Xyricthis novacula genome annotation report

Jèssica Gómez-Garrido[1], Tyler S. Alioto[1,2]
[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

Methods

Genome annotation

Repeats present in the genome assembly were annotated with RepeatMasker v4-1-2 (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[1] search against Uniprot) from the resulting library, RepeatMasker was run again with this new library in order to annotate the specific repeats.

The gene annotation of the mackerel genome assembly was obtained by combining transcript alignments, protein alignments and *ab initio* gene predictions. A flowchart of the annotation process is shown in **Figure ANN1**.

Firstly, RNA from four different tissues was obtained and sequenced with both Illumina RNAseq and ONT direct cDNAseq. After sequencing, the long and short reads were aligned to the genome using, respectively, STAR[2] v-2.7.10a and MINIMAP2[3] v2.24 with the splice option. Transcript models were subsequently generated using Stringtie[4] v2.2.1 on each BAM file and then all the models produced were combined using TACO[5] v0.7.3. High-quality junctions to be used during the annotation process were obtained by running Portcullis[6] v1.2.4 after mapping with STAR and MINIMAP2. Finally, PASA assemblies were produced with PASA[7] v2.5.2. 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 proteomes of *Carassius auratus, Cynoglossus semilaevis, Danio rerio, Oryzias latipes, Parambassis ranga, Sparus aurata* and *Scopthalmus maximus* were downloaded from Uniprot in March 2022 and aligned to the genome using Miniprot[8] 0.6. *Ab initio* gene predictions were performed on the repeat-masked assembly with three different programs: GeneID[9] v1.4, Augustus[10] v3.5.0 and Genemark-ET[11] v4.71 with and without incorporating evidence from the RNAseq data. The gene predictors were run with trained parameters for human, except Genemark, which runs in a self-trained mode. Finally, all the data were combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM)[7]. Additionally, UTRs and alternative splicing forms were annotated via two rounds of PASA annotation updates. Functional annotation was performed on the annotated proteins with Blast2go[12]. First, a Blastp[1] search was made against the nr database (last accessed March 2023). Furthermore, Interproscan[13] v5.55_88.0 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 annotation of ncRNAs was obtained by running the following steps. First, the program cmsearch[14] v1.1.4 that is part of the Infernal[15] package was run against the RFAM database of RNA families[15] v12.0. Additionally, tRNAscan-SE[16] v2.0.11 was run in order to detect the transfer RNA genes present in the masked genome assembly. Identification of lncRNAs was done by first filtering the set of PASA-assemblies that had not been included in the annotation of protein-coding genes to retain those longer than 200bp and not covered more than 80% by a small

ncRNA. The resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene.

Results

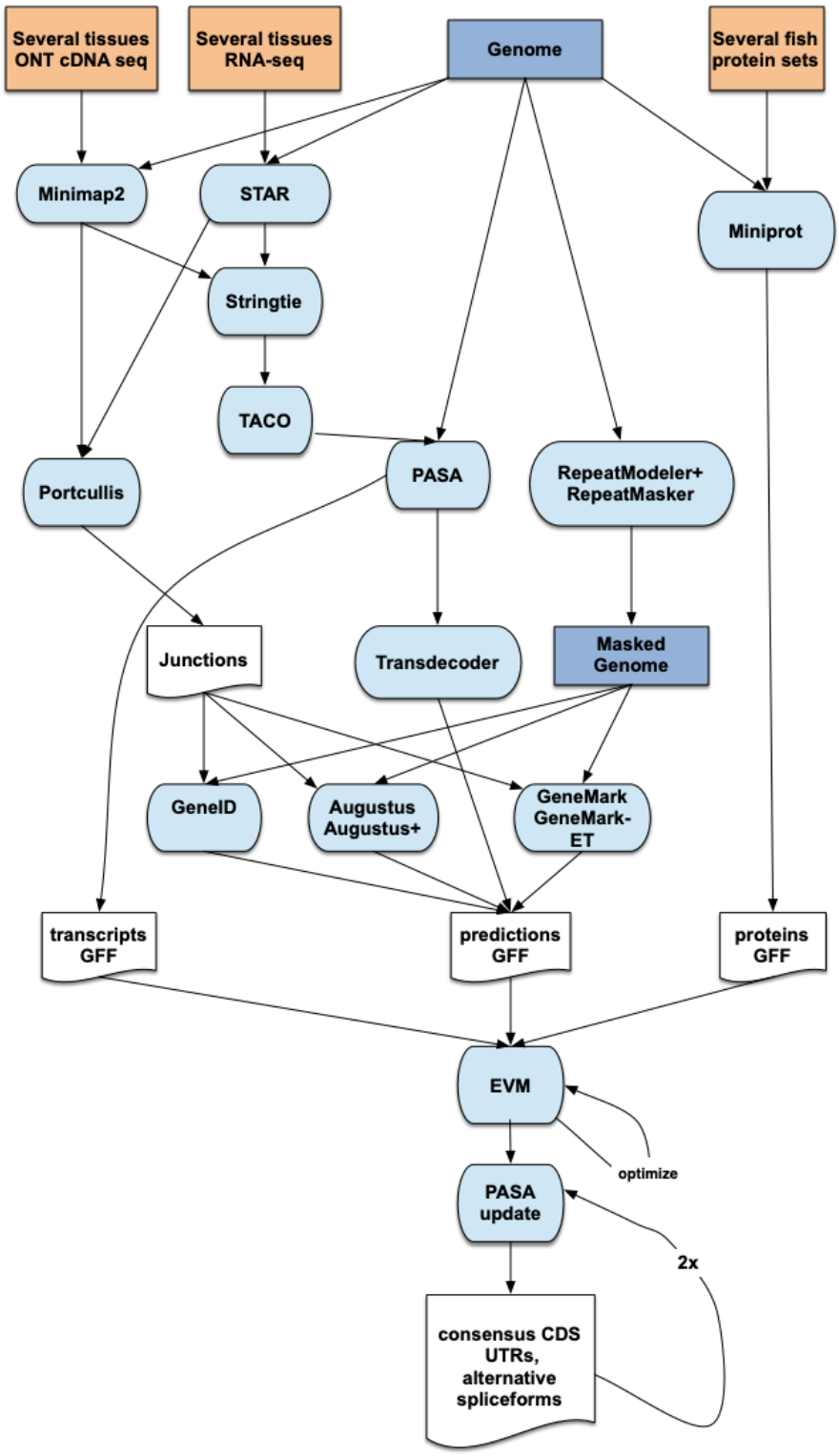
Genome annotation

In total, we annotated 26,428 protein-coding genes that produce 38,000 transcripts (1.44 transcripts per gene) and encode for 35,860 unique protein products. We were able to assign functional labels to 94.2% of the annotated proteins. The annotated transcripts contain 10.88 exons on average, with 93% of them being multi-exonic (**Table ANN1**). In addition, 12,737 non-coding transcripts were annotated, of which 10,450 and 2,287 are long and short non-coding RNA genes, respectively.

Table ANN1: Genome annotation statistics

	XNOV1A annotation
Number of protein-coding genes	26,690
Median gene length (bp)	7,733
Number of transcripts	43,457
Number of exons	281,303
Number of coding exons	263,943
Median UTR length (bp)	1,895
Median intron length (bp)	345
Exons/transcript	10.82
Transcripts/gene	1.63
Multi-exonic transcripts	93%
Gene density (gene/Mb)	34.41

Figure ANN1: workflow of the genome annotation process



Bibliography

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool. *J Mol Biol*, **215**, 403–10.
2. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
3. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–100.
4. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, **33**, 290–5.
5. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. 2017, TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods*, **14**, 68–70.
6. Mapleson, D., Venturini, L., Kaithakottil, G., and Swarbreck, D. 2018, Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience*, **7**.
7. Haas, B. J., Salzberg, S. L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.
8. Li, H. 2023, Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
9. Alioto, T., Blanco, E., Parra, G., and Guigó, R. 2018, Using geneid to Identify Genes. *Curr Protoc Bioinformatics*, **64**, e56.
10. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. 2006, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
11. Lomsadze, A., Burns, P. D., and Borodovsky, M. 2014, Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*, **42**, e119.
12. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–6.
13. Jones, P., Binns, D., Chang, H.-Y., et al. 2014, InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–40.
14. Cui, X., Lu, Z., Wang, S., Jing-Yan Wang, J., and Gao, X. 2016, CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics*, **32**, i332–40.
15. Nawrocki, E. P., and Eddy, S. R. 2013, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–5.
16. Chan, P. P., and Lowe, T. M. 2019, tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol*, **1962**, 1–14.