Ensembl gene annotation project

Danio rerio (Zebrafish)

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

The annotation process of the high-coverage zebrafish assembly began with the raw compute stage [Figure 1] whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1.] (version 3.2.7 with a custom repeat library), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 54.2% of the zebrafish genome.

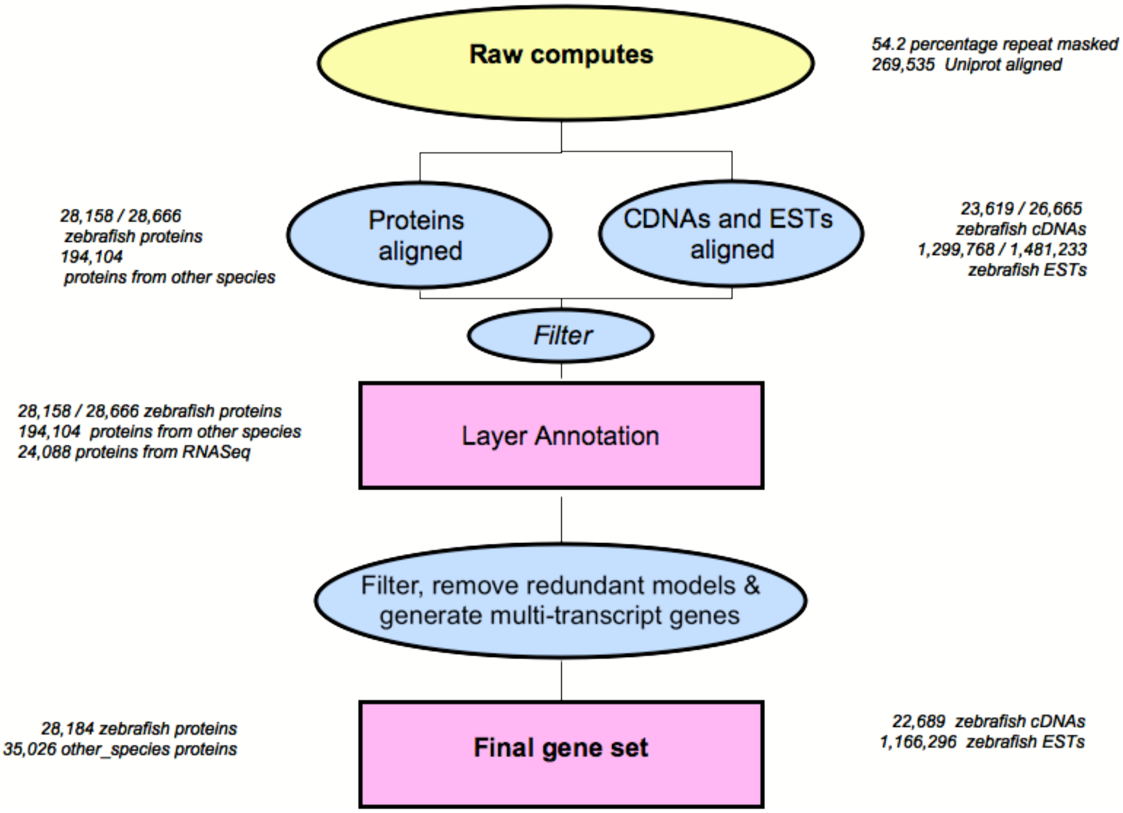
Raw computes
54.2 percentage repeat masked
269,535 Uniprot aligned
Proteins aligned
CDNAs and ESTs aligned
28,158 / 28,666 zebrafish proteins
194,104 proteins from other species
23,619 / 26,665 zebrafish cDNAs
1,299,768 / 1,481,233 zebrafish ESTs
Filter
Layer Annotation
28,158 / 28,666 zebrafish proteins
194,104 proteins from other species
24,088 proteins from RNASeq
Filter, remove redundant models & generate multi-transcript genes
Final gene set
28,184 zebrafish proteins
35,026 other_species proteins
22,689 zebrafish cDNAs
1,166,296 zebrafish ESTs

Figure 1: Summary of zebrafish gene annotation project.

Transcription start sites were predicted using Eponine-scan [4.] and FirstEF [5.]. CpG islands 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 269,535 UniProt, 309,553 UniGene and 284,670 Vertebrate RNA sequences aligning to the genome.

Targeted Stage: Generating coding models from zebrafish evidence

Next, zebrafish protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.]) and filtered to remove sequences based on predictions. The zebrafish 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.] 2 sets of models were produced, one with all consensus splice sites and one where non-consensus splices were allowed, where a single protein sequence had generated two different coding models at the same 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 zebrafish-specific data is referred to as the "Targeted stage". This stage resulted in 28,158 (of 28,666) zebrafish proteins used to build 30,929 coding models.

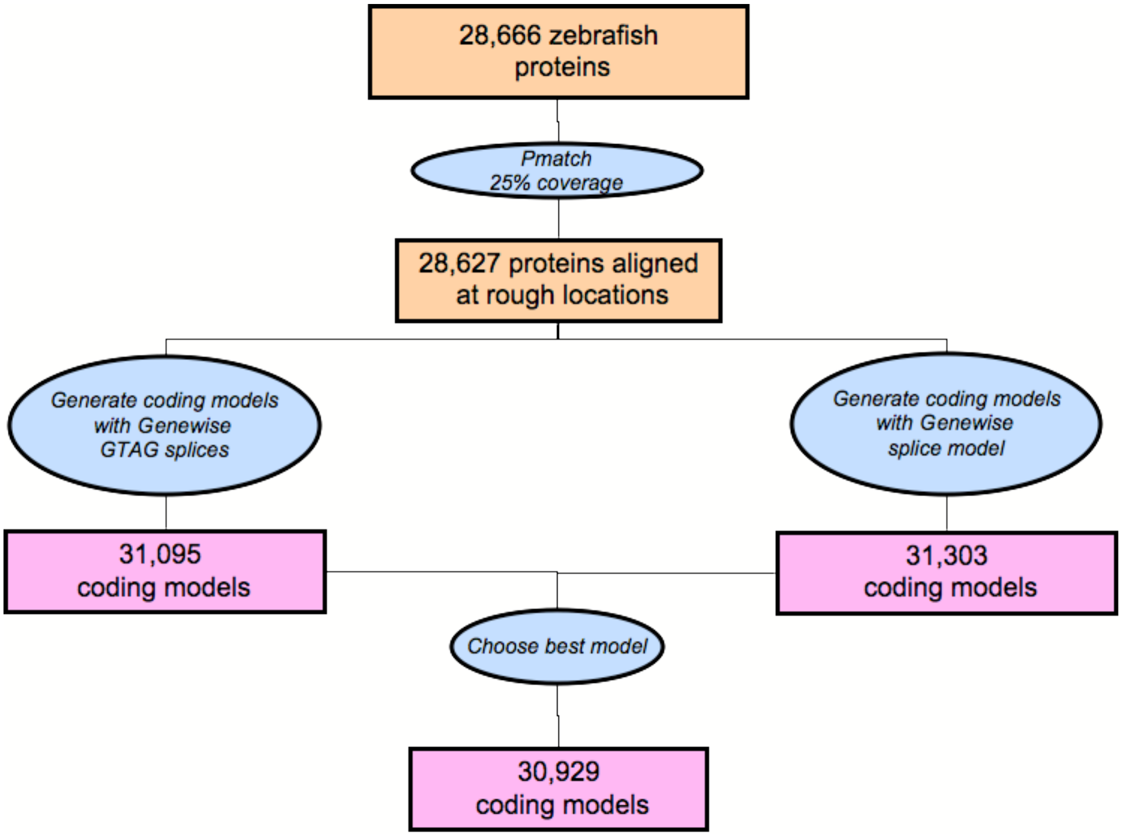
28,666 zebrafish proteins
Pmatch 25% coverage
28,627 proteins aligned at rough locations
Generate coding models with Genewise GTAG splices
Generate coding models with Genewise splice model
31,095 coding models
31,303 coding models
Choose best model
30,929 coding models

Figure 2: Targeted stage using zebrafish specific proteins

cDNA and EST Alignment

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

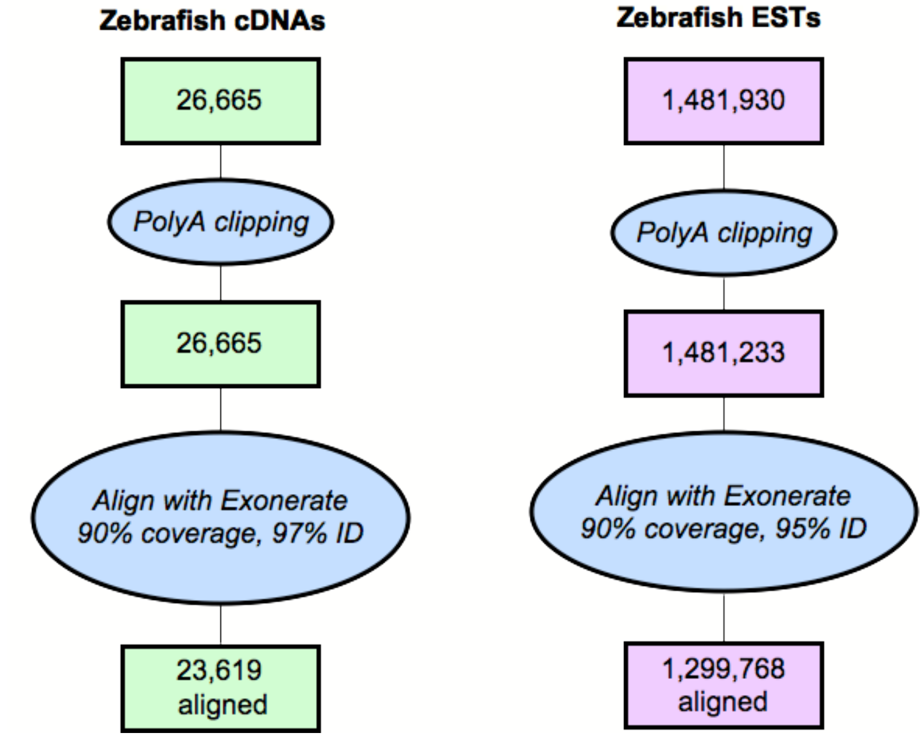

Figure 3: Alignment of zebrafish cDNAs and ESTs, and other species cDNAs to the zebrafish genome.

Of these, 1,299,768 (of 1,481,233) zebrafish ESTs aligned with a cut-off of 90% coverage and 95% identity, and 23,619 (of 26,665) zebrafish cDNAs aligned with a cut-off of 90% coverage and 97% identity.

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

Following the Targeted alignments, additional coding models were generated as follows. The UniProt alignments from the Raw Compute 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 zebrafish is referred to as the "Similarity stage". This stage resulted in 194,104 coding models. [Figure 4]

Filtering Coding Models

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus, RNA-Seq spliced alignments supporting introns were used to help filter the set along with ESTs and cDNAs. The Apollo software [15.] was used to visualise the results of filtering. [Figure 4]

Addition of RNA-Seq models

RNA-Seq models were built from a mixture of 37bp, 54bp and 76bp Illumina paired end reads from 12 zebrafish tissues. 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 discarded. The resulting models were added into the gene set where they produced a novel model or splice variant. [Figure 4]

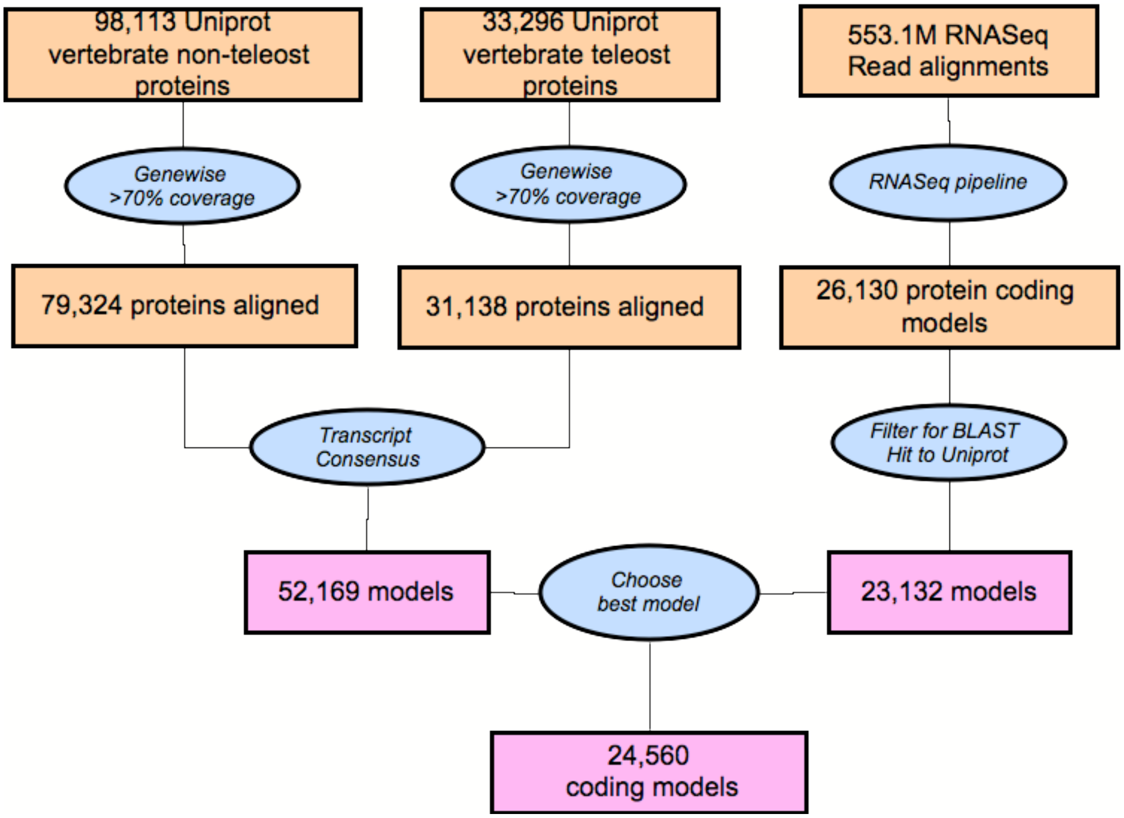

Figure 4: Alignment and filtering of other species proteins and addition of RNASeq models

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 24,818 genes included 10,875 genes built only using zebrafish proteins, a further 4,944 genes built only using proteins from other species, and 7,062 genes built only from RNA-Seq evidence. The remaining 1,937 genes are composed of a mixture of these three evidence sources.

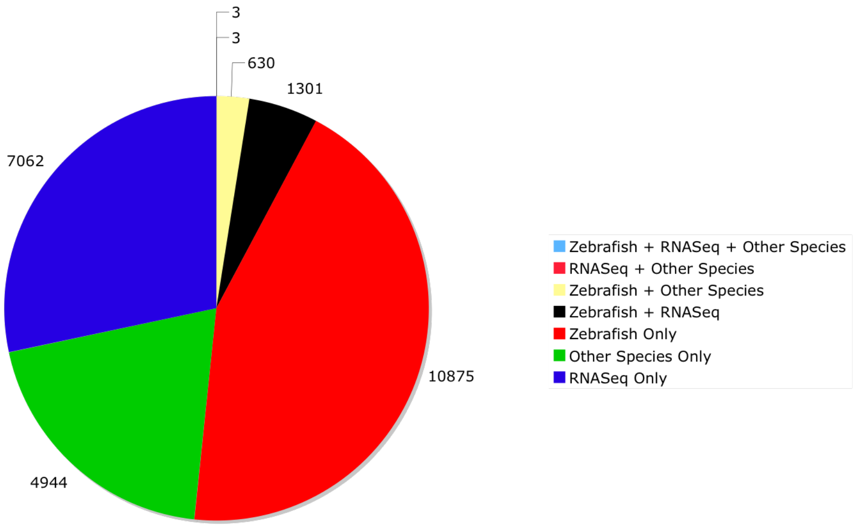

Figure 5: Supporting evidence for zebrafish final gene set.

The final transcript set of 29,324 transcripts included 15,048 transcripts with support from zebrafish proteins, 5,902 transcripts with support from other species proteins and 8,374 transcripts with support from RNA-Seq.

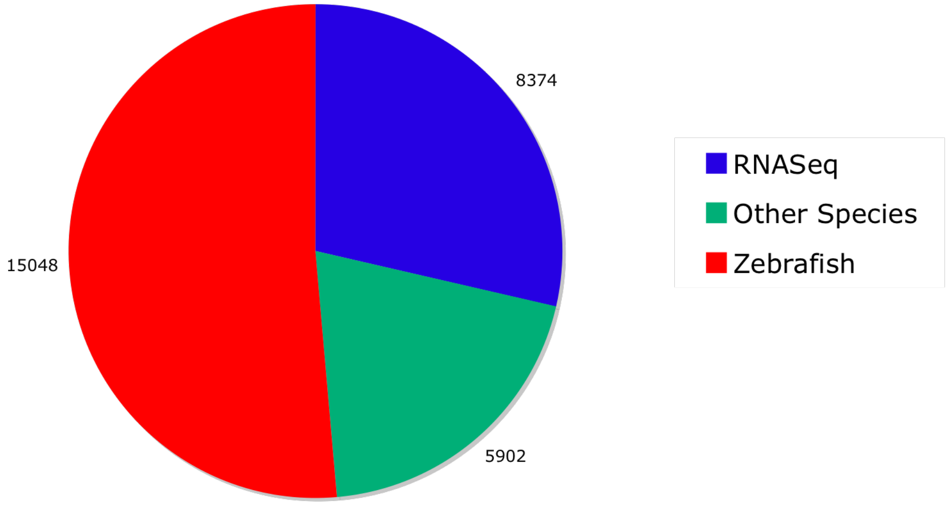

Figure 6: Supporting evidence for zebrafish final transcript set.

Pseudogenes, Protein annotation, non coding genes, Cross-referencing, 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 labelled where appropriate. Stable identifiers were assigned to each gene, transcript, exon and translation. (When annotating a zebrafish for the first time, these identifiers are auto-generated. In all subsequent annotations for zebrafish 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.].

Ensembl Havana Merge Set

Following the completion of the Ensembl gene set Ensembl annotations and Havana annotations from the VEGA database [18.] were merged at the transcript level. Protein-coding transcripts from the two annotation sources were merged if they shared exon-intron structure in the coding and untranslated regions, while non-coding transcripts were merged if they shared identical exon-intron structures.

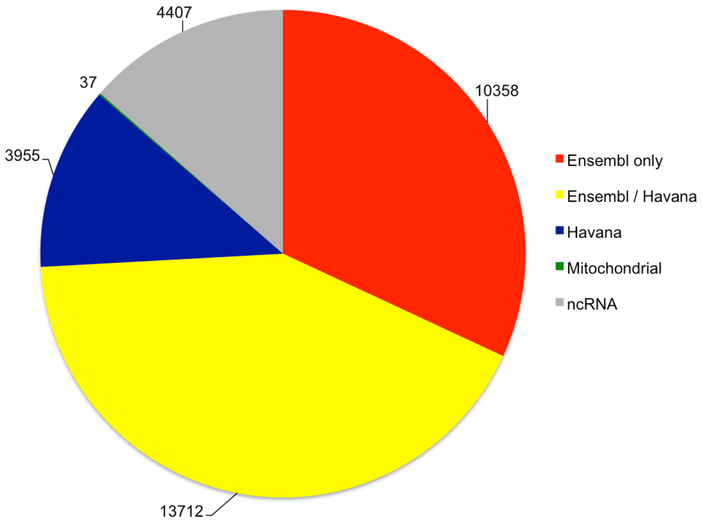

Figure 7: Composition of merged zebrafish gene set.

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 indicate 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 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. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S.R. Eddy: Rfam: an RNA family database. Nucleic Acids Research (2003) 31(1):p439-441.

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