Insect Biochemistry and Molecular Biology 62 (2015) 2–10

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

Integrated modeling of protein-coding genes in the *Manduca sexta* genome using RNA-Seq data from the biochemical model insect☆

Xiaolong Cao, Haobo Jiang*

Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

ARTICLE INFO

Article history:
Received 29 September 2014
Received in revised form
8 January 2015
Accepted 11 January 2015
Available online 20 January 2015

Keywords:
Gene annotation
de novo assembly
Tobacco hornworm
Automated gene modeling
Arthropod genomics

ABSTRACT

The genome sequence of *Manduca sexta* was recently determined using 454 technology. Cufflinks and MAKER2 were used to establish gene models in the genome assembly based on the RNA-Seq data and other species' sequences. Aided by the extensive RNA-Seq data from 50 tissue samples at various life stages, annotators over the world (including the present authors) have manually confirmed and improved a small percentage of the models after spending months of effort. While such collaborative efforts are highly commendable, many of the predicted genes still have problems which may hamper future research on this insect species. As a biochemical model representing lepidopteran pests, *M. sexta* has been used extensively to study insect physiological processes for over five decades. In this work, we assembled *Manduca* datasets Cufflinks 3.0, Trinity 4.0, and Oases 4.0 to assist the manual annotation efforts and development of Official Gene Set (OGS) 2.0. To further improve annotation quality, we developed methods to evaluate gene models in the **M**AKER2, **C**ufflinks, **O**ases and **T**rinity assemblies and selected the best ones to constitute MCOT 1.0 after thorough crosschecking. MCOT 1.0 has 18,089 genes encoding 31,666 proteins: 32.8% match OGS 2.0 models perfectly or near perfectly, 11,747 differ considerably, and 29.5% are absent in OGS 2.0. Future automation of this process is anticipated to greatly reduce human efforts in generating comprehensive, reliable models of structural genes in other genome projects where extensive RNA-Seq data are available.

1. Introduction

With five larval instars, a large body size and hemolymph volume, and a simple larval body structure, the tobacco hornworm *Manduca sexta* has been widely employed as a model organism to study basic physiological processes in insects, such as cuticle formation, neural transmission, hormonal regulation, nutrient transport, intermediary metabolism, and immune responses (Hopkins et al., 2000; Shields and Hildebrand, 2001; Riddiford et al., 2003; Kanost et al., 1990; Arrese and Soulages, 2010; Jiang et al., 2010). Acquired knowledge of the molecular mechanisms underlying these processes would lead to new means of pest control, because *M. sexta* may be a good representative of some serious agricultural pests in the order of Lepidoptera. Several transcriptome analyses have yielded sequences and expression patterns of genes related to immunity, digestion, and olfaction (Zou et al., 2008; Pauchet et al., 2010; Zhang et al., 2011; Grosse-Wilde et al., 2011; Gunaratna and Jiang, 2013), but the potential of this model species is far from fulfillment partly due to the lack of its genome sequence. The shortage of complete protein sequences based on correctly modeled genes substantially hampers proteomic studies, for instance, of the immune complex formed around entomopathogens.

Recently, the genomic DNA isolated from a single male pupa of *M. sexta* was pyrosequenced at >20-fold coverage and assembled into *Manduca* Genome Assembly 1.0 (Msex 1.0) using Newbler with Atlas-GapFill (X et al., 2015). Sixty cDNA libraries, representing mRNA samples of whole larvae, organs and tissues at various developmental stages, were sequenced using Illumina technology,

Abbreviations: OGS, official gene set; ORF, open reading frame; L, length; ML, match length; QL, query length; SL, subject length; M, MAKER; C, Cufflinks; T, Trinity; O, Oases; U, UniProt Arthropoda; Y, C/T/O; S, similarity ratio of lengths; MLI, match length index; S1/S2, Selection 1 or 2; "P", perfect; "N", near perfect; "O", okay; "B", bad; "W", worst.

☆ The sequence files of MCOT 1.0 transcripts and proteins are available to download at ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/OGS2/OSU_files/. BLAST search of the two datasets can be performed at http://agripestbase.org/manduca/?q=blast.

* Corresponding author. Tel.: +1 405 744 9400; fax: +1 405 744 6039.
E-mail address: haobo.jiang@okstate.edu (H. Jiang).

http://dx.doi.org/10.1016/j.ibmb.2015.01.007
0965-1748/

yielding >350 gigabyte data. Some of these RNA-Seq datasets and other known *M. sexta* cDNA sequences were aligned to the reference genome to generate *Manduca* Cufflinks Assembly 1.0 and 1.0b using Bowtie, TopHat, and Cufflinks. Aided by the available sequence data from *M. sexta* and other arthropod species, approximately 18,000 genes in Msex 1.0 were predicted by MAKER2 generating the *Manduca* Official Gene Set 1.0 (OGS 1.0). Some of the OGS 1.0 models were examined by annotators to detect errors using *Manduca* Cufflinks 1.0/1.0b, Trinity 3.0, and Oases 3.0 sequences. The latter two sets of gene transcripts, assembled solely based on the RNA-Seq datasets, were extensively used along with Cufflinks 1.0/1.0b to improve annotation quality. Over a period of more than one year, 2498 structural genes were successfully curated by approximately 70 researchers (X et al., 2015). PASA2 (http://pasa.sourceforge.net/) was then used to select the best models from the MAKER2, Cufflinks, Trinity, Oases, and manual assemblies to generate *Manduca* OGS 2.0 (X et al., 2015).

During the course of gene cross-examination, we came to realize that some of the lessons learned can be valuable to future genome projects. For example, as extensive RNA-Seq data are becoming a norm, genome-dependent and independent assemblies are critically important in the validation and perfection of MAKER2 gene models. Due to limitations of the programs used to produce OGS 2.0 (Table 1), an integration of their outputs using computer programs may greatly reduce human efforts in sequence cross-examination and considerably increase the percentage of crosschecked gene models. To achieve these goals, we have developed methods to evaluate models in the **M**AKER, **C**ufflinks, **O**ases and **T**rinity assemblies. As proof of principle, a reliable, nearly complete set of protein sequences (MCOT 1.0) is generated to facilitate proteomic research in this model insect. In the following, we report the generation of Cufflinks 3.0, Oases 4.0 and Trinity 4.0 gene models, discuss their advantages, shortcomings and integration, and describe how MCOT 1.0 was developed and compared with OGS 2.0.

2. Materials and methods

2.1. Data and program acquisition

Manduca Genome Assembly 1.0 (Msex 1.0) and gene models in *Manduca* Official Gene Sets 1.0 (OGS 1.0, Table S1) and 2.0 (OGS 2.0) and Cufflinks Assembly 1.0 (Cufflinks 1.0) (X et al., 2015) were downloaded from *Manduca* Base (ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/). Universal protein sequences in UniProtKB Arthropoda (Table S1) were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/. The RNA-Seq datasets (X et al., 2015) were acquired from Dr. Gary Blissard at Cornell University. SAMtools (0.1.19) (Li et al., 2009), Bowtie2 (2.2.1) (Langmead and Salzberg, 2012), TopHat (2.0.11) (Trapnell et al., 2009), Cufflinks (2.1.1) (Trapnell et al., 2012; Roberts et al., 2011), Trinity (20131110) (Grabherr et al., 2011), Oases (0.2.08) (Schulz et al., 2012), and BLAST+ (2.2.29) (Camacho et al., 2009) were downloaded from http://samtools.sourceforge.net/, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, http://ccb.jhu.edu/software/tophat/index.shtml, http://cufflinks.cbcb.umd.edu/, http://trinityrnaseq.sourceforge.net/, https://www.ebi.ac.uk/~zerbino/oases/, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ and installed on a local supercomputer according to their manuals.

2.2. Generation of Cufflinks 3.0

The 60 RNA-Seq datasets were aligned to Msex 1.0 using TopHat at settings for three different read types: single end, paired end, and strand specific. "–read-realign-edit-dist 0" was selected to increase accuracy of read alignments. Cufflinks was used to translate the accepted hits generated by TopHat to separate GTF files, with the "-u" command enabled to allow more accurate handling of multiple reads mapped to the same region. Cuffmerge was employed to combine GTF files of all the libraries to make the final GFF file (see scripts in the Supplemental Materials), from which transcript sequences were extracted using gffread to form Cufflinks 3.0 dataset (Table S1).

2.3. Reads treatment, normalization, and de novo assembling

Paired end reads were trimmed to 80 bp using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), with the forward reads combined in one file and the reverse ones in another. To handle the RNA-Seq data with 256 GB RAM of the supercomputer, the number of the reads was reduced according to Haas et al. (2013). The Perl scripts provided in Trinity were used to perform *in silico* read normalization with maximum coverage set to 500. The single end and strand-specific reads were combined in one file for normalization at the same maximum coverage. After all normalized reads were pooled, Trinity was used to assemble the reads as paired end reads, generating Trinity 4.0 (Table S1). For Oases, four hash lengths (*k*: 25, 27, 29, 31) were chosen to assemble the reads as single end reads in four separate runs. Scaffolding was not allowed, preventing the stretches of Ns in assembled transcripts. The transcript files were then merged according to the Oases manual, generating Oases 4.0 (Table S1). In addition, reads that cannot be aligned to Msex 1.0 by TopHat were combined, trimmed to 80 bp, and assembled as paired end reads using Trinity. This new assembly

Table 1
Comparison of the four gene prediction programs.

Program	Algorithm	Advantages	Disadvantages
Cufflinks	Map reads to the reference genome with TopHat and Bowtie to identify splice sites, and then use outputs of TopHat to create gene models	Most sensitive; accurate splicing sites; GTF file for gene annotation; fast, less computation; more tolerant to low quality reads	Carry errors in the genome scaffolds (gaps, NNNs, misassembling, etc.); many isoforms from closely located and related genes do not exist
Maker2	Align EST and protein sequences to genome to produce *ab initio* gene predictions and can use RNA-Seq data to improve the prediction.	Less redundant; model genes poorly represented in the RNA-Seq datasets; GTF file for gene annotation	Low quality of predictions, such as extra or skipped exons, inaccurate splicing junctions, and merging of adjacent genes; biased on proteins
Trinity	De novo assemble transcripts using RNA-Seq data	Not influenced by problems in the genome assembly	Single hash level (*k*: 25); less sensitive than Cufflinks; redundant transcripts; no GTF file; SNPs etc.
Oases	De novo assemble transcripts using RNA-Seq data, and use Velvet for contig assembling	Accurate, not influenced by problems in the genome assembly, multiple hash levels to improve quality of transcript assembly	Less sensitive than Cufflinks, redundant transcripts; intense computation for large datasets; no GTF file; SNPs and other variations