Ensembl gene annotation project

Pan troglodytes (Common chimpanzee)

Daniel Barrell

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

Approximate time: two weeks

The annotation process of the high-coverage chimpanzee 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, run twice, with parameters '`-nolow -species "pan troglodytes" -s`' and '`-nolow -mammal -s`'), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 48% of the chimpanzee genome.

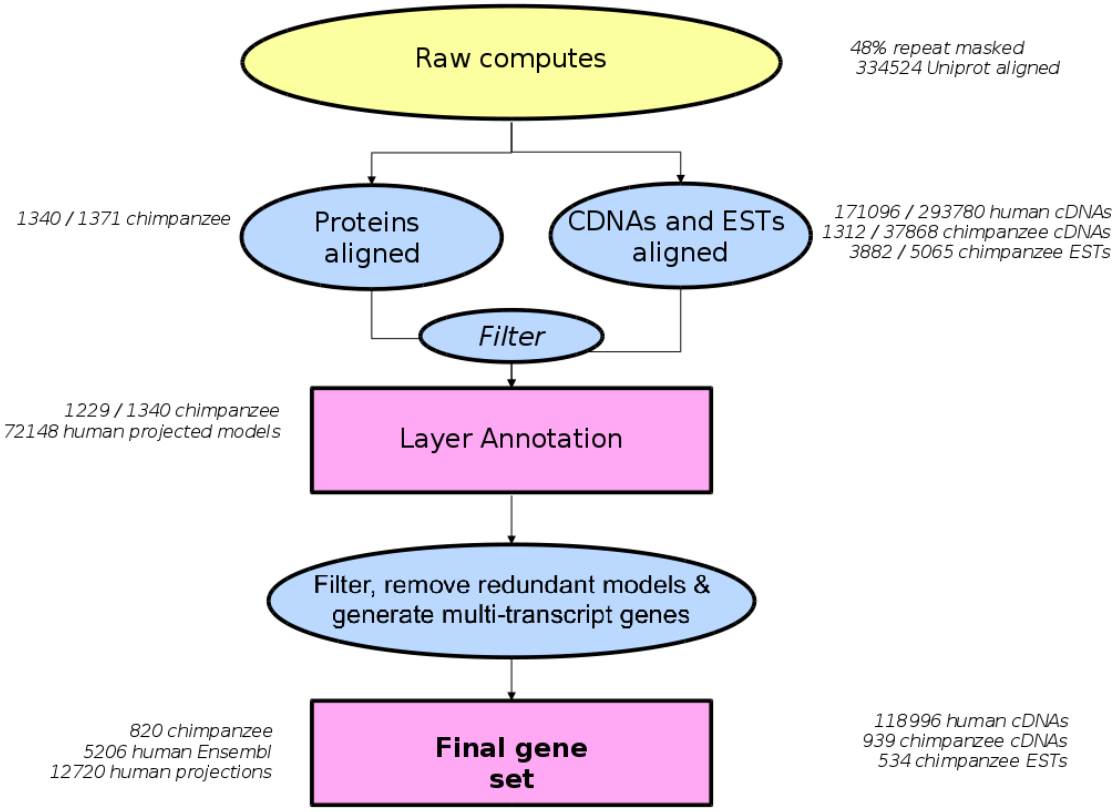

Figure 1: Summary of chimpanzee gene annotation project.

Transcription start sites were predicted using Eponine–scan [4.] and FirstEF [5.]. CpG islands [Micklem, G.] longer than 400 bases 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 334524 UniProt, 351330 UniGene and 344978 Vertebrate RNA sequences aligning to the genome.

Exonerate Stage: Generating coding models from chimpanzee evidence

Approximate time: Two weeks

Next, chimpanzee sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.]). The Uniprot proteins were filtered so that they do not contain fragments and so they only have PE levels of 1 (Evidence at protein level) or 2 (Evidence at transcript level) (See http://www.uniprot.org/docs/pe_criteria for more information). The chimpanzee 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 [15.] 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 chimpanzee) data is referred to as the “Targetted stage”. This stage resulted in 1340 (of 1371) chimpanzee proteins used to build coding models to be taken through to the UTR addition stage.

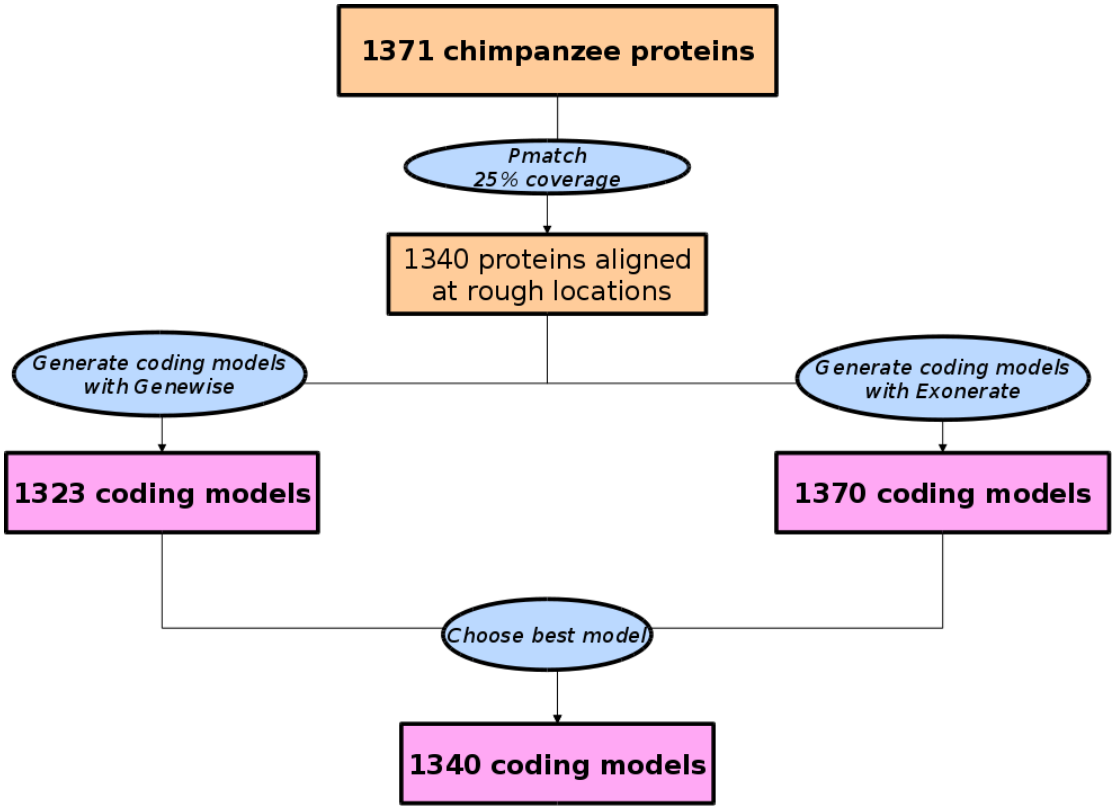

Figure 2: Targetted stage using Chimpanzee protein sequences.

Whole genome alignment (WGA) to human and gene Projection Stages

Approximate time: Three months

The chimpanzee genome was then aligned to the human GRCh37 assembly from Ensembl release 63 using blastz. The alignments were then organised into chains (a sequence of gapless aligned blocks) and nets (a hierarchical collection of chains) [13.] in a pipeline run within the eHive [14.]. A total of 6008865 genomic alignments created 43980991 chains and 271762 nets. The chains and nets were then used to project human genes onto the chimpanzee assembly. The chains generated 37987 protein align features and the nets generated 34161.

cDNA and EST Alignment

Approximate time: One week

Chimpanzee cDNAs and ESTs and human cDNAs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate [Figure 3].

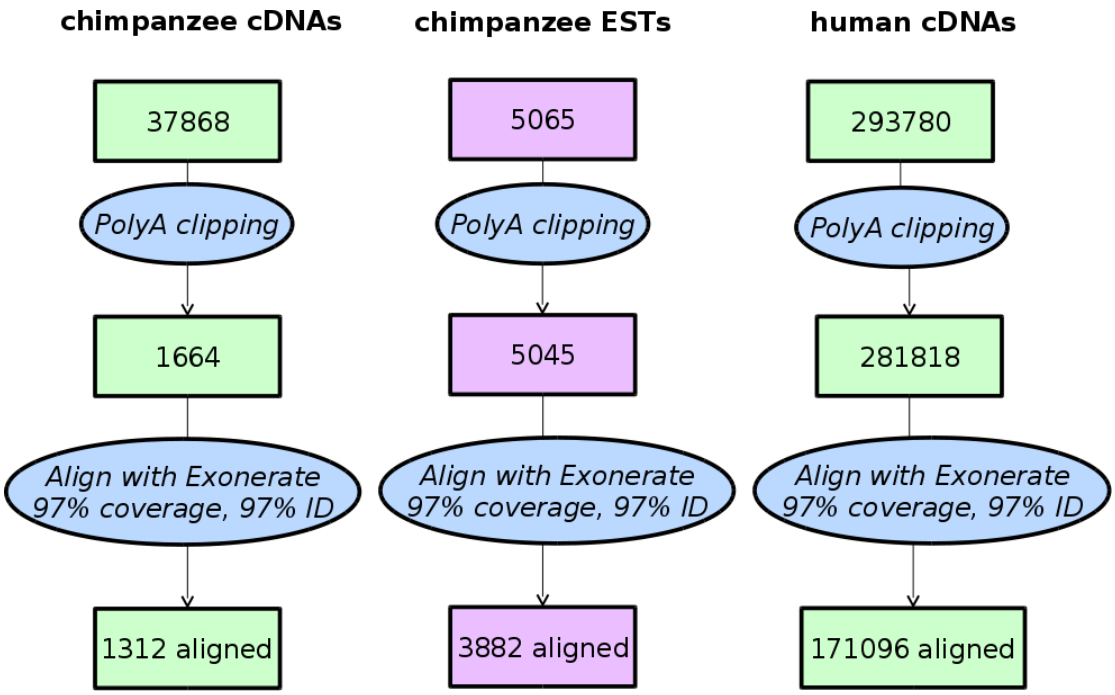

Figure 3: Alignment of chimpanzee cDNAs and ESTs, and human cDNAs to the chimpanzee genome.

Of these, 171096 (of 293780) human cDNAs aligned, 1312 (of 37868) chimpanzee cDNAs aligned, and 3882 (of 5065) chimpanzee ESTs aligned. All alignments were at a cut-off of 90% coverage and 80% identity. The reason for such a low number of chimpanzee cDNAs aligning was because we removed 36202 predicted RefSeq models from the data set (only validated sequences are used as evidence). EST alignments were used to generate EST-based gene models similar to those for [16.] and these are displayed on the website in a separate track from the Ensembl gene set.

Ensembl 63 Human Alignments

Approximate time: One week

Human Ensembl protein alignments with cut off levels of 90% identity and 60% coverage were generated to provide evidence in the filtering stage. 18540 proteins (out of 20282) aligned. 787 of these models had a single internal stop and these were edited into introns so that they would enter the final gene set if used as evidence.

Filtering Coding Models

Approximate time: Three weeks

Coding models from the Targetted and WGA/Projection stages were filtered using modules such as TranscriptConsensus and LayerAnnotation. The Apollo software [17.] was used to visualise the results of filtering. The Ensembl Human alignments were used to provide evidence to projected models. Where the Exonerate model is sufficiently 'better' than the projected model, we use that instead. We deem multi-exon Exonerate models better than single or double exon Exonerate models.

Addition of UTR to coding models

Approximate time: One week

The set of coding models was extended into the untranslated regions (UTRs) using human cDNA, chimpanzee cDNA and chimpanzee EST sequences. This resulted in 1229 chimpanzee coding models with UTR, 4863 human coding models with UTR, and 24850 projected human coding models with UTR (11573 from nets, 13277 from chains).

Generating multi-transcript genes

Approximate time: Three weeks

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. The final gene set of 18746 protein coding genes included 820 genes with at least one transcript supported by chimpanzee proteins, a further 5206 genes without species evidence but with at least one transcript supported by human evidence. The remaining 12720 genes had transcripts supported by proteins from human projections [Figure 4].

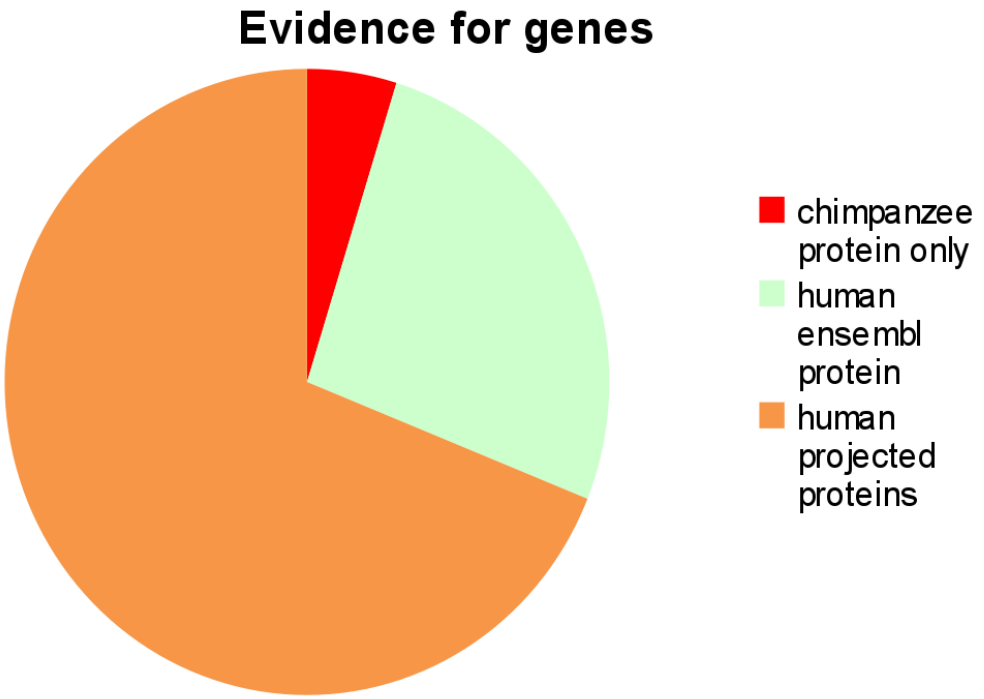

Figure 4: Supporting evidence for chimpanzee final gene set.

The final transcript set of 19894 transcripts included 952 transcripts with support from chimpanzee proteins, 13709 transcripts with support from human projected proteins and 5233 transcripts with support from human Ensembl proteins [Figure 5].

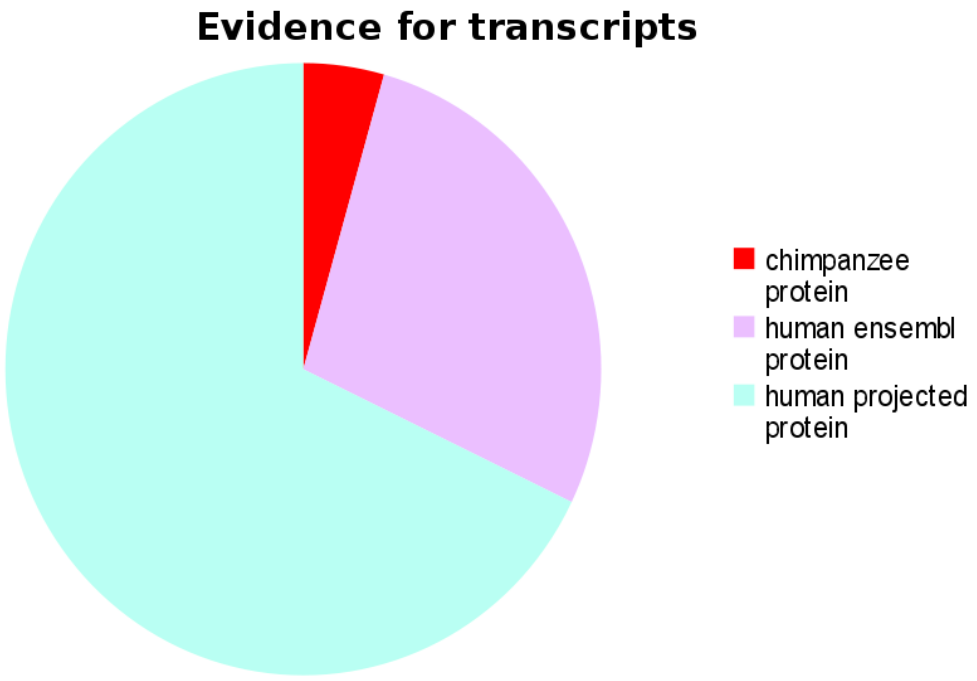

Figure 5: Supporting evidence for chimpanzee final transcript set.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: One 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 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: 1187 E 5034]
5. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the 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 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. W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler: **Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes.** PNAS September 30, 2003 vol. 100 no. 20 11484-11489 [PMID: 14500911]
14. Jessica Severin, Kathryn Beal, Albert J Vilella, Stephen Fitzgerald, Michael Schuster, Leo Gordon, Abel Ureta-Vidal, Paul Flicek and Javier Herrero: **eHive: An Artificial Intelligence workflow system for genomic analysis.** BMC Bioinformatics 2010, 11:240 [PMID: 20459813]
15. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res.* 2004, **14(5):**988-995. [PMID: 15123596]
16. Eyras E, Caccamo M, Curwen V, Clamp M. **ESTGenes: alternative splicing from ESTs in Ensembl.** *Genome Res.* 2004 **14(5):**976-987. [PMID: 15123595]
17. 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]