Merluccius merluccius genome annotation report

Jèssica Gómez-Garrido[1], Tyler S. Alioto[1]
[1]CNAG, Centro Nacional de Análisis Genómico, C/Baldiri Reixac 4, 08028 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 European hake 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**.

RNAseq from five different organs (brain, liver, muscle, spleen and gonad) was used for annotation. The reads were aligned to the genome using STAR[2] v-2.7.2a and transcript models were subsequently generated using Stringtie[3] v2.2.1 and merged using TACO[4] v0.7.3. High-quality junctions to be used during the annotation process were obtained by running ESPRESSO[5] v1.3.0 after mapping with STAR. Finally, PASA assemblies were produced with PASA[6] 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 *Danio rerio, Chanos chanos* and *Carassius auratus* were downloaded from Uniprot in May 2023 and aligned to the genome using Miniprot[7] v0.6. *Ab initio* gene predictions were performed on the repeat-masked assembly with three different programs: GeneID[8] v1.4, Augustus[9] v3.5.0 and Genemark-ET[10] 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)[6]. Additionally, UTRs and alternative splicing forms were annotated via two rounds of PASA annotation updates. Functional annotation was performed on the annotated proteins with Pannzer's[11] online server.

The annotation of ncRNAs was obtained by running the following steps on the repeat masked version of the genome assembly. First, the program cmsearch[12] v1.1 that is part of the Infernal[13] package was run against the RFAM database of RNA families v12.0. Additionally, tRNAscan-SE[14] v2.08 was run in order to detect the tranfer RNA genes present in the 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

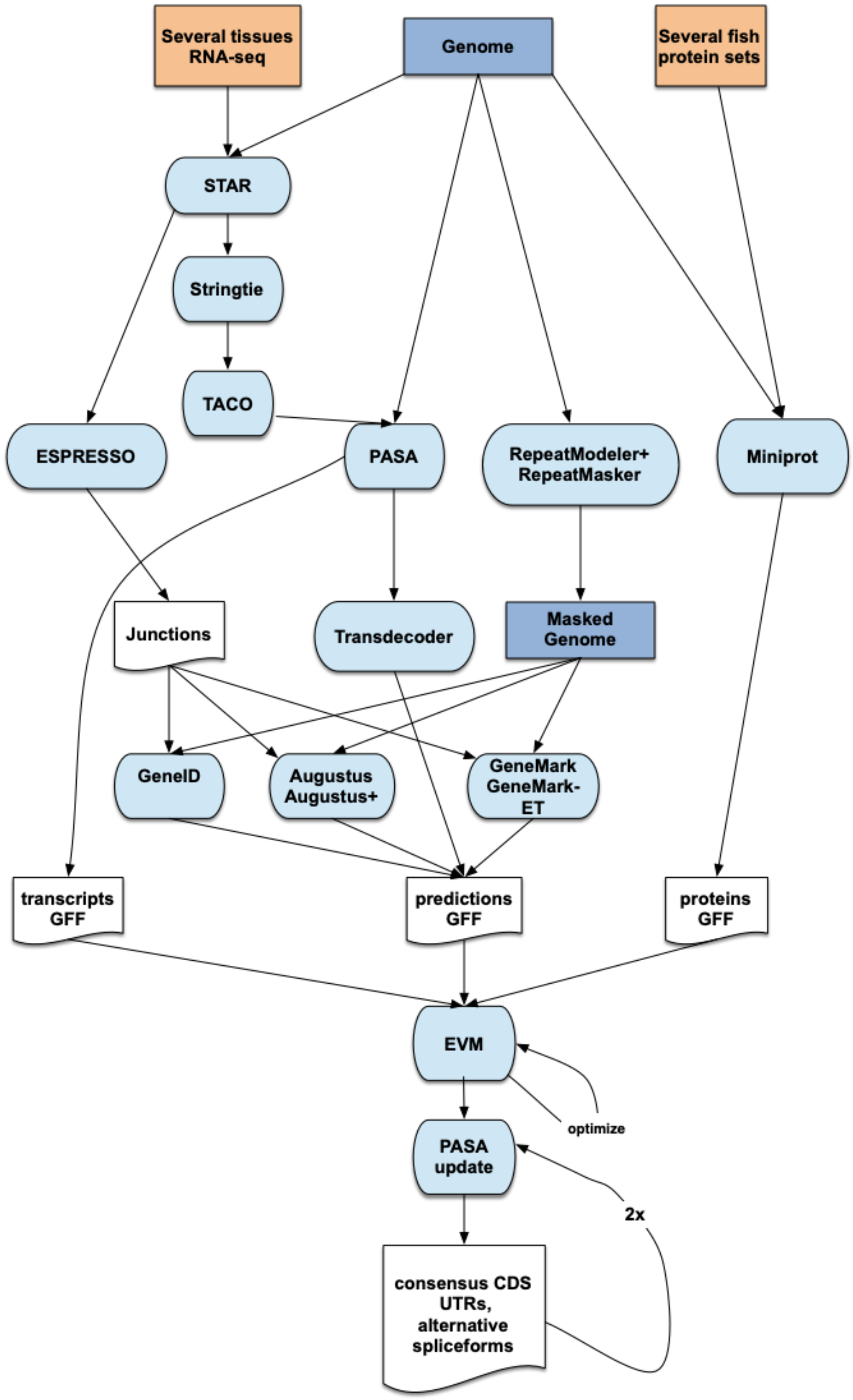
Genome annotation

In total, we annotated 26,625 protein-coding genes that produce 41,543 transcripts (1.56 transcripts per gene) and encode for 37,855 unique protein products. We were able to assign functional labels to 78% of the annotated proteins. The annotated transcripts contain 11.84 exons on average, with 97% of them being multi-exonic (**Table ANN1**). Additionally, 11,083 non-coding transcripts were annotated, of which 5,683 as lncRNA.

Table ANN1: Genome annotation statistics

	MERME1A annotation
Number of protein-coding genes	26,625
Median gene length (bp)	8,357
Number of transcripts	41,543
Number of exons	285,286
Number of coding exons	266,171
Median UTR length (bp)	1,254
Median intron length (bp)	501
Exons/transcript	11.84
Transcripts/gene	1.56
Multi-exonic transcripts	0.97
Gene density (gene/Mb)	37.21

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. 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.
4. 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.
5. Gao, Y., Wang, F., Wang, R., et al. 2023, ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci. Adv.*, **9**, eabq5072.
6. 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.
7. Li, H. 2023, Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
8. Alioto, T., Blanco, E., Parra, G., and Guigó, R. 2018, Using geneid to Identify Genes. *Curr Protoc Bioinformatics*, **64**, e56.
9. 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.
10. 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.
11. Törönen, P., and Holm, L. 2022, PANNZER —A practical tool for protein function prediction. *Protein Science*, **31**, 118–28.
12. 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.
13. Nawrocki, E. P., and Eddy, S. R. 2013, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–5.
14. Chan, P. P., and Lowe, T. M. 2019, tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol*, **1962**, 1–14.