Ensembl gene annotation project (*e!*64)

Bos taurus (cow, UMD3.1 assembly)

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

Approximate time: 2 weeks

The annotation process of the high-coverage cow 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 "cow" -s`'), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 48.95% of the species genome.

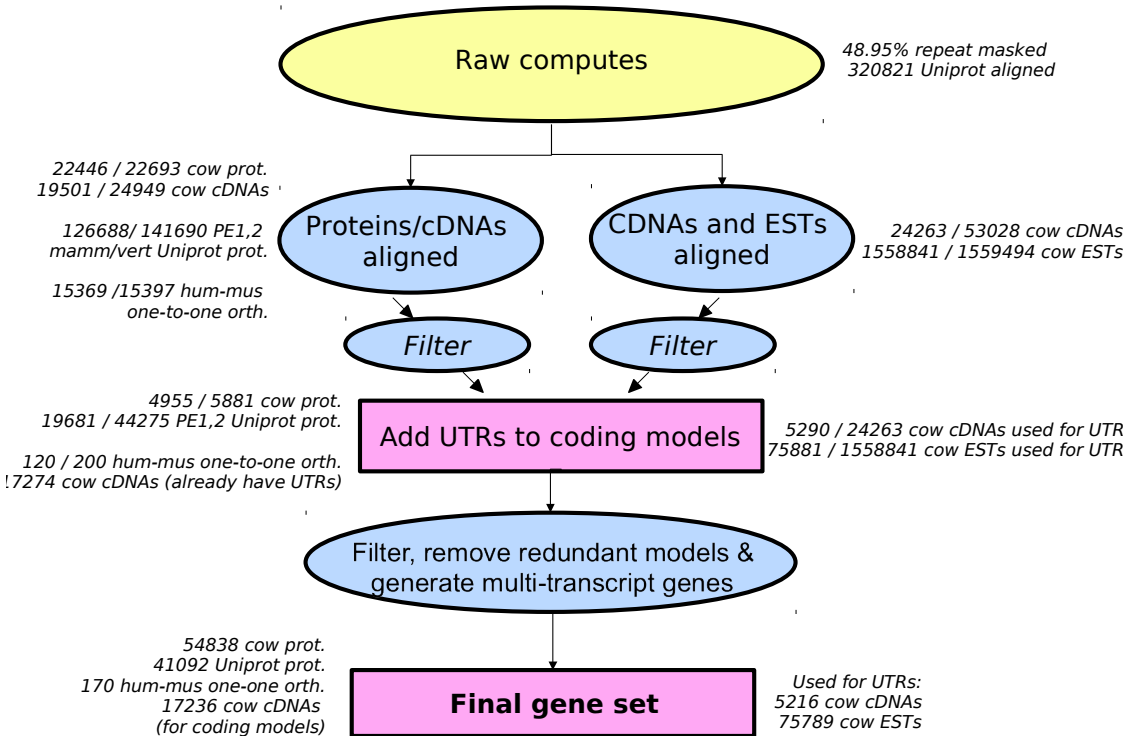

Figure 1: Summary of cow gene annotation project.

Transcription start sites were predicted using Eponine–scan [4.] and FirstEF [5.]. CpG islands longer than 400 bases (Micklem, G., pers. Comm.) 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 320821 UniProt, 344434 UniGene and 336326 Vertebrate RNA sequences aligning to the genome.

Targetted Stage: Generating coding models from cow evidence

Approximate time: 2-3 weeks

Next, cow protein and cDNA sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.] for proteins, ENA/Genbank/DDBJ and RefSeq [9.] for cDNAs). The downloaded UniProt sequences were filtered so only those falling into Uniprot's Protein Existence (PE) classification level 1 and 2 were kept, while the RefSeq sequences were filtered to remove those based on predictions (e.g. RefSeq accession starting with “XP” and “XM” for cDNAs). Cow cDNAs were also filtered to remove those associated with the discarded cow proteins. Following the filtering of input protein/cDNA sequences, the cow protein sequences were first mapped to rough locations in the genome using Pmatch to reduce the search space for the subsequent Genewise step, as indicated in [Figure 2]. Models of the coding sequence (CDS) were produced from the proteins using Genewise [13.], which was run with four different sets of parameters to accommodate for cases where some coding models contain non-canonical (non GT/AG) splice sites. In parallel to the Genewise step, cow cDNAs with known CDS start coordinate and CDS length were aligned to the genome using Exonerate (*cdna2genome* model) [12.] to generate coding models [Figure 2]. Because all cDNAs used in this step had known pairing with proteins (e.g. RefSeq cDNAs with accession prefix “NM_” matching RefSeq proteins with “NP_” prefix), it

allowed the comparison of coding models generated by Exonerate for a given cDNA to those generated by Genewise using its counterpart protein. Models supported by dubious cow protein/cDNA evidence (e.g. cDNA fragments with wrongly annotated short open-reading frames) were removed manually on a case-by-case basis. In addition, coding models supported by cDNAs (generated by Exonerate) were filtered to keep those which passed a 98% identity and 98% coverage cut-off in the cDNAs' alignments to the genome. The Apollo software [15.] was used to visualise the results of filtering.

Where one protein sequence (and its corresponding cDNA sequence, if any) had generated more than one candidate 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, cow) data is referred to as the “Targetted stage” (Figure 2). This stage resulted in 24791 coding models (6866 built from 5881 cow proteins and 17925 built from 17274 cow cDNAs) which were taken through to the UTR addition stage.

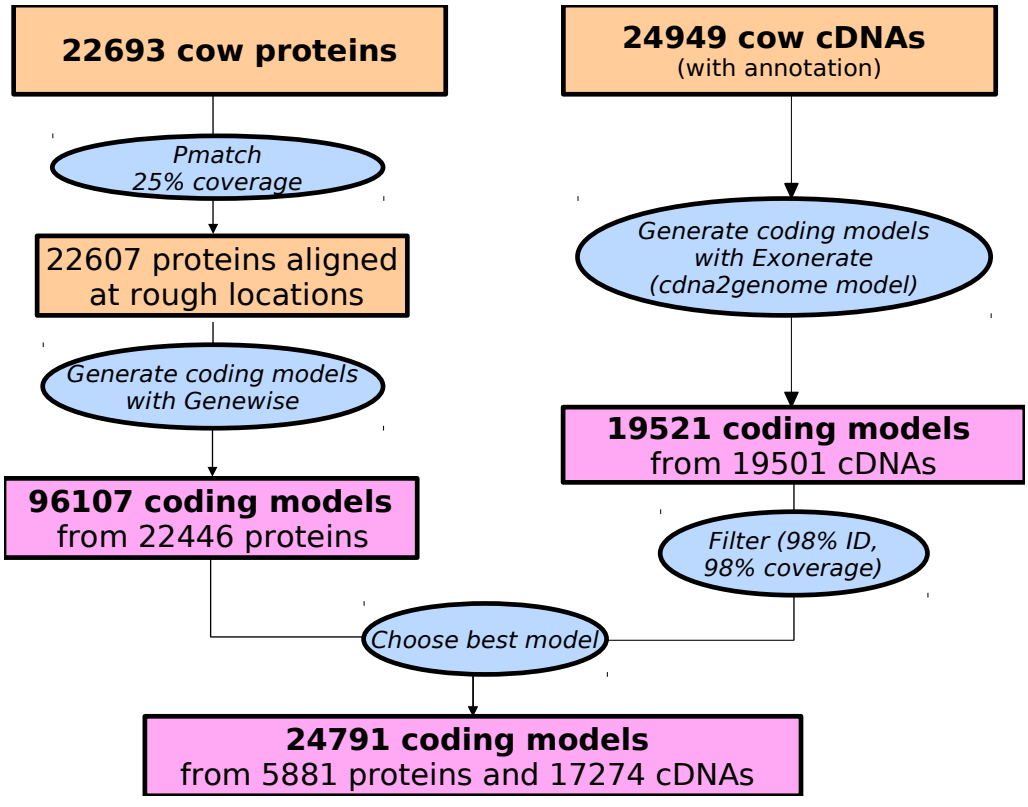

Figure 2: Targetted stage using cow protein and cDNA sequences.

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

Approximate time: 4-5 weeks

Following the cow Targetted alignments, additional coding transcript models were generated using data from related species in a process referred to as the "Similarity stage". The "Similarity" models were generated in three tiers. To start with, UniProt alignments from the Raw Computes step were filtered to retain only those sequences belonging to UniProt's "Mammalia" and "Vertebrata" taxonomical classes as well as Uniprot's Protein Existence (PE) classification level 1 and 2. For the first tier, in genomic regions which were not covered by any coding models from Targetted alignments, WU-BLAST was rerun for the Uniprot protein sequences and the results were passed to Genewise [13.] to build coding models. In most cases, multiple coding models built from different Uniprot proteins were generated in a single locus, each model with a slightly different exon-intron structure. To filter for the best supported structures, the TranscriptConsensus module was used to compare each Genewise model against cow cDNA and EST alignments in the region (see next section on how these alignments were generated), where exons in the Genewise model were scored for overlapping with exons of cDNA/EST alignments, and model(s) with the highest combined score in a region were kept.

Genewise models which passed the TranscriptConsensus filter were then further filtered based on the coverage of supporting evidence to ensure top quality Similarity models were taken as the first tier. For models supported by UniProt/SwissProt curated proteins, minimum coverage thresholds were set at 70% and 90% for single-exon and multi-exon models respectively. For models supported by UniProt/trEMBL proteins, the minimum coverage thresholds were more stringent and were set at 90% (single-exon) and 95% (multi-exon). Models which passed both the TranscriptConsensus and minimum-coverage filters were then used to "fill in" the gaps in genomic regions without Targetted alignments. The vast majority of "Similarity" models (20653) were generated this way in tier 1 [Figure 3].

While assessing the quality of Targetted and gap-filling Similarity models using Apollo, we noticed that some apparently truncated/split models could be "extended" or "bridged" by mammalian and vertebrate UniProt PE 1,2 proteins aligned to the entire genome (including regions which already have Targetted and gap-filling Similarity models) and filtered with a less stringent alignment coverage threshold. In addition, the relaxed coverage threshold would also recover some good-quality, multi-exon models missed when the coverage threshold for UniProt/trEMBL proteins was set at 95%. Therefore, in tier 2, UniProt proteins were aligned to the entire genome in exactly the same way as described for the gap-filling models, with the exception of setting the minimum coverage thresholds for both UniProt/SwissProt and UniProt/trEMBL evidence at 70% and 90% for single-exon and multi-exon models respectively. As a result, 4029 models were generated in tier 2 [Figure 3], which cover over 900 genomic regions were also covered by coding models for the first time, 663 extension cases and 94 bridging cases were identified.

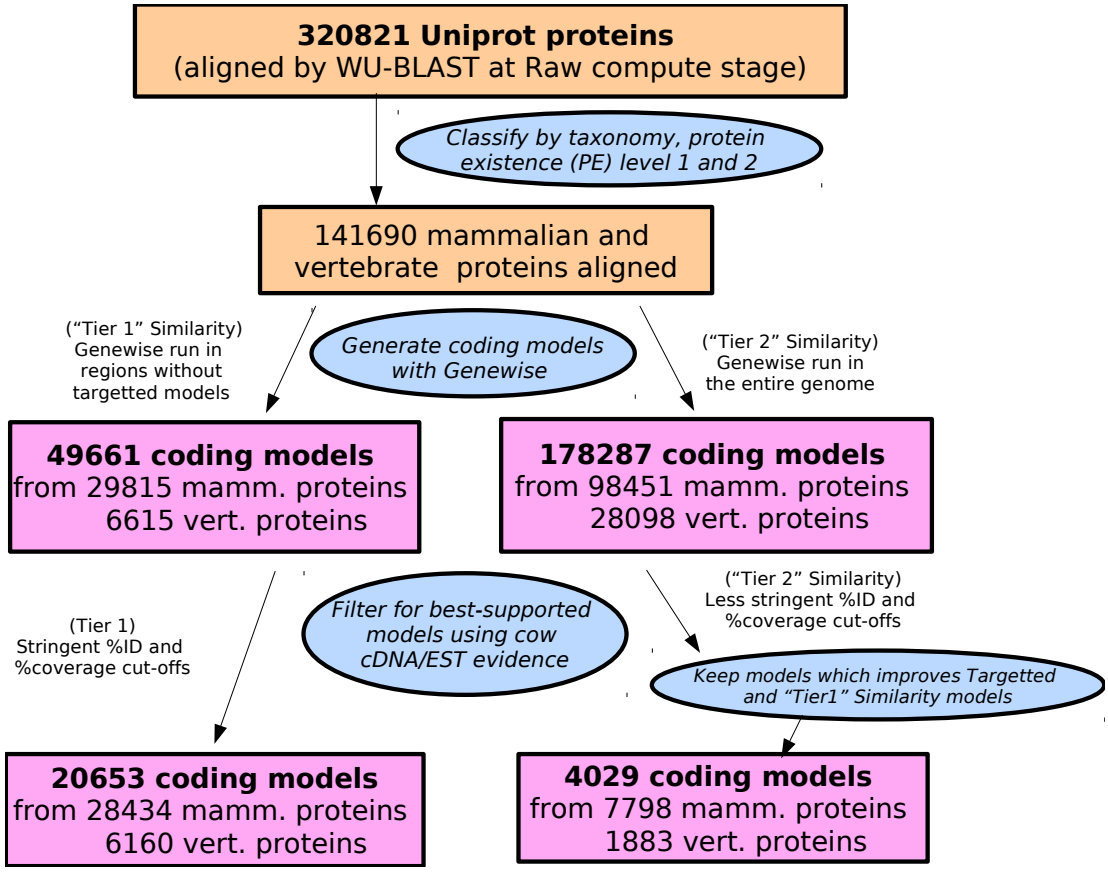

Figure 3: Alignment and filtering of mammalian and vertebrate UniProt proteins.

The last tier of Similarity models were generated by aligning the translations of Ensembl human genes which have one-to-one orthologues with human (Ensembl release 62 data) to the entire genome using Exonerate *protein2genome* model. The Exonerate models were filtered by a 80% identity and 95% coverage cut-off. As a result, 200 models (from 182 orthologues) which filled "gaps" without any Targetted/Similarity tiers 1-2 models or substantially extend/bridge long (e.g. 10-exon +) existing models were included in tier 3 [Figure 4].

Taking tiers 1-3 together, the Similarity stage resulted in 24682 coding models supported by Uniprot proteins (36232 mammalian proteins and 8043 vertebrate proteins) and 200 coding models supported by 182 human Ensembl translations. These similarity models were through to the UTR addition stage.

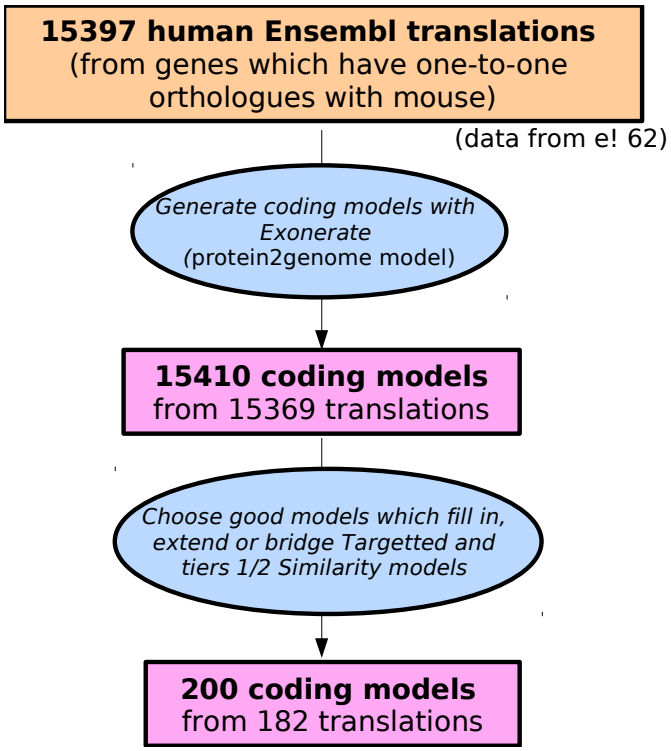

Figure 4: Alignment and filtering of human Ensembl translations

cDNA and EST Alignment

Approximate time: 2 weeks

Cow cDNAs and ESTs were downloaded from ENA/Genbank/DDBJ and RefSeq [9.], clipped to remove polyA tails, and aligned to the genome using Exonerate (*est2genome* model) [Figure 5].

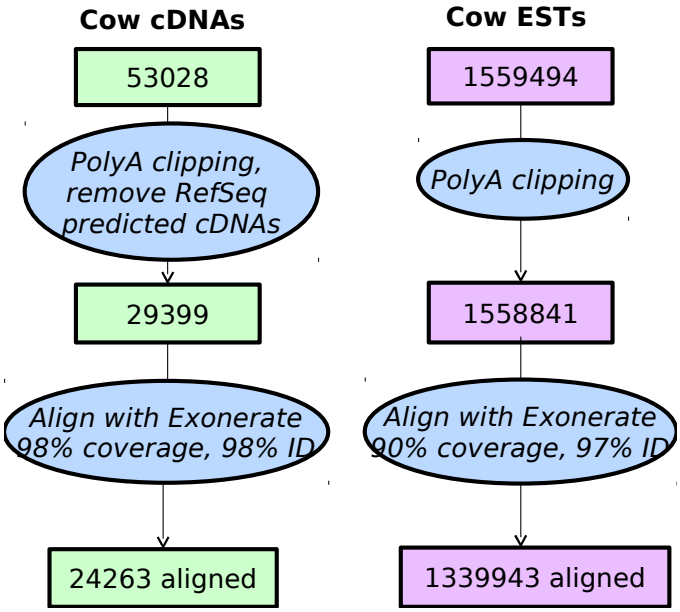

Figure 5: Alignment of cow cDNAs and ESTs to the cow genome.

24263 (of 29399) cow cDNAs aligned and 1339943 (of 1559494) cow ESTs aligned. A 98% cut-off was applied to the cDNA alignments for both sequence identity and coverage. The alignment filtering cut-offs for ESTs were set at 90% (coverage) and 97% (sequence 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.

Addition of UTR to coding models

Approximate time: 2 weeks

After finalising the set of coding models, those generated by Genewise alignments were extended into the untranslated regions (UTRs) using cow cDNAs. (Coding models generated by Exonerate's *cdna2genome* model

already contained UTR annotations and hence did not go through this UTR addition step.) Where available, RefSeq "NM" cDNA vs "NP" protein pairing information was used to ensure the correct matching of cDNAs to coding models supported by RefSeq proteins. This resulted in 5362 (of 6866) coding models from 4955 cow proteins with UTR, 24480 (of 24682) coding models from 19681 Uniprot mammalian/vertebrate proteins with UTR, and 120 (of 200) coding models from 120 Ensembl human tranlsations from with UTR.

Generating multi-transcript Ensembl genes

Approximate time: 3 weeks

The above steps generated a large set of potential transcript models (with or without UTR), 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. The resulting gene set contained 20778 genes, of which 11409 contained transcripts supported by cow cDNAs/proteins only (from the "Targetted" stage of the build), 8801 contained transcripts supported by Uniprot proteins only (from tiers 1 and 2 of the "Similarity" stage). [Figure 6]. The 20778 cow genes were associated with a total of 22902 transcripts, of which 12670 were supported by cow cDNAs/proteins, 10063 had support from Uniprot proteins, and 169 had Ensembl human translation as supporting evidence. [Figure 7].

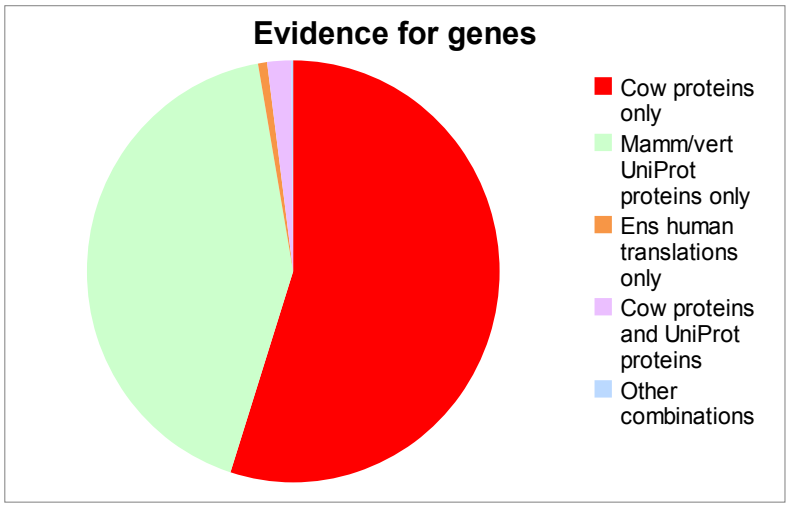

Figure 6: Supporting evidence for cow Ensembl gene set.

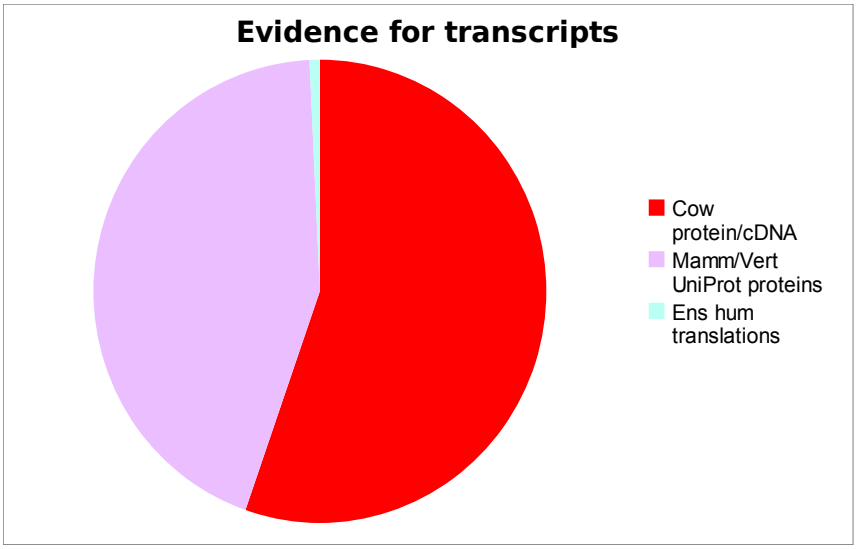

Figure 7: Supporting evidence for cow Ensembl transcript set.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 3 weeks

The gene set was screened for potential pseudogenes and retrotransposed genes. Also imported into the Ensembl gene set were annotation of mitochondrial genes in INDSC [18.] and short non-coding RNAs (e.g. miRNAs, snoRNAs) generated by the ncRNA pipeline [19.]. 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 on the Ensembl gene set

The main focus of the Ensembl automatic gene annotation pipeline is to generate a conservative set of protein-coding gene models, although some 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 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]
16. http://www.ensembl.org/info/docs/genebuild/ig_tcr.html
17. ftp://ftp.cines.fr/IMGT/IMGT.zip
18. http://www.ncbi.nlm.nih.gov/nuccore/NC_006853
19. http://www.ensembl.org/info/docs/genebuild/ncrna.html