Ensembl gene annotation project

Pelodiscus sinensis (Chinese soft-shell turtle)

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

The annotation process of the high-coverage Chinese soft-shell turtle 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 "pelodiscus_sinensis" -s`'), RepeatModeler [4.] (version open-1.0.5, to obtain a repeats library, then filtered for an additional RepeatMasker run), Dust [2.] and TRF [3.]. Combination of all repeat analyses, RepeatMasker, RepeatModeler, Dust and TRF brings the total proportion of the masked genome to 43.59%.

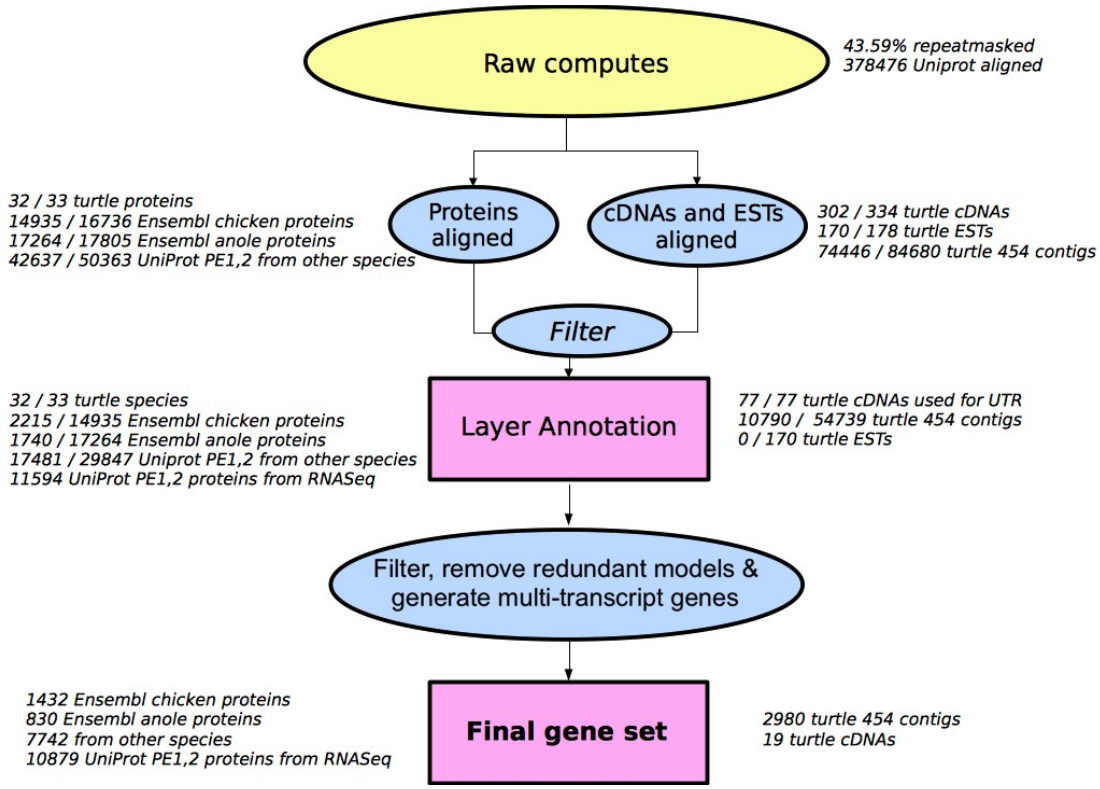


Figure 1: Summary of turtle gene anntation project.

Transcription start sites were predicted using Eponine–scan [5.] and FirstEF [6.]. CpG islands 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 [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 378476 UniProt, 328450 UniGene and 322092 Vertebrate RNA sequences aligning to the genome.

Targeted Stage: Generating coding models from Chinese soft-shell evidence

Next, turtle protein sequences were downloaded from public databases, UniProt SwissProt/TrEMBL [9.] and RefSeq [10.]. The turtle 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 [14.] and Exonerate [13.]. The generation of transcript models using turtle specific data is referred to as the “Targeted stage”. This stage resulted in 32 of the 33 turtle proteins used to build coding models. However, none of these models were used in subsequent analyses as they were overridden with longer models from the Similarity stage.

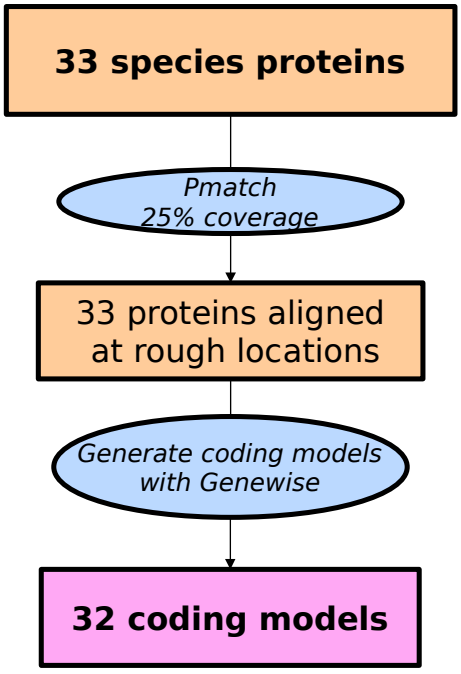


Figure 2: Targetted stage using turtle protein sequences.

cDNA and EST Alignment

Turtle cDNAs and ESTs were downloaded from GenBank, clipped to remove polyA tails, and aligned to the genome using Exonerate. Of 334 turtle cDNAs, 216 sequences aligned while 142 of the 178 ESTs aligned. The cutoffs for both data sets were 90% coverage and 90% identity. Contig sequences generated by the Chinese softs-shell turtle Consortium using 454 sequencing method were also aligned to the genome. Of 84680 initial set, 54739 aligned with a cut-off of 90% coverage and 95% identity [Figure 3].

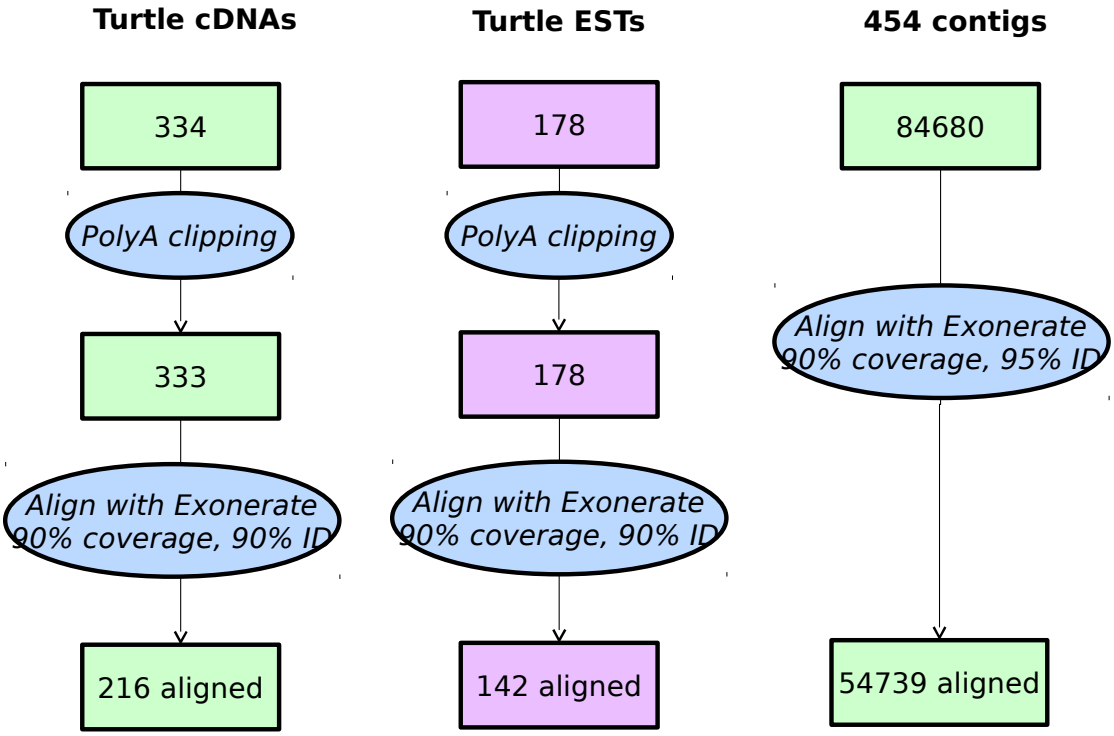


Figure 3: Alignment of turtle cDNAs and ESTs, and 454 contigs to the turtle genome.

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

Due to the small number of turtle specific protein and cDNA evidence the majority of the gene models were based on proteins from other species. 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 53646 coding models [Figure 4].

Alignment of Ensembl chicken and anole lizard translations

Ensembl chicken and anole lizard translations were aligned against the turtle genome. The cutoff values for coverage and identity were set at 80% and 60% respectively. Of the chicken translations, 14935 of the 16736 retrieved translations aligned. From the 17805 lizard translations, 17264 sequences aligned above the set thresholds. The resulting coding models were taken through to the all subsequent steps [Figure 4].

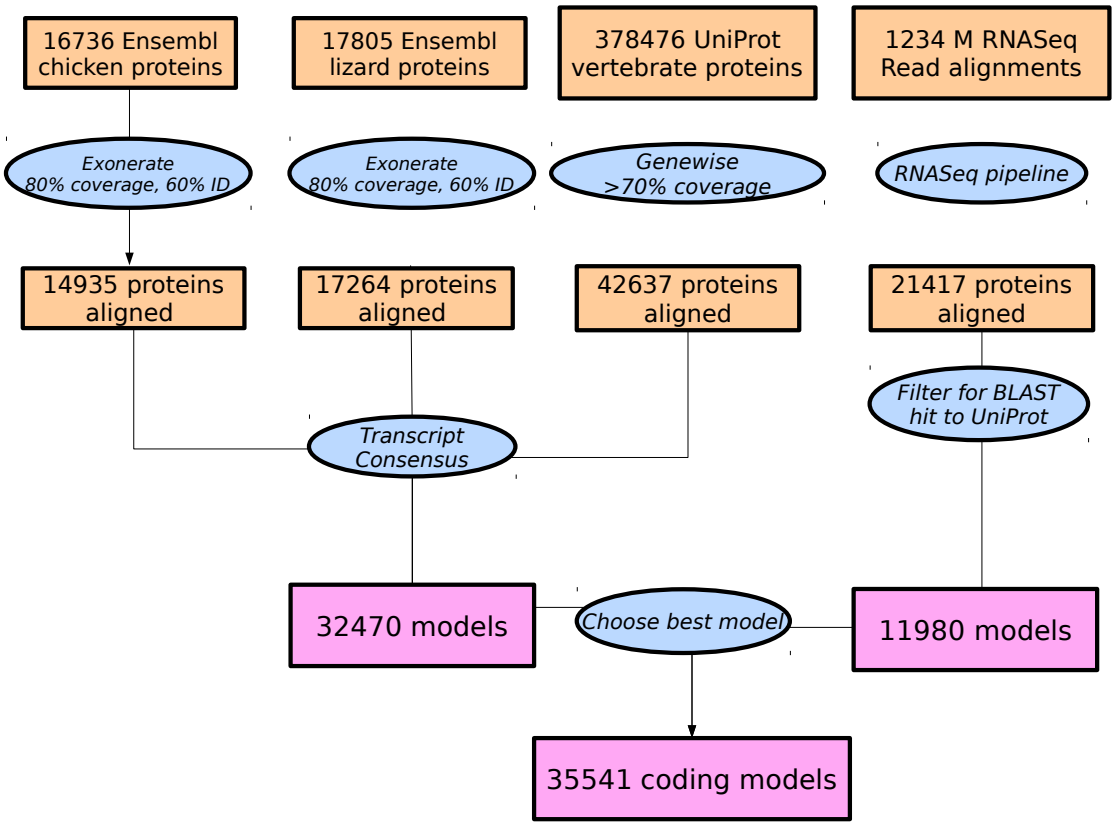


Figure 4: Alignment and filtering of other species proteins and addition of RNASeq models.

Filtering Coding Models

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus and LayerAnnotation. RNA-Seq spliced alignments supporting introns were used to help filter the set. The Apollo software [16.] was used to visualise the results of filtering.

Addition of RNA-Seq models

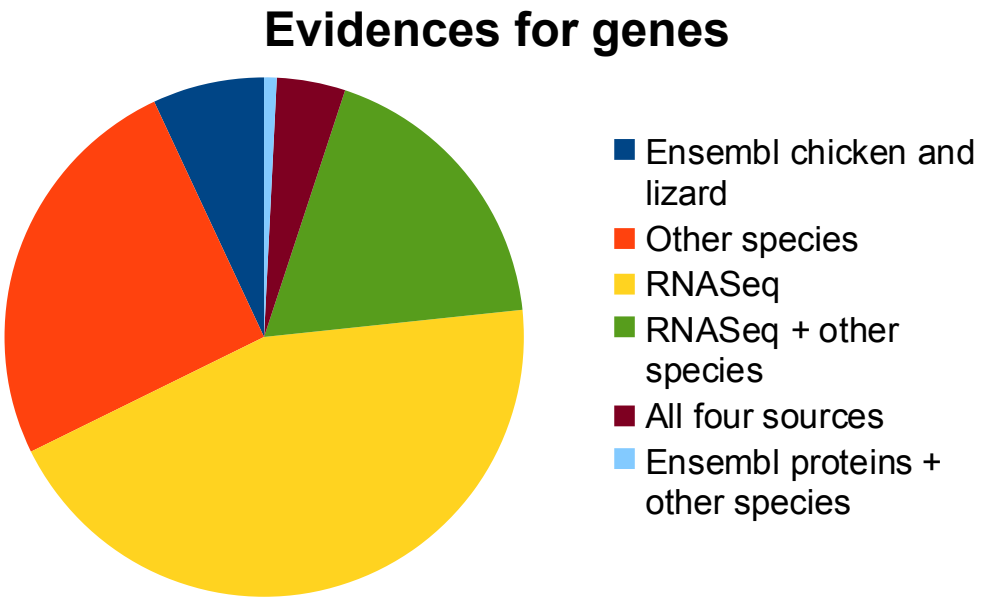
The largest set of turtle specific evidence was from paired end RNASeq, this was used where appropriate to help inform our gene annotation. A set of 1.2 billion reads that passed QC were aligned to the genome using BWA resulting in 1.1 billion (87.6%) reads aligning and properly pairing. The Ensembl RNASeq pipeline was used to process the BWA alignments and create a further 120 million split read alignments using Exonerate. The split reads and the processed BWA alignments were combined to produce 21417 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 WUBLAST, 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 10892 models were added.

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using turtle cDNA and contigs from the 454 sequencing project. This resulted in 5935 of 32470 coding models with UTR. In addition, 10892 RNASeq models also contributed to the UTR addition of the final 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 of 18272 genes included 4603 genes built only using proteins from other species and 8070 genes built only from RNASeq evidence. 3322 genes had a mixture of RNASeq and evidence from other species proteins. A further 1263 genes were supported only by Ensembl chicken or Ensembl lizard translations. The remaining 917 genes contained transcripts from all four sources [Figure 5].



The final set of 20752 transcripts included 12384 transcripts with support from RNASeq evidence, 8616 transcripts with support from other species proteins and 2236 transcripts with support from Ensembl chicken or lizard data [Figure 6]. A small set of the transcripts, 2581, were supported by evidences from two sources.

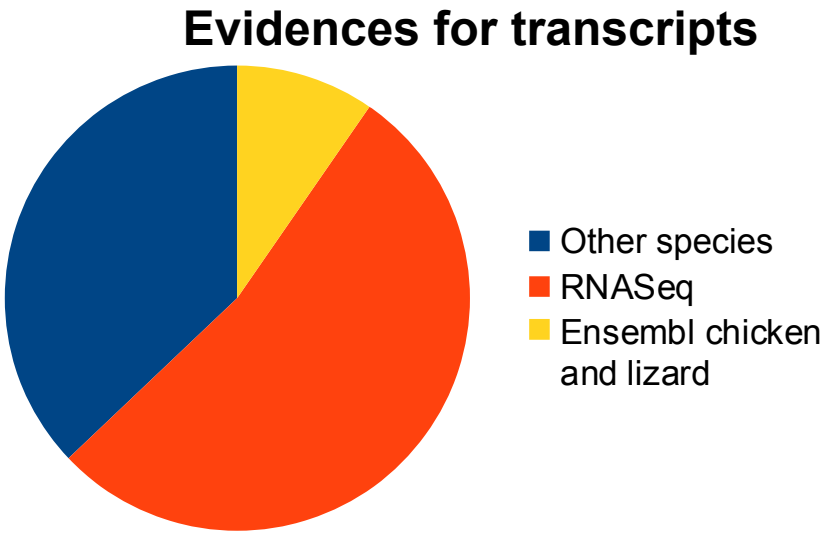


Figure 6: Supporting evidence for turtle final transcript set.

Pseudogenes, non-coding genes, 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 [17.] and miRBase [18.].

The final gene set consists of 18188 protein coding genes including mitochondrial genes, these contain containing 20752 transcripts. A total of 97 pseudogenes were identified and 1018 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 this pipeline is to generate a conservative set of protein-coding gene models, although noncoding 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=log

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. Smit, AFA, Hubley, R . **RepeatModeler Open-1.0.** 2008-2010. http://www.repeatmasker.org
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 Bioinformatic*s 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, 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]
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