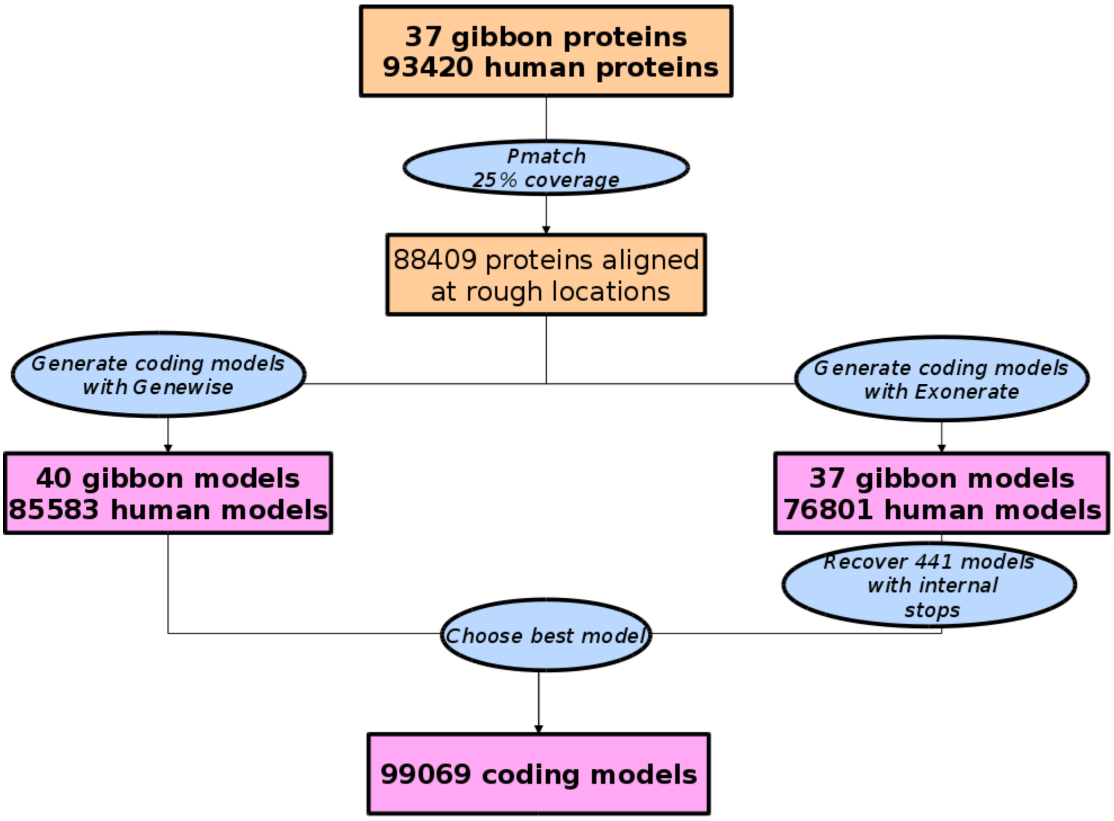

Figure 2: Targetted stage using Gibbon protein sequences.

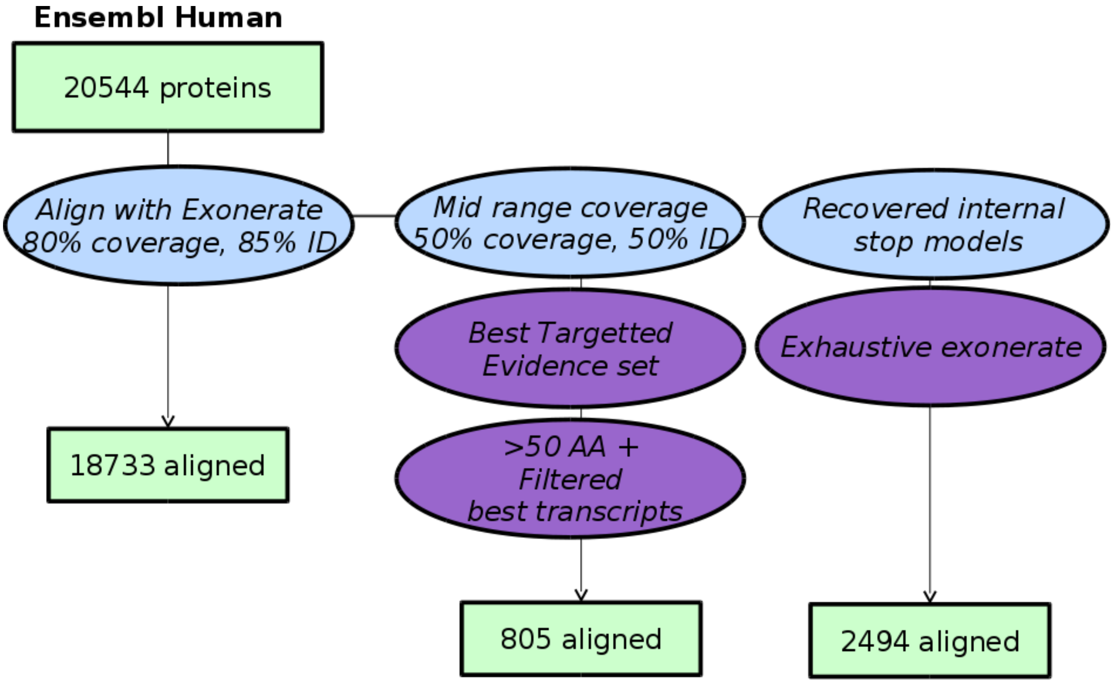
Ensembl Human
20544 proteins
Align with Exonerate 80% coverage, 85% ID
Mid range coverage 50% coverage, 50% ID
Recovered internal stop models
Best Targetted Evidence set
Exhaustive exonerate
>50 AA + Filtered best transcripts
18733 aligned
805 aligned
2494 aligned

Figure 3: Filtering of Human Ensembl proteins.

Recovery of internal stop entries

The exonerate alignments can produce transcript models with stop codons, which cannot be used in the final gene set because the GeneBuilder module removes models which include internal stops. For models with only one stop we attempt to replace the stops with small introns where they lie in the middle of the exon. For models wirth more than one stop attempts to get a better alignment are then made on the region using exonerate in 'exhaustive' mode.

CDNA Alignments

Approximate time: two weeks

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

Of these, 108557 (of 277212) Human cDNAs aligned and 4 (of 4) Gibbon cDNAs aligned. Human alignments were at a cut-off of 90% coverage and 90% identity.

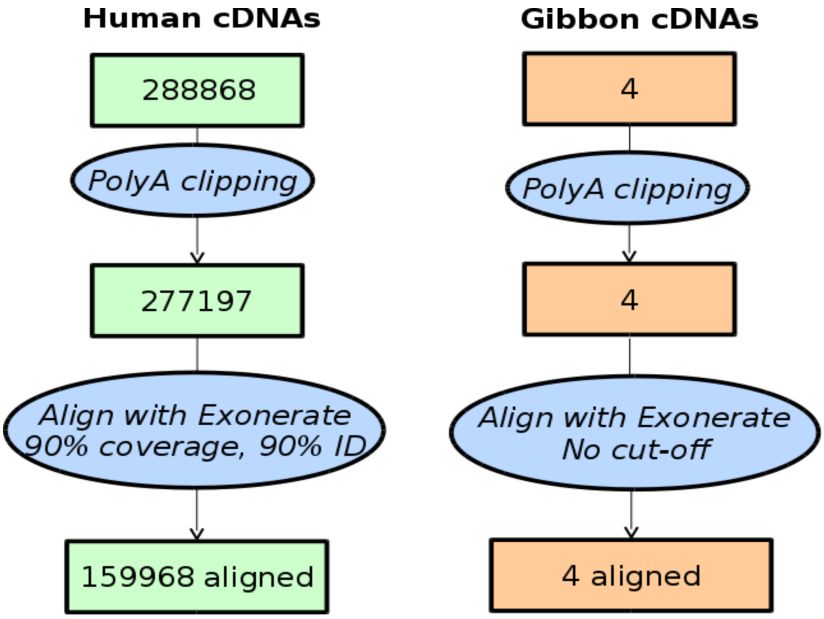

Figure 4: Alignment of Gibbon and Human cDNAs

Addition of UTR to coding models

Approximate time: one week

The set of coding models was extended into the untranslated regions (UTRs) using Human and Gibbon cDNA sequences. This resulted in 19 (of 26) Gibbon coding models with UTR and 21427 (of 77888) Human coding models with UTR.

Generating multi-transcript genes

Approximate time: one week

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 set of 19461 coding genes included 13 genes with at least one transcript supported by Gibbon proteins with the remaining having at least one transcript supported by Human evidence. [Figure 5].

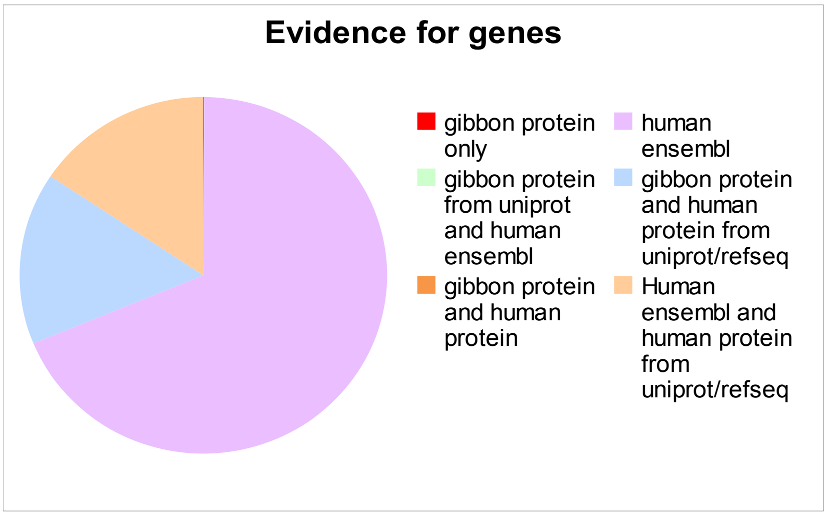

Figure 5: Supporting evidence for Gibbon final gene set.

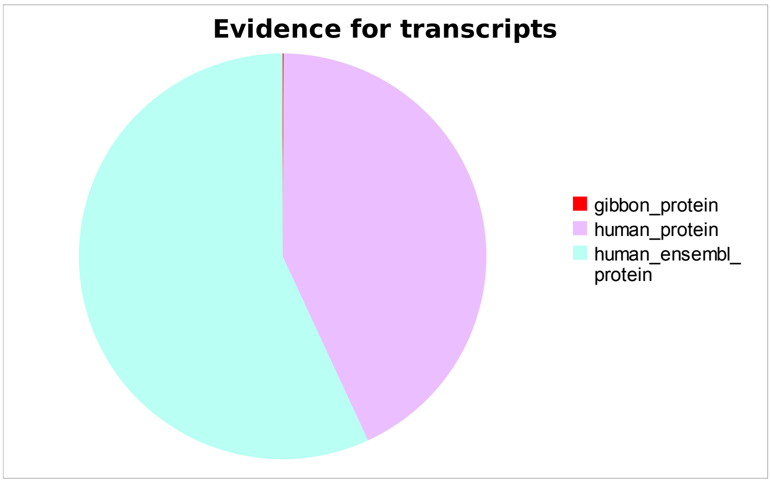

Figure 6: Supporting evidence for Gibbon final transcript set.

The final transcript set of 24554 transcripts included 18 transcripts with support from gibbon proteins, 13190 transcripts with support from Human Ensembl proteins and 19974 transcripts with support from UniProt/RefSeq [Figure 6].

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: one 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 Gibbon, the stable identifiers will be 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 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 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 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. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res.** 2011, **39: D214-D219**. http://www.uniprot.org/downloads [PMID: 21051339]
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. 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]