

Ensembl gene annotation project (e!67)

Oreochromis niloticus (Nile tilapia)

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

Approximate time: 4 days

The annotation sdf process of the high-coverage tilapia 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 danio -s'), RepeatModeler [2] (version open-1.0.5, to obtain a repeats library, then filtered for an additional RepeatMasker run), Dust [3] and TRF [4]. Both executions of RepeatMasker and Dust combined masked 5.3 of the species genome.

Transcription start sites were predicted using Eponine-scan [5] and FirstEF [6]. CpG islands [Micklem, G.] longer than 400 bases and tRNAs [7] were also predicted. Genscan [8] was run across RepeatMasked sequence and the results were used as input for UniProt [9], UniGene [10] and Vertebrate RNA

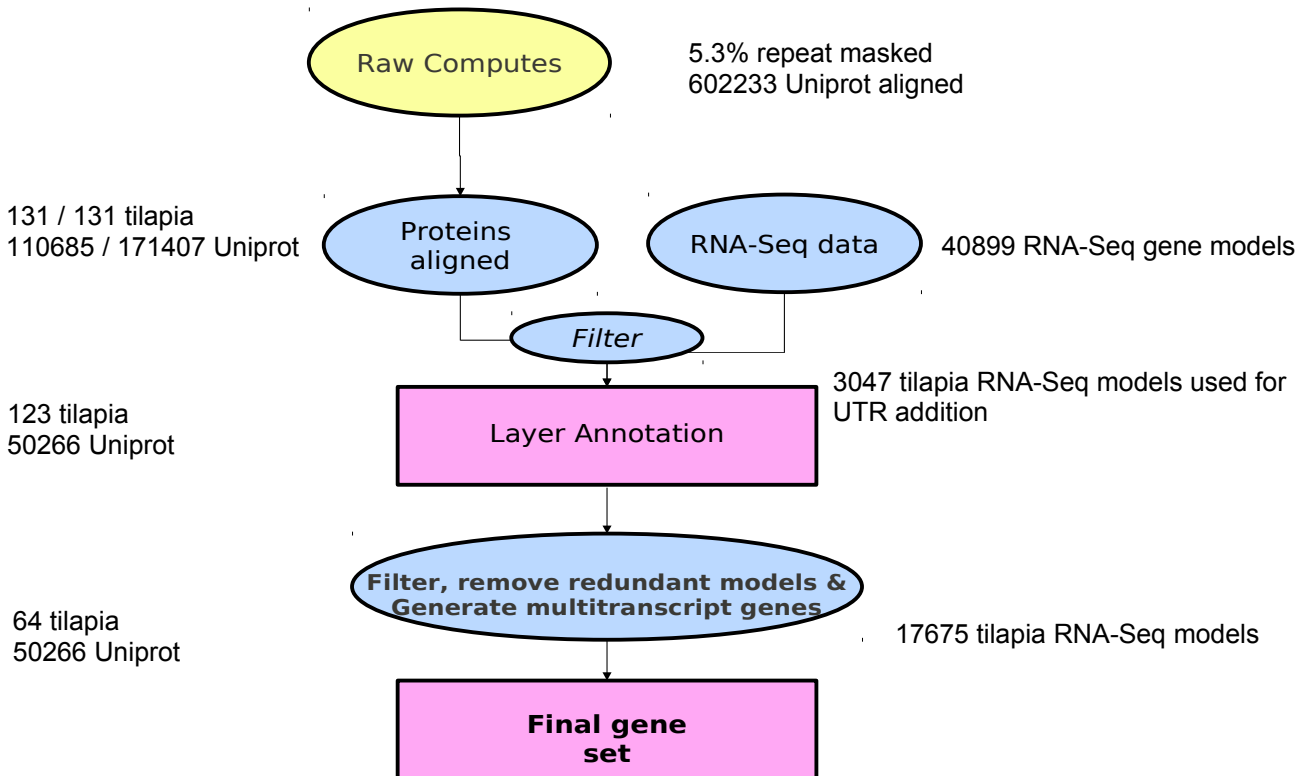


Figure 1: Summary of tilapia gene annotation project

[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 602233 UniProt, 346308 UniGene and 324070 Vertebrate RNA sequences aligning to the genome.

Exonerate Stage: Generating coding models from tilapia proteins

Approximate time: 1 week

Next, tilapia protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [9] and RefSeq [10]). The tilapia protein sequences were mapped to the genome using Pmatch as indicated in [Figure 2].

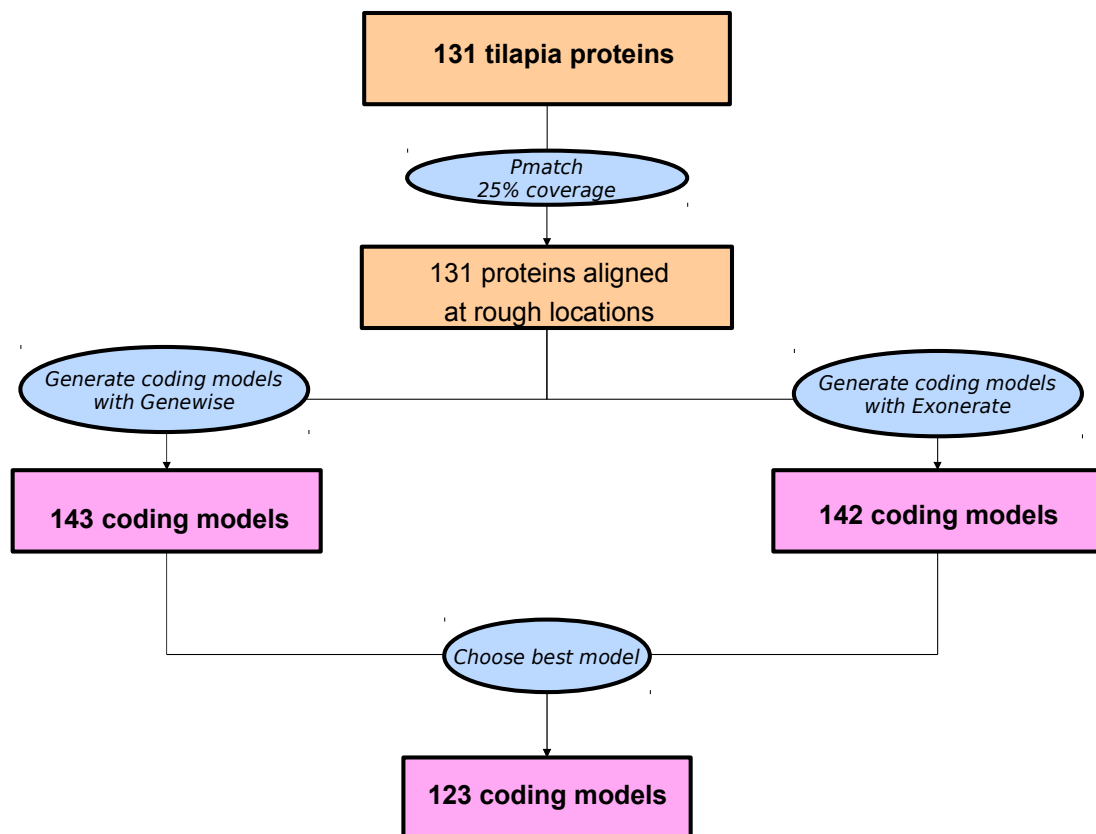


Figure 2: Targetted stage using tilapia protein sequences.

Models of the coding sequence (CDS) were produced from the proteins using Genewise [14] and Exonerate [13]. 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 data is referred to as the “Targetted stage”. This stage resulted in 123 (of 131) tilapia proteins used to build coding models to be taken through to the UTR addition stage.

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

Approximate time: 2 weeks

Following the tilapia 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 [14] 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 110685 coding models.

cDNA and EST Alignment

Approximate time: 2 days

Tilapia cDNAs and ESTs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate [Figure 3].

Of these, 35(of 35) tilapia cDNAs aligned, and 113606 (of 119531) tilapia ESTs aligned. EST and cDNAs alignments are displayed on the website in a separate track from the Ensembl gene set.

RNA-Seq models

Approximate time: 2 months

RNA-Seq data provided by the Broad Institute was used in the annotation. This comprised paired end data from 11 tissues type: blood, brain, embryo, eye, heart, kidney, liver, muscle, ovary, skin, testis. The 700,016,850 available reads from all the tissues were aligned to the genome using BWA, resulting in 548,573,947 reads aligning. Subsequently, the Ensembl RNA-Seq pipeline was used to process the BWA alignments and create a further 95,651,450 split read alignments using Exonerate. The split reads and the processed

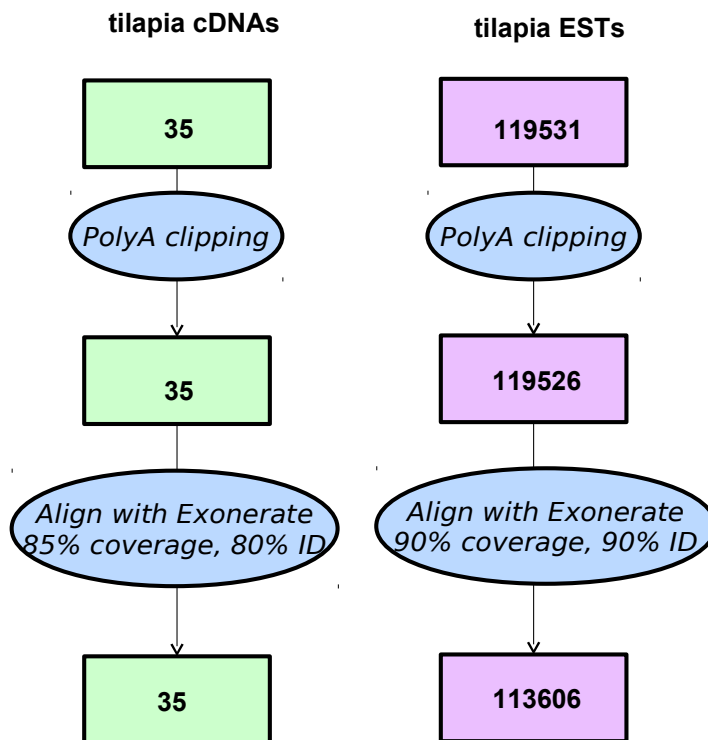


Figure 3: Alignment of tilapia cDNAs and EST to the tilapia genome

BWA alignments were combined to produce 40899 transcript models in total. 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 split into a separate class.

Filtering Coding Models

Approximate time: 1 week

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus and LayerAnnotation. The Apollo software [16] 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 tilapia RNA-Seq data. This resulted in 4 (of 123) tilapia coding models with UTR and 6293 (of 50266) UniProt coding models with UTR. All RNA-Seq

models acquired their UTR during the RNA-Seq Pipeline.

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 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 final gene set of 21462 genes included 26 genes with at least one transcript supported by tilapia proteins. The remaining 21436 genes had transcripts supported by proteins from other sources [Figure 4].

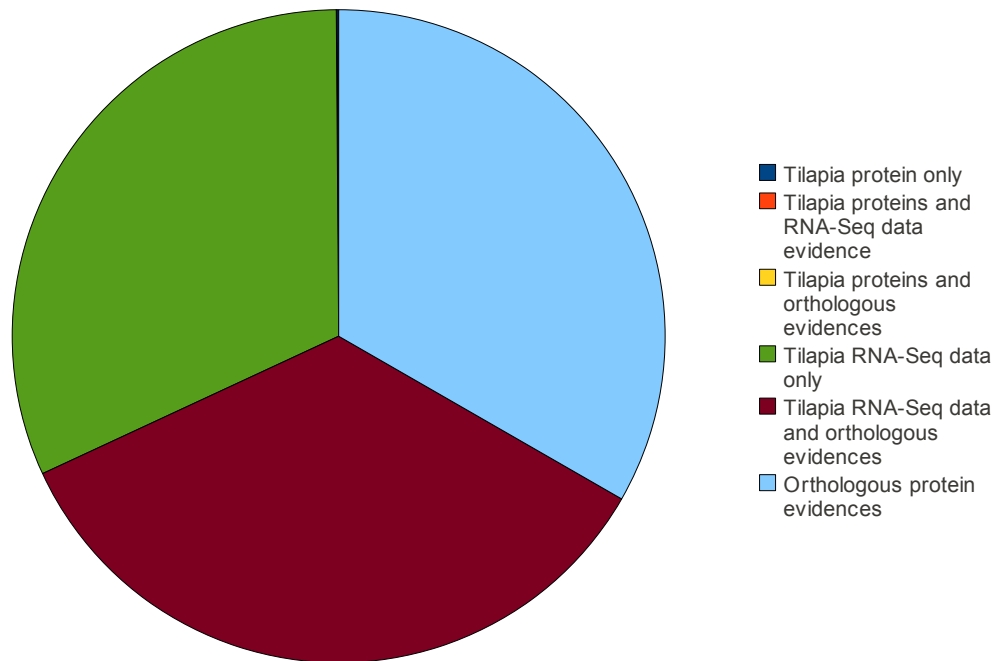


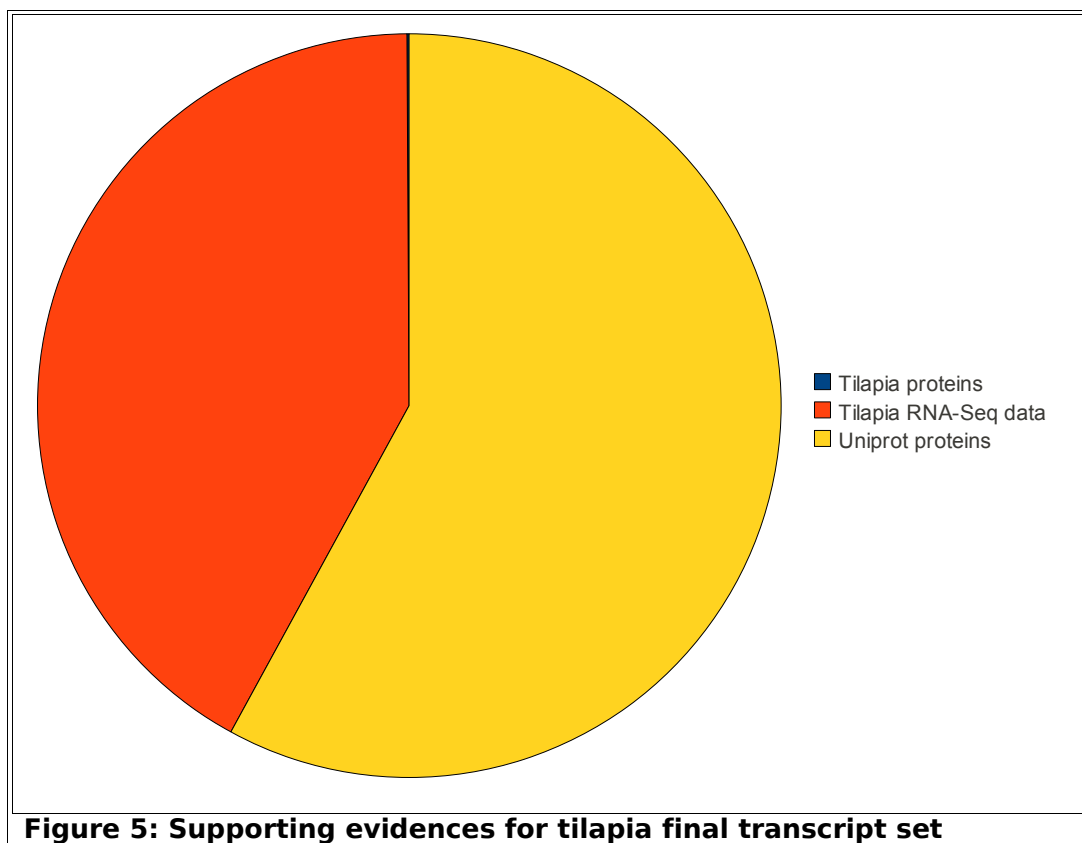
Figure 4: Supporting evidence for tilapia final gene set

The final transcript set of 27609 transcripts included 26788 transcripts with support from 26 tilapia proteins, 11235 transcripts with support from RNA-Seq data and 15527 transcripts with support from UniProt SwissProt [Figure 5].

Pseudogenes, Protein annotation, Cross-referencing, Stable identifiers, non coding RNAs

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.)

Small structured non-coding genes were added using annotations taken from RFAM [17] and miRBase [18].

The final gene set consists of 21437 protein coding genes, these contain 26763 transcripts. A total of 22 pseudogenes were identified, 3 retrotransposed and 821 ncRNAs.

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 Nile tilapia gene annotation summary this pipeline is to generate a conservative set of protein-coding gene models, although non coding 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 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/-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 Smit, AFA, Hubley, R: **RepeatModeler Open-1.0**. 2008-2010. www.repeatmasker.org
- 3 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.
- 4 Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. Nucleic Acids Res. 1999, **27(2)**:573-580. <http://tandem.bu.edu/trf/trf.html> [PMID: [9862982](#)]
- 5 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](#)]
- 6 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](#)]
- 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 Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. BMC Bioinformatics 2005, **6**:31. [PMID: [15713233](#)]
- 14 Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. Genome Res. 2004, **14(5)**:988-995. [PMID: [15123596](#)]
- 15 Eyras E, Caccamo M, Curwen V, Clamp M: **ESTGenes: alternative splicing from ESTs in Ensembl**. Genome Res. 2004 **14(5)**:976-987. [PMID: [15123595](#)]
- 16 Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglu 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](#)]
- 17 Griffiths-Jones S., Bateman A., Marshall M., Khanna A., Eddy S.R: **Rfam: an RNA family database**. Nucleic Acids Research (2003) **31(1)**:p439-441. [PMID: [12520045](#)]
- 18 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 [PMID: [16381832](#)]
- 19 Wilming L. G., Gilbert J. G. R., Howe K., Trevanion S., Hubbard T. and Harrow J. L: **The vertebrate genome annotation (Vega) database**. Nucleic Acid Res. 2008 Jan; Advance Access published on November 14, 2007; doi:10.1093/nar/gkm987 [PMID: [18003653](#)]