Genome annotation report

Authors: Jèssica Gómez-Garrido, Tyler Alioto

Methods

Repeats present in the fSolSen1 genome assembly were annotated with RepeatMasker v4-0-7 (http://www.repeatmasker.org) using the custom repeat library available for *Danio rerio*. Moreover, a new repeat library specific for our assembly was made with RepeatModeler v1.0.11. After excluding those repeats that were part of repetitive protein families (performing a blast against uniprot) from the resulting library, Repeat Masker was run again with this new library in order to annotate the specific repeats.

The gene annotation of the assembly was obtained by combining transcript alignments, protein alignments and *ab initio* gene predictions.

Firstly, RNAseq reads were obtained from several tissues and developmental stages and aligned to the genome with STAR [1](v-2.7.2a). Transcript models were subsequently generated using Stringtie [2] (v2.0.1) on each BAM file and then all the models produced were combined using TACO v0.6.2. Finally, PASA assemblies were produced with PASA [3] (v2.4.1). The *TransDecoder* program, which is part of the PASA package, was run on the PASA assemblies to detect coding regions in the transcripts. Secondly, the complete *Danio rerio*, *Scopthalmus maximus* and *Cynoglossus semilaevis* proteomes were downloaded from Uniprot in April 2020 and aligned to the genome using spaln [4] (v2.4.03). *Ab initio* gene predictions were performed on the repeat masked SolSen1 assembly with three different programs: GeneID [5] v1.4, Augustus [6] v3.3.4 and Genemark-ES [7] v2.3e with and without incorporating evidence from the RNAseq data. The gene predictors were run with trained parameters for human, except Genemark that runs on a self-trained manner. Finally, all the data was combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM [3]). Additionally, UTRs and alternative splicing forms were annotated through two rounds of PASA annotation updates. Functional annotation was performed on the annotated proteins with Blast2go [8]. First, a Diamond blastp [9] search was made against the nr database (last accessed May 2021). Furthermore, Interproscan [10] was run to detect protein domains on the annotated proteins. All these data were combined by Blast2go which produced the final functional annotation results.

The non-coding RNA annotation required several steps. First, we annotated as long-non-coding RNAs (lncRNAs) those Pasa-assemblies that had not been included into the protein-coding annotation, that did not match any protein-coding gene and that were longer than 200bp.

We also sequenced small RNAs (sRNAs) from several tissues and developmental stages. The corresponding reads were aligned with STAR [1] (v-2.7.2a) with parameters (--outFilterMultimapNmax 25 --alignIntronMax 1 --alignMatesGapMax 1000000 --outFilterMismatchNoverLmax 0.05 --outFilterMatchNmin 16 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0). The resulting mappings were processed to produce the annotation of small non-coding RNAs. First, TACO was run to assemble the reads into transcripts. Transcripts overlapping exons from the protein-coding or lncRNA annotations were removed from the set of small non-codig RNAs.
Finally, the program cmsearch [11] (v1.1.4) that comes with Infernal [12] was run on the sncRNAs against the RFAM [13] database of RNA families (v14.6) in order to annotate products of those genes.

The final non-coding annotation contains the lncRNAs and the sncRNAs. The resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene.

Results

In total, we have annotated 24,264 protein-coding genes, that produce 40,511 transcripts (1,67 transcripts per gene) and encode for 37,259 unique protein products. We have been able to assign functional labels to 70,39% of the annotated proteins. The annotated transcripts contain 12.79 exons on average, with 95% of them being multi-exonic (Table 1).

In addition, 52,888 non-coding RNAs have been annotated, of which 6,871 and 46,017 are long and short non-coding RNA genes, respectively.

Table 1: Genome annotation statistics

	SolSen1A annotation
Number of protein-coding genes	24,264
Median gene length (bp)	7,566
Number of transcripts	40,511
Number of exons	277,235
Number of coding exons	263,350
Median UTR length (bp)	950
Median intron length (bp)	389
Exons/transcript	12.79
Transcripts/gene	1.67
Multi-exonic transcripts	95.47%
Gene density	39.53

References

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
2. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol* 2015, **33:**290-295.
3. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol* 2008, **9:**R7.
4. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.
5. Parra G, Blanco E, Guigo R: **GeneID in Drosophila.** *Genome Res* 2000, **10:**511-515.
6. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.
7. Lomsadze A, Burns PD, Borodovsky M: **Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.** *Nucleic Acids Res* 2014, **42:**e119.
8. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
9. Benjamin Buchfink 1 KRaH-GD: **Sensitive protein alignments at tree-of-life scale**
using DIAMOND.
10. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
11. Cui X, Lu Z, Wang S, Jing-Yan Wang J, Gao X: **CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction.** *Bioinformatics* 2016, **32:**i332-i340.
12. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics* 2013, **29:**2933-2935.
13. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD: **Rfam 12.0: updates to the RNA families database.** *Nucleic Acids Res* 2015, **43:**D130-137.