Clodip2 genome annotation report

Authors: Jèssica Gómez-Garrido, Tyler Alioto

Methods

Repeats present in the clodip2 genome assembly were annotated with RepeatMasker v4-0-6 (http://www.repeatmasker.org/) using the repeat library specific for our assembly that was built with RepeatModeler v1.0.11. Repeats that were part of repetitive protein families (detected by running a Blast of the Repeat library against swissprot) were removed from the library before masking the genome.

The gene annotation of the *Cleon dipterum* 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 1.

Firstly, RNAseq reads from several conditions (subprojects MAYFLY_04 and MAYFLY_05) were aligned to the genome with STAR [1](v-2.6.1b). Transcript models were subsequently generated using Stringtie [2] (v1.0.4) and PASA assemblies were produced with PASA [3] (v2.3.3) by adding also the transcripts obtained in a previous annotation attempt. 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 *Drosophila melanogaster* and *Anopheles gambiae* transcriptomes were downloaded from Uniprot in February 2019 and aligned to the genome using Spaln [4] (v2.3.1). *Ab initio* gene predictions were performed on the repeat masked clodip2 assembly with three different programs: GeneID [5] v1.4, Augustus [6] v3.2.3 and Genemark-ES [7] v2.3e with and without incorporating evidence from the RNAseq data. The gene predictors were run with parameters trained for drosophila, except Genemark that runs on a self-trained manner. Finally, all the data was combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM [3]). Additionally, UTRs and alternative splicing forms were annotated through two rounds of PASA annotation updates. Functional annotation was performed on the annotated proteins with Blast2go [8]. First, a Blastp [9] search was made against the nr database (last accessed February 2019). Furthermore, Interproscan [10] 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 produced by running the following steps. First, the program cmsearch [11] (v1.1) that comes with Infernal [12] was run against the RFAM [13] database of RNA families (v12.0). Also, tRNAscan-SE [14] (v1.23) was run in order to detect the tranfer RNA genes present in the genome assembly. To detect the lncRNAs we selected those Pasa-assemblies that had not been included into the annotation of protein-coding genes in order to get all those expressed genes that were not translated into a protein. Finally, those Pasa-assemblies without protein-coding gene annotation that were longer than 200bp and whose length was not covered at least in an 80% by a small ncRNA were incorporated into the ncRNA annotation as lncRNAs. The resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene.

Results

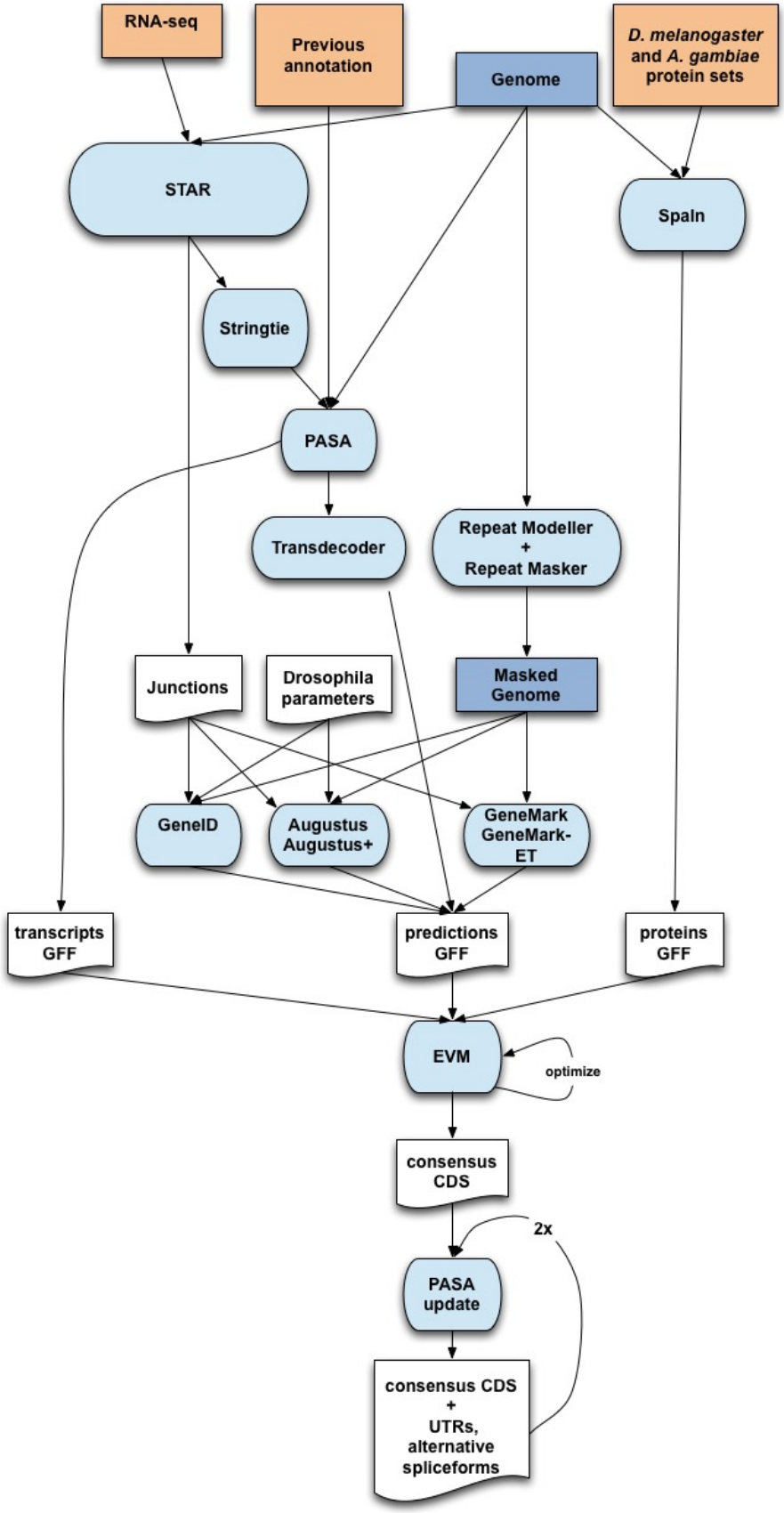
In total, we have annotated 14687 protein-coding genes, that produce 48186 transcripts (3.28 transcripts per gene) and encode for 34373 unique protein products. We have been able to assign functional labels to 68.48% of the annotated proteins. The annotated transcripts contain

11 exons on average, with 47559 of them being multi-exonic (Table 1). In addition, 5785 non-coding transcripts have been annotated, of which 4951 and 834 are long and short non-coding RNA genes, respectively.

Table 2: Genome annotation statistics

	Cdip2A annotation
Number of protein-coding genes	14687
Median gene length (bp)	4479
Number of transcripts	48186
Number of exons	181598
Number of coding exons	152833
Coding GC content	51.42%
Exons/transcript	11
Transcripts/gene	3.28
Multi-exonic transcripts	47559 (95.9%)
Gene density	12.28 kb

Table 1: Genome annotation pipeline flowchart



References

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
2. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol* 2015, **33:**290-295.
3. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol* 2008, **9:**R7.
4. Gotoh O: **A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence.** *Nucleic Acids Res* 2008, **36:**2630-2638.
5. Parra G, Blanco E, Guigo R: **GeneID in Drosophila.** *Genome Res* 2000, **10:**511-515.
6. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.
7. Lomsadze A, Burns PD, Borodovsky M: **Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.** *Nucleic Acids Res* 2014, **42:**e119.
8. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
10. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
11. 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, **32:**i332-i340.
12. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics* 2013, **29:**2933-2935.
13. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD: **Rfam 12.0: updates to the RNA families database.** *Nucleic Acids Res* 2015, **43:**D130-137.
14. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.