Podarcis lilfordi 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 rPodLil1.1 genome assembly were annotated with RepeatMasker v4-1-2 (http://www.repeatmasker.org) using the custom repeat library available for podarcis. 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 Lilford's wall lizard 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 five different tissues (heart, kidney, liver, lungs and tail) was obtained and sequenced with both Illumina RNAseq and PacBio IsoSeq. After sequencing, the long and short reads were aligned to the genome using, respectively, STAR (2) v-2.7.2a and MINIMAP2 (3) v2.14 with the splice option. Transcript models were subsequently generated using Stringtie (4) v2.1.4 on each BAM file and then all the models produced were combined using TACO (5) v0.6.3. High-quality junctions to be used during the annotation process were obtained by running Portcullis (6) v1.2.0 after mapping with STAR and MINIMAP2. Finally, PASA assemblies were produced with PASA (7) 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 proteomes of *Podarcis muralis, Pogona vitticeps* and *Pantherophis guttatus* were downloaded from Uniprot in April 2002 and aligned to the genome using Spaln (8) v2.4.03. *Ab initio* gene predictions were performed on the repeat-masked rPodLil1.1 assembly with three different programs: GeneID (9) v1.4, Augustus (10) v3.3.4 and Genemark-ES (11) v2.3e 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 Diamond Blastp (13) search was made against the nr database (last accessed May 2022). Furthermore, Interproscan (14) 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 (15) v1.1 that is part of the Infernal (16) package was run against the RFAM database of RNA families (16) v12.0. Additionally, tRNAscan-SE (17) 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

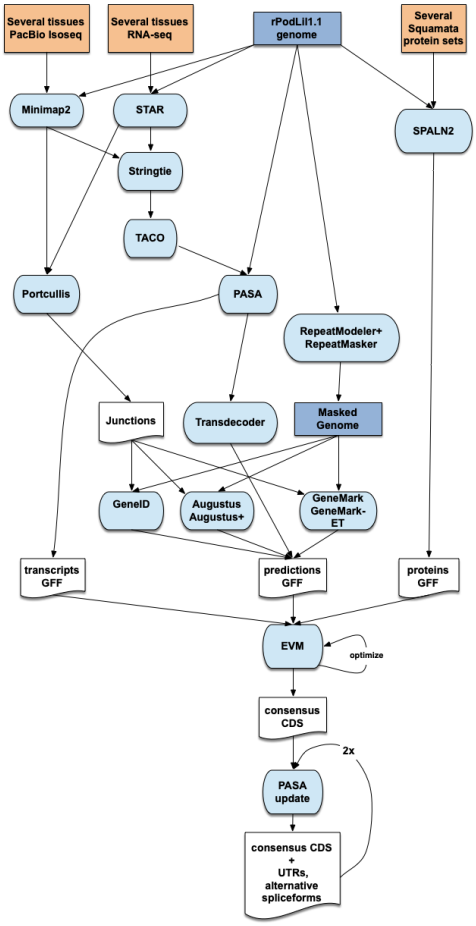
Genome annotation

In total, we annotated 25,678 protein-coding genes that produce 43,594 transcripts (1.7 transcripts per gene) and encode for 38,631 unique protein products. We were able to assign functional labels to 72% of the annotated proteins. The annotated transcripts contain 10.9 exons on average, with 91% of them being multi-exonic (**Table ANN1**). In addition, 47,087 non-coding transcripts were annotated, of which 12,794 and 34,293 are long and short non-coding RNA genes, respectively.

Table ANN1: Genome annotation statistics

	PODLIA annotation
Number of protein-coding genes	25,678
Median gene length (bp)	13,439
Number of transcripts	43,594
Number of exons	247,241
Number of coding exons	232,510
Median UTR length (bp)	2,138
Median intron length (bp)	1,296
Exons/transcript	10.9
Transcripts/gene	1.7
Multi-exonic transcripts	91%
Gene density (gene/Mb)	17.58

Figure ANN1: workflow of the genome annotation process



Bibliography

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

3. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

4. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar;33(3):290–5.

5. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods. 2017 Jan;14(1):68–70.

6. Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. Gigascience. 2018 Dec 1;7(12).

7. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008 Jan 11;9(1):R7.

8. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012 Nov 1;40(20):e161.

9. Alioto T, Blanco E, Parra G, Guigó R. Using geneid to Identify Genes. Curr Protoc Bioinformatics. 2018 Dec;64(1):e56.

10. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006 Feb 9;7:62.

11. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014 Sep;42(15):e119.

12. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005 Sep 15;21(18):3674–6.

13. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021 Apr 7;18(4):366–8.

14. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236–40.

15. 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 Jun 15;32(12):i332–40.

16. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013 Nov 15;29(22):2933–5.

17. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol. 2019;1962:1–14.