Ensembl gene annotation project (*e!70*)

Felis catus (cat, Felis_catus-6.2)

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 Cat 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 "cat" -s`'), Dust [2] and TRF[3]. Both executions of RepeatMasker and Dust combined masked 39.4% of the species genome.

Transcription start sites were predicted using Eponine–scan [4] and FirstEF [5]. CpG islands longer than 400 bases and tRNAs [6] were also predicted. Genscan [7] was run across RepeatMasked sequence and the results were

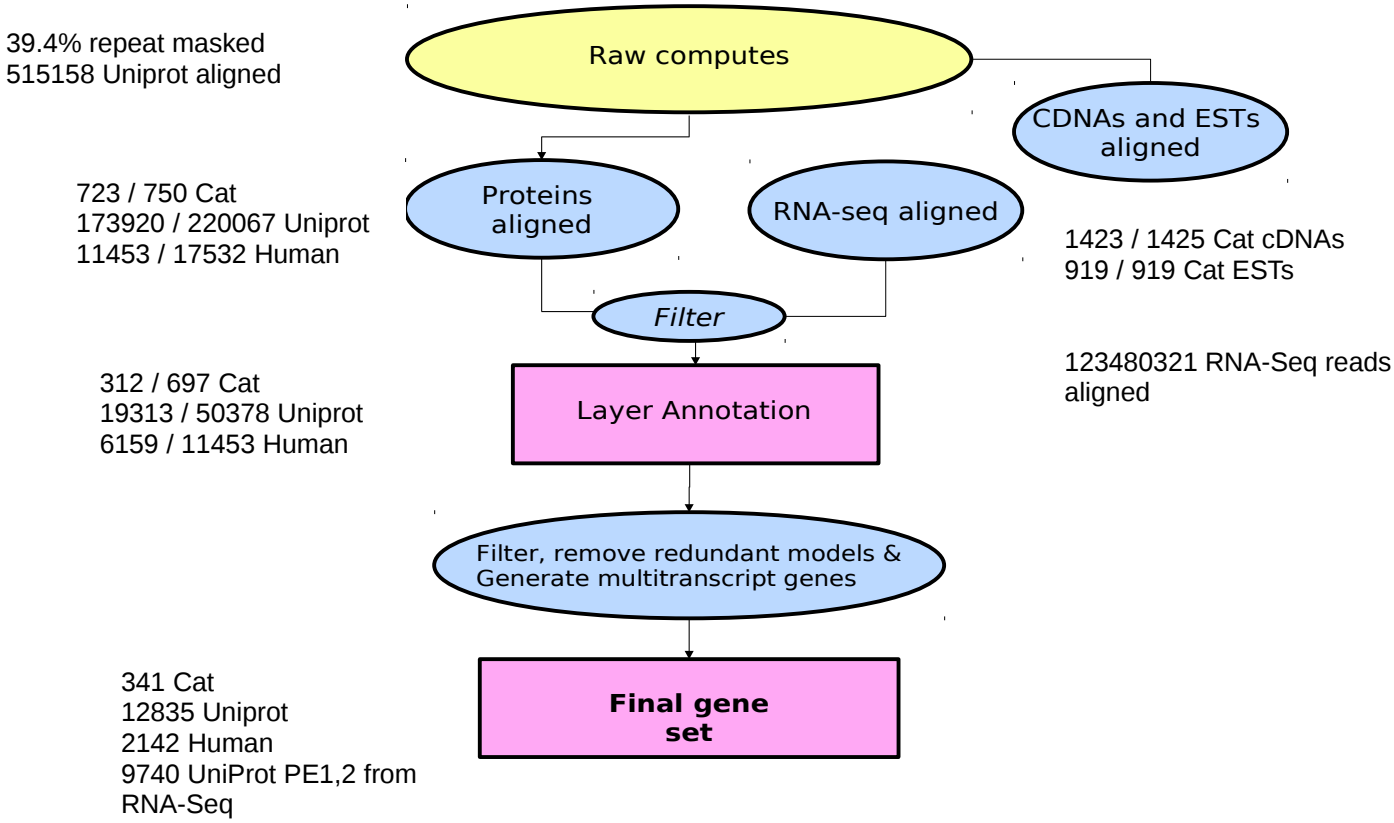

Figure 1: Summary of Cat gene annotation project

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 515158 UniProt, 342654 UniGene and 333476 Vertebrate RNA sequences aligning to the genome.

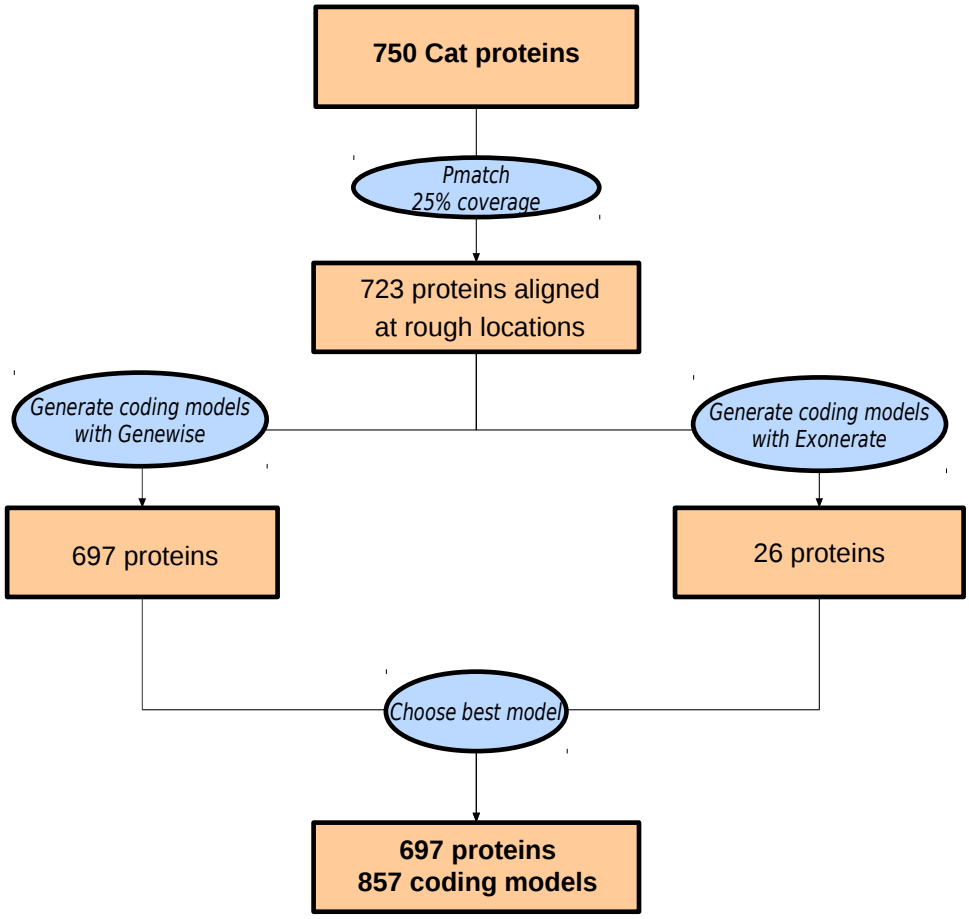

Figure 2: Targeted stage using Cat protein sequences.

Targetted Stage: Generating coding models from Cat evidence

Approximate time: 2-3 days

Next, Cat protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8] and RefSeq [9]). The 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 Cat-specific data is referred to as the "Targeted stage". This stage resulted in a selection of 697 (of 750) Cat proteins.

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

Approximate time: 1 week

Following the 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 [13] 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 173920 coding models.

In addition to these, 11453 (of 17532) Human proteins were aligned to the Cat genome using Exonerate. These came from the longest translation of each human protein coding gene in Ensembl 65. These Human proteins along with the 697 Cat proteins from the previous stage together produced 12310 coding models that were taken through to the UTR addition stage.

cDNA and EST Alignment

Approximate time: 2-3 days

Cat 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 1423(of 1425) Cat cDNAs aligned, and 919 (of 919) Cat ESTs aligned. All alignments were at a cut-off of 90% coverage and 80% identity.

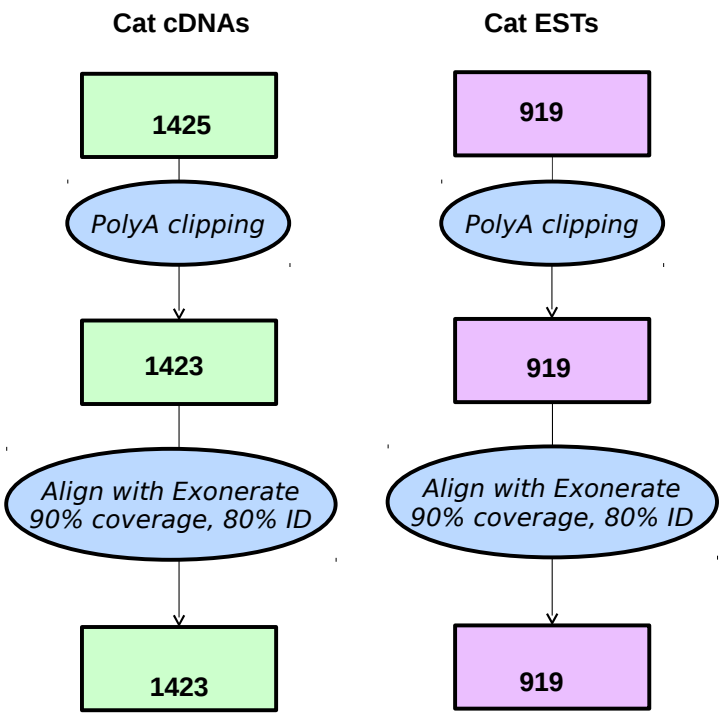

Figure 3: Alignment of Cat cDNAs and ESTs to the Cat genome

RNA-Seq models

Approximate time: 1-2 weeks

RNA-Seq data provided by The Texas A&M College of Veterinary Medicine & Biomedical Sciences was used in the annotation. This was composed of unpaired 50bp reads of testis samples from four individuals. The available reads were aligned to the genome using BWA, resulting in 123480321 reads aligning. The processed BWA alignments produced 25709 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 removed from the pipeline system. Models with good BLAST alignments were kept and used in the layer annotation stage. These were also used along with models that had a moderately good BLAST alignment, in the UTR addition stage.

Filtering Coding Models

Approximate time: 2 days

Coding models from the Similarity stage and from the RNA-seq data were filtered using modules such as TranscriptConsensus and LayerAnnotation. The Apollo software [15] was used to visualise the results of filtering.

Addition of UTR to coding models

Approximate time: 1 day

The set of coding models was extended into the untranslated regions (UTRs) using Cat RNA-Seq data. This resulted in 312 (of 697) Cat coding models with UTR, 6159 (of 11453) Human coding models with UTR, and 19313 (of 50378) UniProt coding models with UTR.

Generating multi-transcript genes

Approximate time: 1-2 days

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

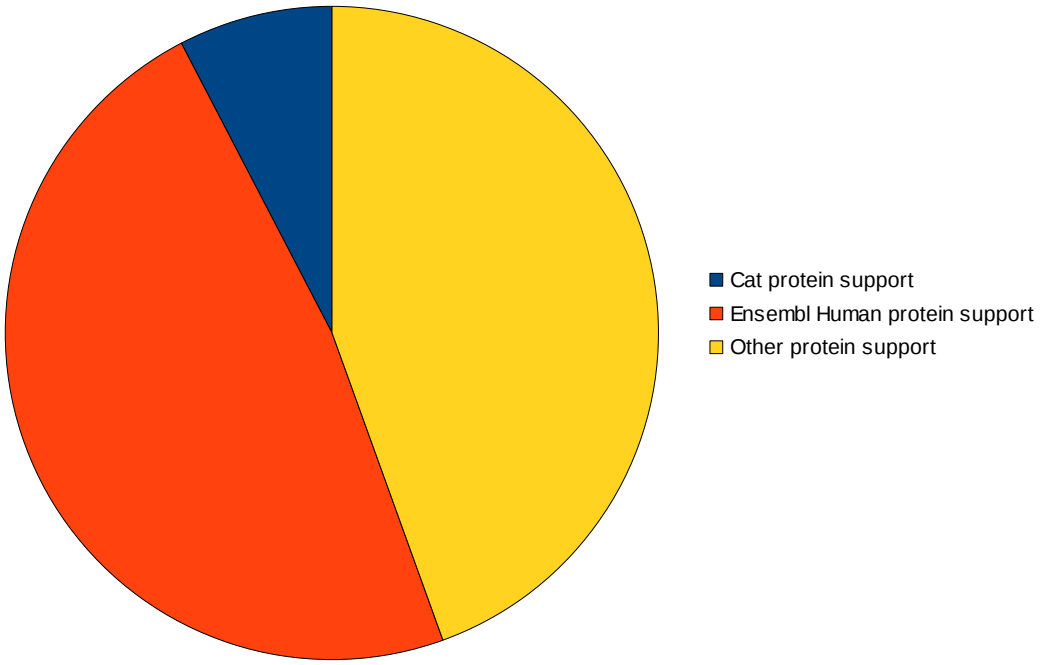

Figure 4: Supporting evidence for Cat final gene set

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 preliminary gene set of 21890 genes included 328 genes with at least one transcript supported by Cat proteins, 2072 genes with at least one transcript supported by Human evidence and the remaining 19490 genes had transcripts supported by proteins from other sources [Figure 4].

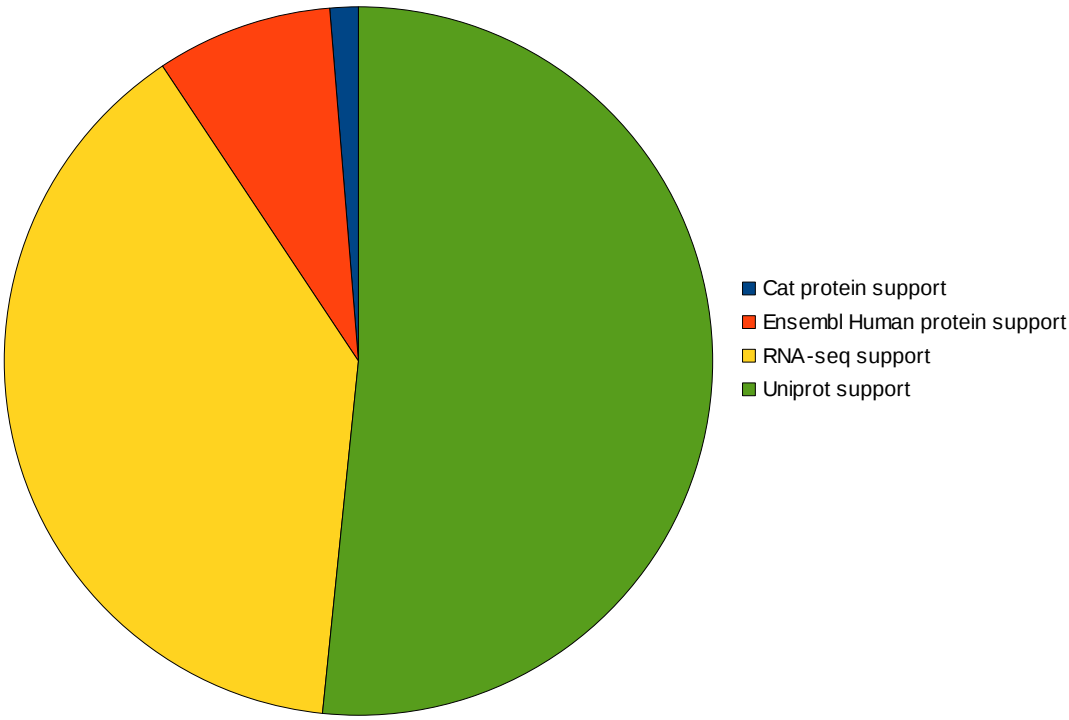

Figure 5: Supporting evidences for Cat final transcript set

The transcript set of 22656 transcripts included 338 transcripts with support from Cat proteins, 2109 transcripts with support from Human translations, 10411 transcripts with support from RNA-Seq data and 13402 transcripts with support from UniProt SwissProt [Figure 5].

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 1-2 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 [16] and miRBase [17].

The final gene set consists of 19493 protein coding genes, including mitochondrial genes, these contain 20259 transcripts. A total of 542 pseudogenes were identified and 1831 ncRNAs. Of the protein coding transcripts 13 transcripts were mitochondrial.

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 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 of 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 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 Bioinformatics* 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 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]

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 [PMID: 16381832]

18 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]