Ensembl gene annotation project

Meleagris_gallopavo (Turkey)

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

Approximate time: 1 week

The annotation process of the high-coverage turkey 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 "Galliformes" -s`'), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 9.48% of the turkey genome.

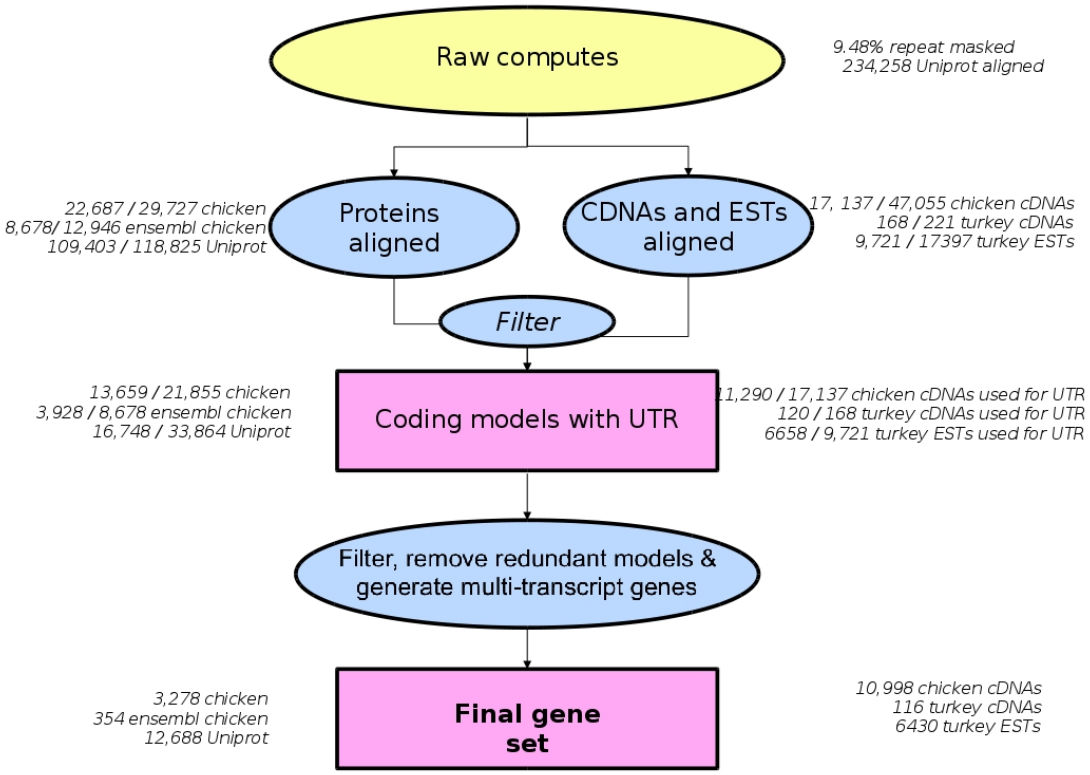

Figure 1: Summary of turkey gene annotation project.

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 283.488 UniProt, 272.363 UniGene and 267.286 Vertebrate RNA sequences aligning to the genome.

Targetted Stage: Generating coding models from chicken evidence

Approximate time: 1 week

Next, chicken protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.]). The chicken 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 [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 species-specific (in this case chicken) data is referred to as the “Targetted stage”. This stage resulted in 21,855 (of 29,727) chicken proteins used to build coding models to be taken through to the UTR addition stage.

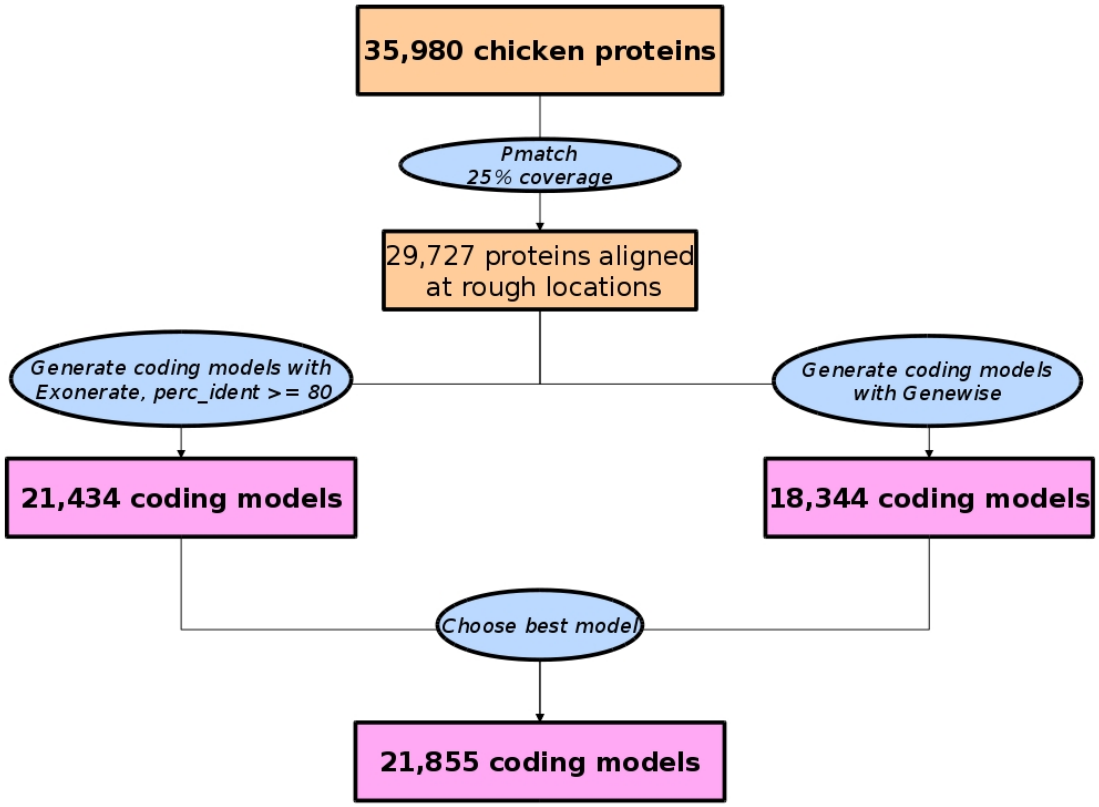

Figure 2: Alignment and filtering of chicken proteins.

Exonerate Stage: Generating coding models from ensembl chicken evidence

Approximate time: 1 week

Also, chicken protein sequences were downloaded from Ensembl (e! 59). The ensembl chicken protein sequences were aligned to the genome using Exonerate as indicated in [Figure 3].

This stage resulted in 8,678 (of 12,946) ensembl chicken proteins used to build coding models to be taken through to the UTR addition stage.

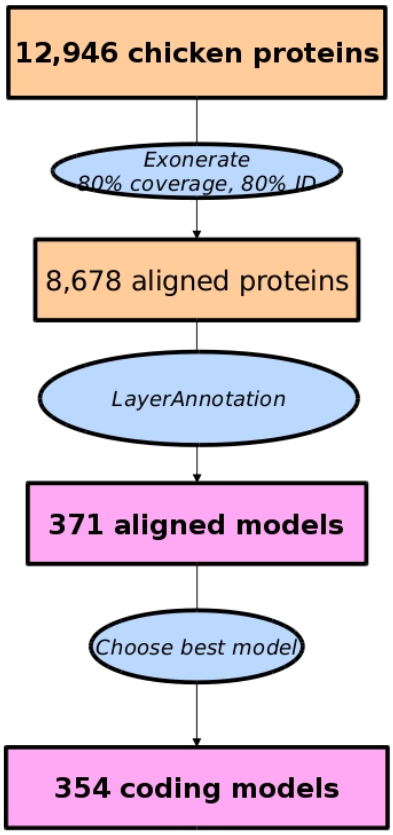

Figure 3: Alignment and filtering of ensembl chicken proteins.

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

Approximate time: 1 week

Following the chicken 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 [11] 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 5,209 galliforme, 1, 177 bird and 112,439 vertebrate coding models.

cDNA and EST Alignment

Approximate time: 1 week

Turkey cDNAs and ESTs and chicken cDNAs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate [Figure 4].

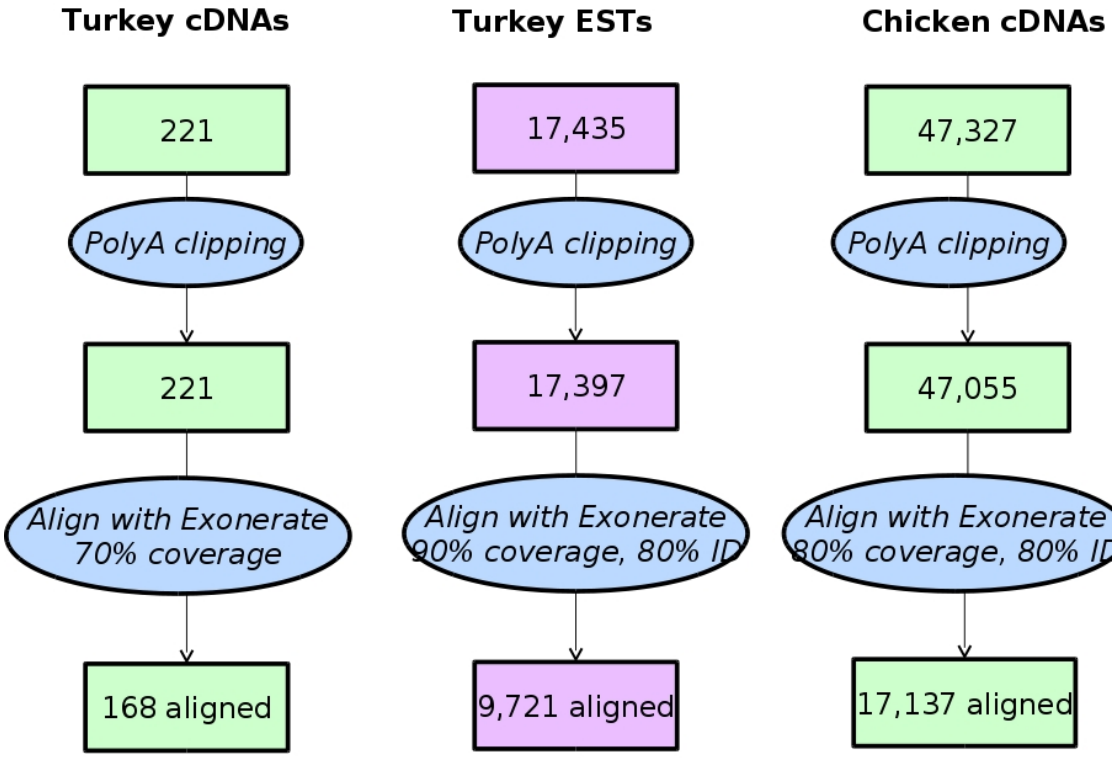

Figure 4: Alignment of turkey cDNAs and ESTs, and chicken cDNAs to the turkey genome.

Of these, 17,137 (of 47,327) chicken cDNAs aligned, 168(of 221) turkey cDNAs aligned, and 9.721 (of 17,435) turkey ESTs aligned. All alignments were at a cut-off of 90% coverage and 80% identity. EST alignments were used to generate EST-based gene models similar to those for human [14.] and these are displayed on the website in a separate track from the Ensembl gene set.

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 [15.] 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 chicken cDNA, turkey cDNA and turkey EST sequences. This resulted in 13,659 (of 21,855) chicken coding models with UTR, 3,928 (of 8,678) ensembl chicken coding models with UTR, and 16,748 (of 33,864) UniProt coding models with UTR.

Generating multi-transcript genes

Approximate time: 5 weeks

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 14,112 genes included 2,822 genes with at least one transcript supported by chicken proteins, a further 349 genes with at least one transcript supported by ensembl chicken evidence. The remaining 10,941 genes had transcripts supported by proteins from other sources [Figure 5].

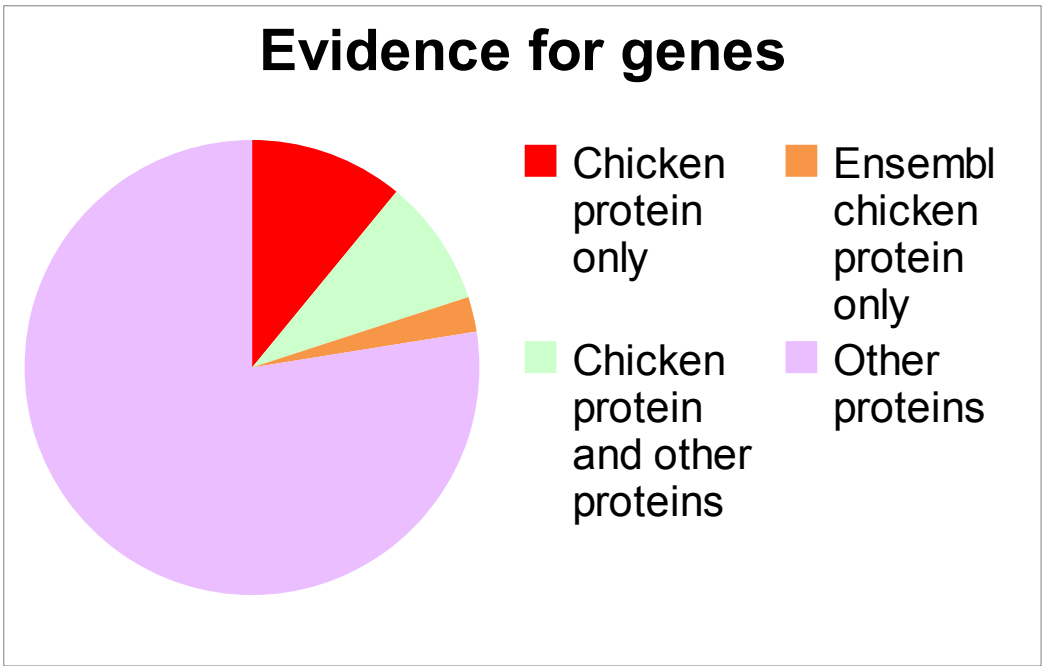

Figure 5: Supporting evidence for turkey final gene set.

The final transcript set of 16483 transcripts included 3,200 transcripts with support from chicken proteins, 349 transcripts with support from esembl chicken proteins and 12,934 transcripts with support from UniProt SwissProt [Figure 6].

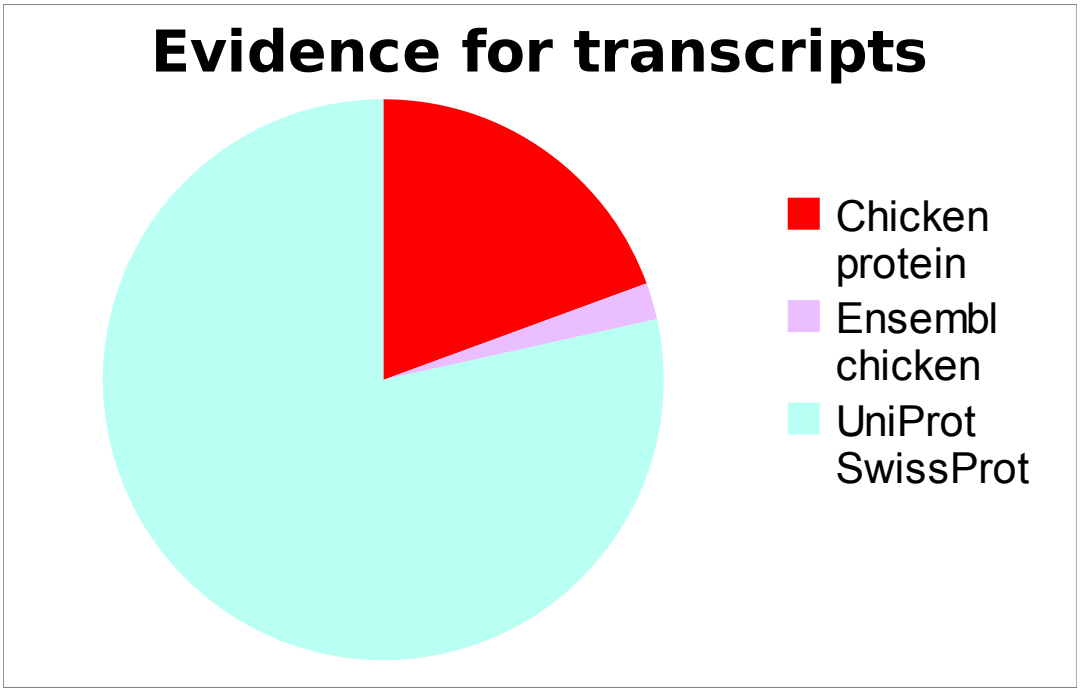

Figure 6: Supporting evidence for turkey final transcript set.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 1 week

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.)

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=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 Bioinformatic*s 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]