Panthera onca genome annotation report

Jèssica Gómez-Garrido[1,2], Tyler S. Alioto[1,2]
[1]Centro Nacional de Análisis Genómico, C/ Baldiri Reixac 4, 08028 Barcelona, Spain
[2]Universitat de Barcelona (UB), Barcelona, Spain

Methods

Genome annotation

Repeats present in the mPanOnc1.3 genome assembly were annotated with RepeatMasker v4-1-2 (http://www.repeatmasker.org) using the custom repeat library available for cat. The gene annotation of the jaguar 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**.

First, due to the absence of transcriptomic datasets for the jaguar, Illumina RNA-seq reads from several tissues of Panthera leo were downloaded from the NCBI SRA archive (SRR13380817, SRR13380818, SRR5485090, SRR5485091, SRR5485092, SRR5485093, SRR5485094). The reads were then aligned to the genome using STAR[1] v-2.7.2a and transcript models were subsequently generated using Stringtie[2] v2.2.1 on each BAM file. All the models produced were combined using TACO[3] v0.7.3. High-quality junctions to be used during the annotation process were obtained by running ESPRESSO[4] v1.3.0 after mapping with STAR. Finally, PASA assemblies were produced with PASA[5] v2.5.2 by adding also the Panthera onca transcripts annotated on a previous genome assembly produced by the DNAZOO. 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 *Panthera leo, P. tigris altaica, Felis catus, Acinox jubatus, Lynx pardinus* and *L. canadiensis* were downloaded from Uniprot in October 2022 and aligned to the genome using Miniprot[6]. *Ab initio* gene predictions were performed on the repeat-masked assembly with three different programs: GeneID[7] v1.4, Augustus[8] v3.5.0 and GeneMark-ET[9] v4.71 with and without incorporating evidence from the RNAseq data. The gene predictors were run with human specific parameters, except Genemark that runs in a self-trained way. Finally, all the data were combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM)[5]. 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[10] online server.

Comentado [JG1]: Change this and add the citation

The annotation of ncRNAs was obtained by running the following steps on the repeat masked version of the genome assembly. First, the program cmsearch[11] v1.1 that is part of the Infernal[12] package was run against the RFAM database of RNA families v12.0. Additionally, tRNAscan-SE[13] 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

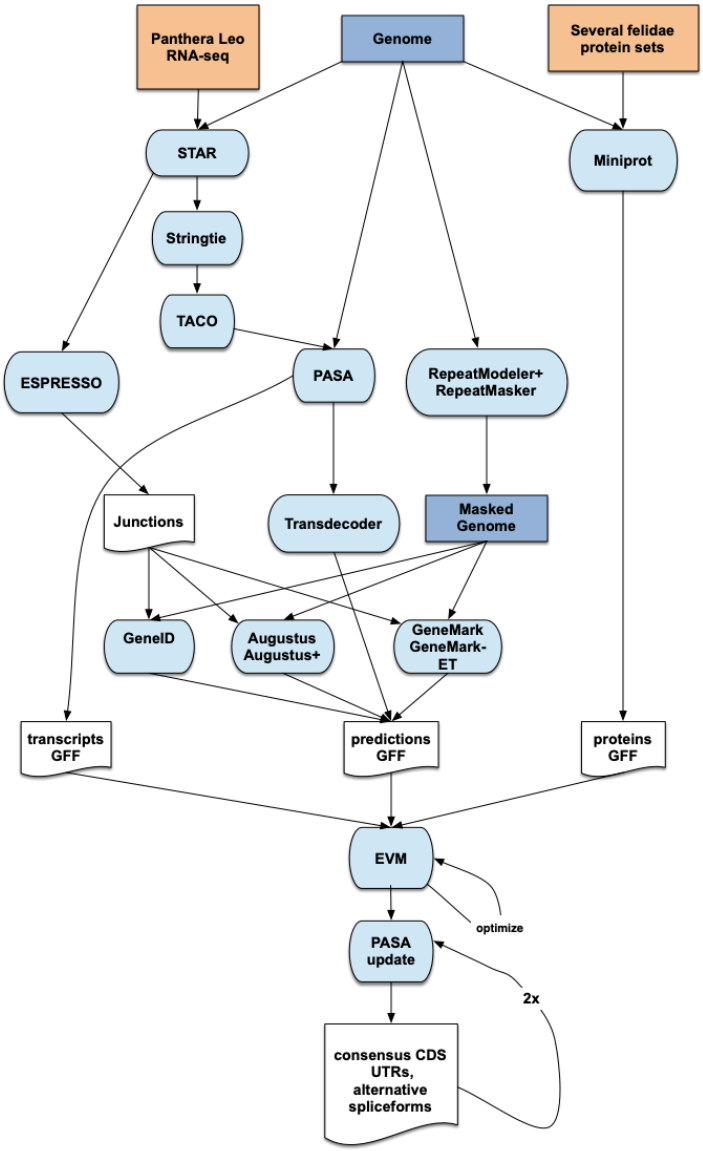
Genome annotation

In total, we annotated 23,886 protein-coding genes that produce 40,347 transcripts (1.69 transcripts per gene) and encode for 35,590 unique protein products. We were able to assign functional labels to 85.3% of the annotated proteins. The annotated transcripts contain 10.14 exons on average, with 88% of them being multi-exonic (**Table ANN1**). In addition, 44,770 non-coding transcripts were annotated, of which 25,998 and 18,772 are long and short non-coding RNA genes, respectively.

Table ANN1: Genome annotation statistics

	PAON1A annotation
Number of protein-coding genes	23,886
Median gene length (bp)	13,036
Number of transcripts	40,347
Number of exons	231,818
Number of coding exons	213,568
Median UTR length (bp)	1,358
Median intron length (bp)	1,377
Exons/transcript	10.14
Transcripts/gene	1.69
Multi-exonic transcripts	88%
Gene density (gene/Mb)	9.94

Figure ANN1: workflow of the genome annotation process



Bibliography

1. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
2. 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.
3. 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.
4. 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.
5. 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.
6. Li, H. 2023, Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
7. Alioto, T., Blanco, E., Parra, G., and Guigó, R. 2018, Using geneid to Identify Genes. *Curr Protoc Bioinformatics*, **64**, e56.
8. 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.
9. 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.
10. Törönen, P., and Holm, L. 2022, PANNZER —A practical tool for protein function prediction. *Protein Science*, **31**, 118–28.
11. 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.
12. Nawrocki, E. P., and Eddy, S. R. 2013, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–5.
13. Chan, P. P., and Lowe, T. M. 2019, tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol*, **1962**, 1–14.