Helleia helle genome annotation report

Francisco Câmara[1], 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 *insecta*. 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 *species* 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**.

RNA from two different tissues (20220704_Lycaena_Wing_CBP18 and 20220704_Lycaena_Leg_CBP19) was obtained and sequenced with ONT direct cDNAseq. the long reads were aligned to the genome using MINIMAP2[2] v2.14 with the splice option. Transcript models were subsequently generated using Stringtie. 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 MINIMAP2. 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 proteomes of the Kamehameha, Danaus and "Squinting bush" butterflies and the "swissprot invertebrates" proteins were downloaded from Uniprot in February 2023. Furthermore, we also downloaded the RefSeq proteomes of all butterflies of the superfamily Papilionoidea from the NCBI available in February 2023. These five protein data sets were combined and aligned to the genome using Miniprot[7] 0.6. *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 the honey bee, 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.

Results

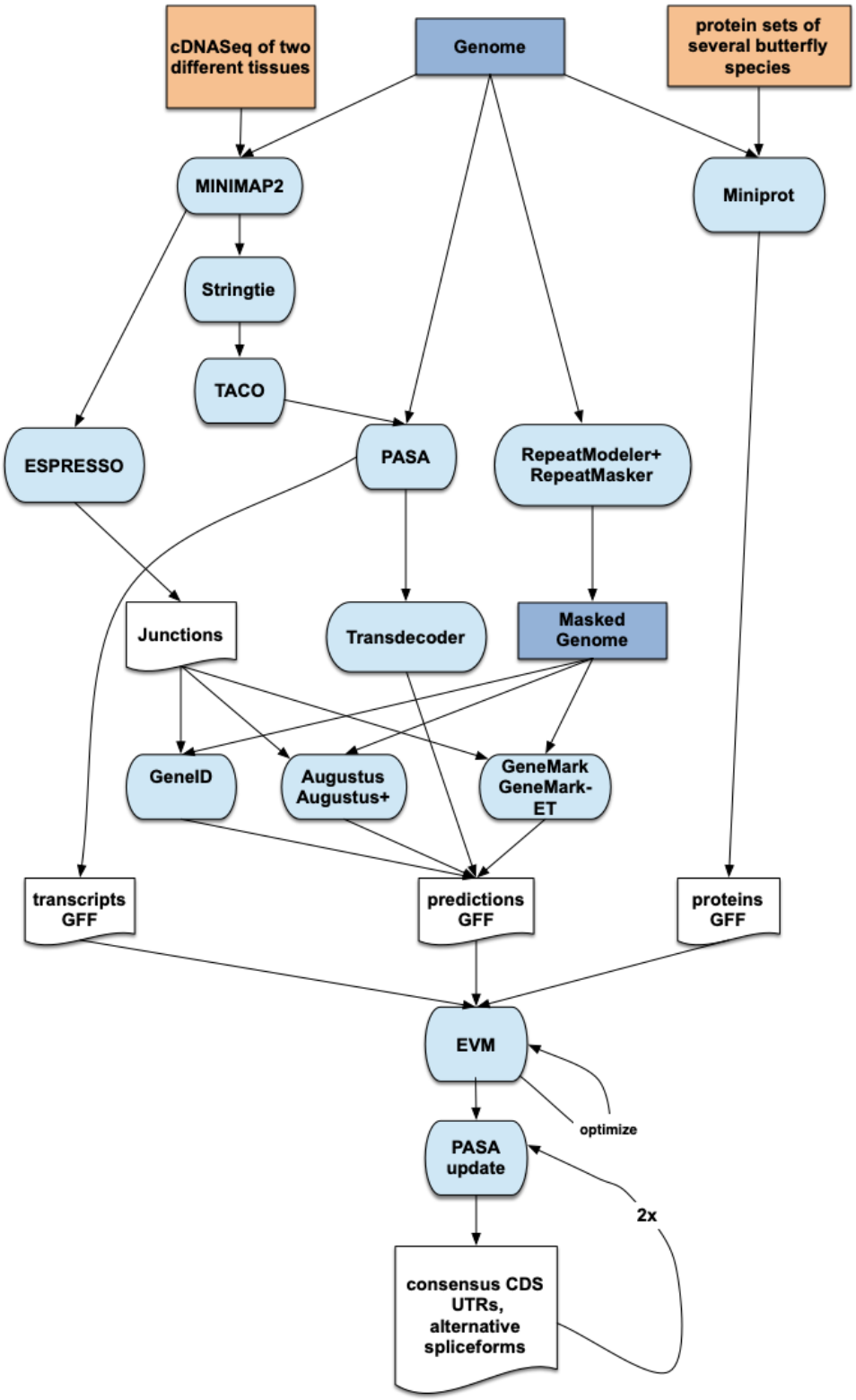
Genome annotation

In total, we annotated 20,122 protein-coding genes that produce 22,758 transcripts (1.13 transcripts per gene) and encode for 22,349 unique protein products. We were able to assign functional labels to 60.5% of the annotated proteins. The annotated transcripts contain 6.1 exons on average, with 81% of them being multi-exonic (**Table ANN1**).

Table ANN1: Genome annotation statistics

	HEHEL1A annotation
Number of protein-coding genes	20,112
Median gene length (bp)	4,624
Number of transcripts	22,758
Number of exons	118,891
Number of coding exons	115,084
Median UTR length (bp)	863
Median intron length (bp)	711
Exons/transcript	6.09640565954829
Transcripts/gene	1.13100089454329
Multi-exonic transcripts	0.809209948150101
Gene density (gene/Mb)	36.7655208363154

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. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–100.
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.