Ensembl gene annotation project (*e!74*)

Ovis aries (Sheep)

This document describes the annotation process of the high-coverage sheep assembly, described in Figure 1. The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.

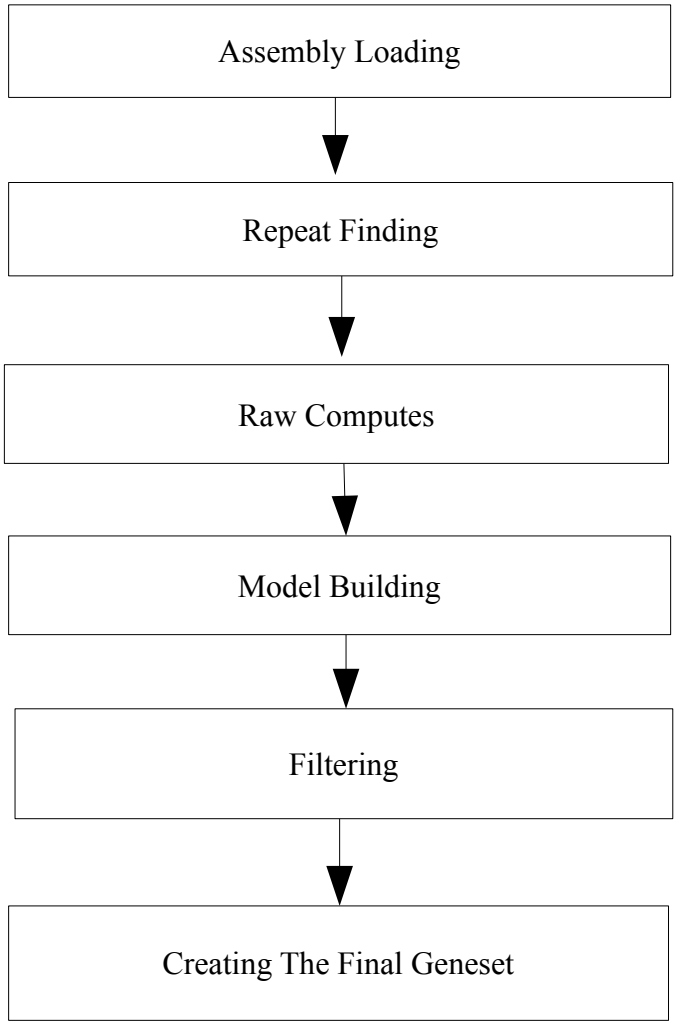

Figure 1: The Gene Annotation Pipeline

Repeat Finding

After loading into a database, the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.2.8 with parameters '`-nolow -species "ovis_aries" -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 47.3% of the species genome.

Raw Computes

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. The results of Eponine-scan, FirstEF, CpG, and tRNAscan are for display purposes only; they are not used in the gene annotation process.

Genscan [8] was run across repeat-masked 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 7,815,610 UniProt, 9,367,621 UniGene and 8,938,147 Vertebrate RNA sequences aligning to the genome.

cDNA and EST Alignments

Sheep cDNAs and ESTs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate (Table 1). These alignments provide supporting evidence for models.

Species	Type	Sequences Downloaded	Sequences Aligned
sheep	cDNA	25,477	2,420
	EST	338,483	264,124

Table 1: cDNA/EST alignments

All alignments were at a cut-off of 90% coverage and 97% identity.

Model Generation

Various sources of transcript and protein data were investigated and used to generate gene models using a variety of techniques. The data and techniques employed to generate models are outlined here. The numbers of gene models generated are described in Table 2.

Pipeline	Source	Number of Models
Targeted	1,001 UniProt sheep proteins 793 RefSeq sheep proteins	1,883
Similarity	350,188 UniProt vertebrate only proteins	66,797
RNASeq	International Sheep Genome Consortium (ISGC)	23,357
Ensembl Longest Translations	22,529 Ensembl Release 69 proteins for human 19,994 Ensembl Release 69 proteins for cow	22,317 20,335

Table 2: Gene Model Generation Overview

Targeted Pipeline: Generating coding models using species specific proteins

Protein sequences for sheep were downloaded from public databases (UniProt SwissProt/TrEMBL [9] with Protein Existence (PE) classification level 1 or 2 and RefSeq [10]). The sheep protein sequences were mapped to the genome using Pmatch set at a low threshold (-T 14). Two sets of coding models were then produced from the proteins using Exonerate [13] and Genewise [14].

Where one protein sequence had generated more than one coding model at a locus, the BestTargeted 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. This pipeline is shown in Figure 2.

Similarity Pipeline: Generating coding models using proteins from related species

Coding models were generated using data from related species. The UniProt alignments from the Raw Computes step were filtered and only vertebrate sequences were kept. WU-BLAST was rerun for these sequences and the results were passed to Genewise [14] to build coding models.

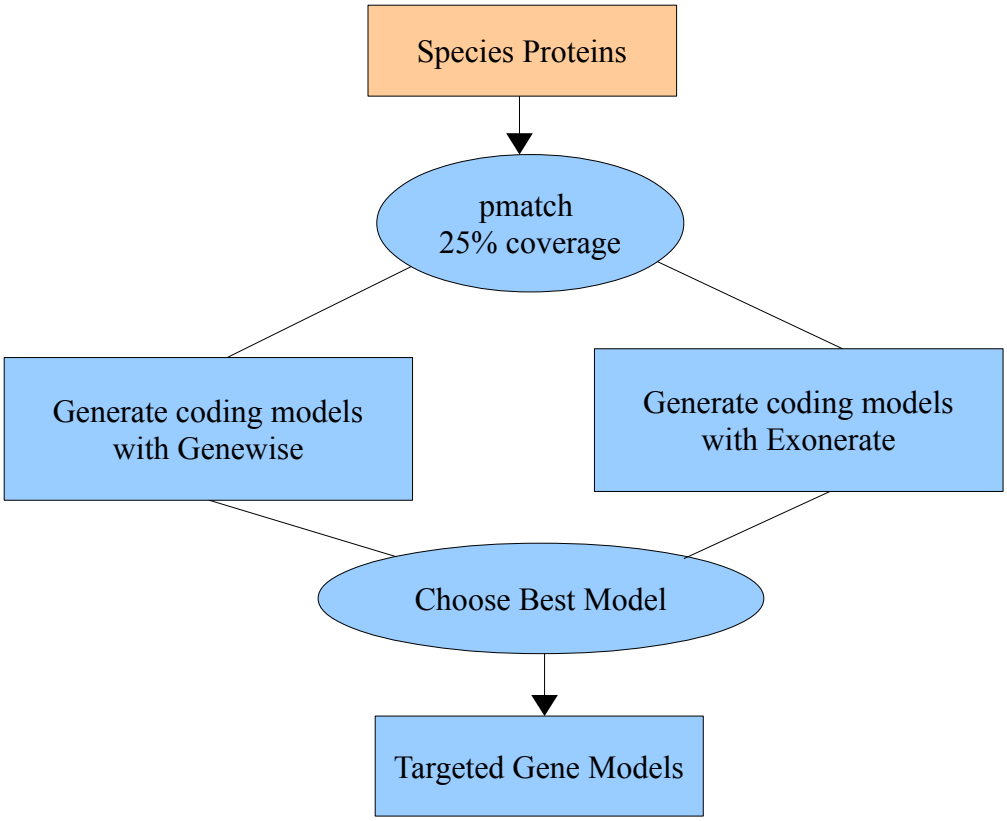

Figure 2: Targeted Pipeline

RNASeq Pipeline

RNASeq data provided by the International Sheep Genome Consortium (ISGC) was used in the annotation. This comprised paired-end data from a pool of 94 tissues samples including: a range of different tissue types between a trio (ram, ewe and lamb), 7 tissue types from the reference sheep and tissue types from different breeds (Table 3). The available reads were aligned to the genome using BWA. The Ensembl RNASeq pipeline was used to process the BWA alignments and create further split read alignments using Exonerate.

The split reads and the processed BWA alignments were combined to produce 25,832 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 poorly scoring or no BLAST alignments were split into a separate class.

Ensembl Longest Translations

The longest translation for each protein coding gene in Ensembl proteins release 69 for human and cow were downloaded. These proteins were

aligned against the sheep genome using Exonerate [13] to produce a set of coding models.

Filtering the Models

The filtering phase decided the subset of protein-coding transcript models, generated from the model-building pipelines, that comprise the final protein-coding gene set.

Models were filtered using the TranscriptConsensus, LayerAnnotation and GeneBuilder modules.

Apollo software [16] was used to visualise the results of filtering.

LayerAnnotation

The LayerAnnotation module was used to define a hierarchy of input data sets, from most preferred to least preferred. The output of this pipeline included all transcript models from the highest ranked input set. Models from lower ranked input sets are included only if their exons do not overlap a model from an input set higher in the hierarchy (Figure 3).

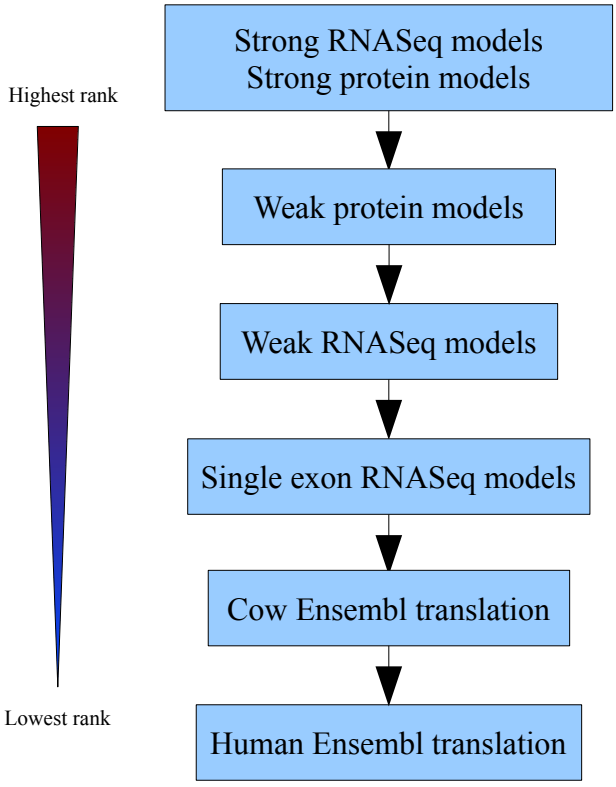

Figure 3: Layer annotation hierarchy

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNASeq, cDNA and EST sequences. At the UTR addition stage 36,717 gene models out of total of 66,797 non-RNASeq pipeline generated gene models had UTR added.

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 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.

At this stage the gene set comprised 21,715 genes with 23,630 transcripts.

Pseudogenes

The Pseudogene module was run to identify processed pseudogenes from within the set of gene models – these were labelled as pseudogenes. A total of 291 genes were labelled as pseudogenes.

Creating The Final Gene Set

ncRNAs

Small structured non-coding genes were added using annotations taken from RFAM [17] and miRBase [18]. WU-BLAST was run for these sequences and models built using the Infernal software suite [20].

Cross-referencing

Before public release the transcripts and translations were given external references (cross-references to external databases). Translations were searched for signatures of interest and labelled where appropriate.

Stable Identifiers

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.

Final Gene Set Summary

The final gene set consists of 20,921 protein coding genes, including 13 mitochondrial genes. These contain 22,823 transcripts. A total of 291 pseudogenes were identified. 3,985 ncRNAs were added by the ncRNA pipeline, of which 24 are mitochondrial.

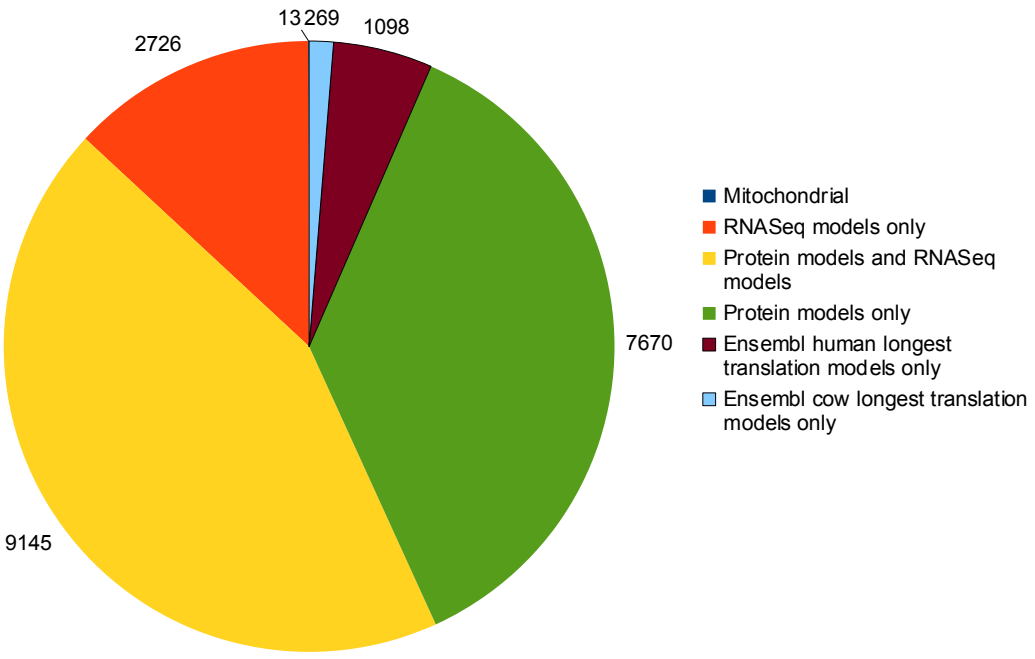

Figure 4: Supporting evidences for the protein coding gene models

Institute	Tissue type	Total number of reads	Number of aligned reads
Roslin	Ram abomasum mucosa	99,005,126	89,344,177
	Ram adrenal gland	63,698,910	57,617,837
	Ram alveolar macrophages	85,657,840	75,769,110
	Ram brain stem	86,771,358	79,294,162
	Ram caecum	149,878,848	131,716,966
	Ram cerebellum	83,204,438	77,644,934
	Ram cerebrum	89,357,620	83,035,375
	Ram colon	70,758,896	62,435,422
	Ram duodenum	96,591,908	87,338,756
	Ram hypothalamus	81,489,222	74,657,046
	Ram kidney cortex	88,866,990	81,109,519
	Ram kidney medulla	103,387,408	94,911,815
	Ram liver	74,980,724	65,111,956
	Ram lung	87,609,566	78,522,629
	Ram lymph node mesenteric	62,223,930	52,011,490
	Ram lymph node prescapular	51,429,782	43,871,301
	Ram muscle biceps	73,206,112	64,324,999
	Ram muscle long dorsal	59,850,862	53,344,755
	Ram omentum	93,599,460	84,926,940
	Ram pituitary gland	97,467,102	90,701,897
	Ram rectum	81,292,246	73,778,212
	Ram rumen	126,008,160	112,040,264
	Ram skin back	93,021,642	83,694,859
	Ram spleen	88,293,714	80,975,545
	Ram testes	111,359,246	96,930,826
	Ram testes epididymis	89,330,764	81,553,121
	Ram thyroid gland	60,638,860	54,187,652
	Ram tonsil	79,036,824	68,401,334
	Ram ventricle	137,711,684	126,216,587
	Lamb abomasum	76,028,632	65,256,622
	Lamb adrenal gland	69,584,336	64,641,763
	Lamb caecum	64,006,906	49,664,011
	Lamb cerebellum	73,606,310	68,824,706
	Lamb cerebrum	98,220,768	92,250,075

	Lamb cervix	82,137,082	76,342,022
	Lamb colon	61,811,046	53,785,166
	Lamb hypothalamus	81,760,226	74,592,276
	Lamb kidney cortex	78,258,174	72,295,830
	Lamb kidney medulla	147,231,286	128,998,323
	Lamb lung	77,684,220	69,247,420
	Lamb lymph node mesenteric	78,045,452	71,919,016
	Lamb lymph node prescapular	111,730,248	103,417,350
	Lamb mammary gland	169,583,902	156,462,840
	Lamb muscle biceps	141,313,386	128,142,672
	Lamb muscle long dorsal	108,662,348	98,315,370
	Lamb omentum	63,845,402	58,613,597
	Lamb ovarian follicles	83,296,756	68,005,443
	Lamb ovary	70,295,788	64,073,883
	Lamb peyer's patch	151,036,272	136,352,610
	Lamb pituitary gland	79,924,110	74,573,197
	Lamb rectum	84,197,222	77,843,469
	Lamb rumen	139,214,568	124,258,217
	Lamb skin back	80,893,222	74,061,392
	Lamb spleen	89,479,968	82,993,603
	Lamb thyroid gland	68,388,800	63,223,286
	Lamb uterus	100,198,926	92,150,618
	Lamb ventricle	119,525,042	104,215,570
	Ewe abomasum	114,292,718	101,181,513
	Ewe adrenal gland	100,055,390	93,392,107
	Ewe alveolar macrophages	108,464,290	92,949,218
	Ewe cerebellum	67,080,764	58,561,729
	Ewe cervix	44,125,388	35,923,569
	Ewe colon	143,926,686	132,018,588
	Ewe corpus luteum	57,079,798	52,205,428
	Ewe heart vertricle	77,997,224	69,154,769
	Ewe kidney medulla	80,023,904	72,150,907
	Ewe liver	94,692,758	78,646,168
	Ewe lung	103,459,074	93,157,006
	Ewe lymph node mesenteric	61,224,904	55,748,997
	Ewe mammary gland	157,588,662	146,768,983

	Ewe muscle biceps	102,732,442	92,053,325
	Ewe muscle long dorsal	78,979,872	70,723,067
	Ewe omentum	67,757,636	60,687,081
	Ewe ovary	56,319,328	51,867,830
	Ewe peyers patch	93,101,228	83,634,563
	Ewe pituitary	69,662,304	64,683,731
	Ewe placenta membranes	37,031,834	33,484,064
	Ewe rectum	79,945,450	72,970,900
	Ewe rumen	109,945,320	98,182,253
	Ewe skin side	87,361,100	80,134,463
	Ewe thyroid gland	76,015,486	70,362,538
	Ewe uterus	92,753,772	84,946,899
	Whole embryo	380,997,588	351,232,681
BGI	Reference brain	31,846,364	30,156,826
	Reference heart	28,311,302	26,237,966
	Reference kidney	28,659,972	26,521,475
	Reference liver	25,976,374	23,237,649
	Reference lung	27,771,950	25,397,246
	Reference ovarian	30,207,260	27,590,781
	Reference white adipose	32,516,614	29,870,436
	Merino skin	20,778,330	18,920,984
USDA	Polypay individual 1	71,983,228	54,354,891
	Rambouillet individual 1	37,405,600	28,604,394
	Rambouillet individual 2	95,992,076	72,291,050
	Merged	8,189,753,630	7,359,995,908

Table 3: Tissue type used for the RNASeq pipeline

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 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 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. [PMID: 9862982] http://tandem.bu.edu/trf/trf.html

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., 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]

20 Eddy, SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure**. BMC Bioinformatics 2002, 3:18. [PMID:12095421]