Insect Biochemistry and Molecular Biology xxx (2015) 1–9

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*

A. Marchant [a, *], F. Mougel [a, b], V. Mendonça [a, c], M. Quartier [a, d], E. Jacquin-Joly [e], A. da Rosa [c], E. Petit [a, b], M. Harry [a, b, *]

[a] UMR EGCE (Laboratoire Evolution, Génomes, Comportement, Ecologie), Univ. Paris-Sud, CNRS, IRD, IDEEV, Univ. Paris-Saclay, Avenue de la Terrasse, Bâtiment 13, BP1 – 91198 Gif-sur-Yvette, France
[b] Université Paris Sud, Orsay, France
[c] Faculdade de Ciências Farmacêuticas, Universidade Estadual Paulista (UNESP), Araraquara, Brazil
[d] University of Neuchâtel, Institute of Biology, Neuchâtel, Switzerland
[e] INRA, UMR 1392, Institut d'Ecologie et des Sciences de l'Environnement de Paris (iEES-Paris), Versailles, France

ARTICLE INFO

Article history:
Received 21 October 2014
Received in revised form
8 April 2015
Accepted 15 May 2015
Available online xxx

Keywords:
Rhodnius prolixus
Chagas disease vectors
Reference transcriptome
de novo assembly
Referenced-based assembly
OBP
CSP

ABSTRACT

High Throughput Sequencing capabilities have made the process of assembling a transcriptome easier, whether or not there is a reference genome. But the quality of a transcriptome assembly must be good enough to capture the most comprehensive catalog of transcripts and their variations, and to carry out further experiments on transcriptomics. There is currently no consensus on which of the many sequencing technologies and assembly tools are the most effective.

Many non-model organisms lack a reference genome to guide the transcriptome assembly. One question, therefore, is whether or not a reference-based genome assembly gives better results than *de novo* assembly. The blood-sucking insect *Rhodnius prolixus*—a vector for Chagas disease—has a reference genome. It is therefore a good model on which to compare reference-based and *de novo* transcriptome assemblies.

In this study, we compared *de novo* and reference-based genome assembly strategies using three datasets (454, Illumina, 454 combined with Illumina) and various assembly software. We developed criteria to compare the resulting assemblies: the size distribution and number of transcripts, the proportion of potentially chimeric transcripts, how complete the assembly was (completeness evaluated both through CEGMA software and *R. prolixus* proteome fraction retrieved). Moreover, we looked for the presence of two chemosensory gene families (Odorant-Binding Proteins and Chemosensory Proteins) to validate the assembly quality. The reference-based assemblies after genome annotation were clearly better than those generated using *de novo* strategies alone. Reference-based strategies revealed new transcripts, including new isoforms unpredicted by automatic genome annotation. However, a combination of both *de novo* and reference-based strategies gave the best result, and allowed us to assemble fragmented transcripts.

1. Introduction

Blood-sucking insects in the sub-family Triatominae (Hemiptera, Reduviidae) are vectors of *Trypanosoma cruzi* (Kinetoplastida, Trypanosomatidae), which causes Chagas disease. Triatomines are obligate hematophages at all larval stages. *Rhodnius prolixus* has been a model for insect physiology and biochemistry for many decades (beginning with the early study by Wigglesworth, 1936). Because it was the main vector of Chagas disease in Central America, vector control programs have been developed to target domestic populations. Strong evidence now indicates that *R. prolixus* no longer transmits *Trypanosoma cruzi* in some northern countries of Latin America (Coura and Coura, 2013; Hashimoto and Schofield, 2012). But new vectors are now emerging that bring concerns about the potential health impacts.

Besides *R. prolixus*, some other *Rhodnius* species are now domiciliated—they are colonizing human dwellings. These include *Rhodnius ecuadoriensis* in Ecuador and *Rhodnius pallescens* in

* Corresponding authors. UMR EGCE (Laboratoire Evolution, Génomes, Comportement, Ecologie), Univ. Paris-Sud, CNRS, IRD, IDEEV, Univ. Paris-Saclay, Avenue de la Terrasse, Bâtiment 13, BP1 – 91198 Gif-sur-Yvette, France.
E-mail addresses: axelle.marchant@egce.cnrs-gif.fr (A. Marchant), myriam.harry@egce.cnrs-gif.fr (M. Harry).

http://dx.doi.org/10.1016/j.ibmb.2015.05.009
0965-1748/

Panama. Species like *Rhodnius nasutus* and *Rhodnius neglectus* are in the process of domiciliation in Brazil. A better understanding of the ecology, biochemistry and physiology of Triatomes will be helpful for developing new tools to control these disease vectors by disrupting chemical communication. In particular, we need to understand what is happening at the molecular level, by targeting chemosensory and behavioral genes that allow insects to interact with their environment (chemosensory recognition of food sources, breeding and oviposition sites) and with their congeners. Our goal here was to facilitate that understanding by building a reference transcriptome of *R. prolixus.*

Recent studies using High Throughput Sequencing (HTS) technologies have revolutionized the field of functional genomics and gene expression, and significantly improved genome annotation (see references in Carver et al., 2012). They opened the field to a wide range of applications from *de novo* sequencing to functional genomics and transcriptomics. Transcriptomics can be used to provide sequence information for *de novo* transcript assembly to improve gene annotations in species where the genome is not sequenced; detect differential expression at cell, tissue or population levels; detect new coding transcripts, non-coding RNAs or alternative splicing isoforms; and discover structural variations in transcripts such as SNPs or indels (see reviews: Burgess, 2013; Marguerat and Bähler, 2010; Mortazavi et al., 2008; Oshlack et al., 2010). Different sequencing technologies (Illumina, SOLiD, 454) and protocols (single end or paired ends) are currently available to perform RNA-Seq experiments. Several articles and reviews describe the current usage and limitations of these biological techniques (Wang et al., 2009; Wilhelm et al., 2010). Though many new algorithms and tools have been developed to efficiently process and analyze transcriptomic data (see for review Carver et al., 2012), there is no clear consensus on the best ways to perform assembly analysis.

To date, transcriptome studies on Triatominae have been mostly targeted sialotranscriptomes of various species because they are of pharmacological interest. Sialomes have been sequenced from salivary glands in species within two genera: *Rhodnius,* including *R. prolixus*, (Bussacos et al., 2011; Ribeiro et al., 2004), *Rhodnius brethesi* and *Rhodnius robustus* (Bussacos et al., 2011), and *Triatoma*, including *T. brasiliensis* (Santos et al., 2007) and *Triatoma infestans* (Assumpção et al., 2008). The gene expression profile of *R. prolixus* ovarian follicle tissue have also been surveyed (Medeiros et al., 2011). Moreover, recently, Ribeiro et al. (2014) analyzed the transcriptome of the *R. prolixus* digestive tract using 454 technology. With the aim to build a reference transcriptome for the non-model insect *R. prolixus*, which lacked a reference genome, Marchant et al. (2014) performed *de novo* hybrid assembly using several programs and merged contigs from Illumina paired-end reads and 454 contigs corrected with Illumina reads.

A genome assembly for *R. prolixus* is available (http://metazoa.ensembl.org/Rhodnius_prolixus/Info/Index). In this study, we compare both *de novo* and reference-based strategies using various criteria, including contig lengths, N50, contig number, chimeric contigs, completeness and the fraction of the *R. prolixus* proteome retrieved. Moreover, to validate the assembly quality, we evaluate the presence of a gene family; namely, Odorant-Binding Proteins (OBPs) and Chemosensory Proteins (CSPs).

2. Materials and methods

2.1. Insects mRNA extraction and next-generation sequencing

We used two *R. prolixus* strains: the 077 strain maintained at the Insetário de Triatominae da Faculdade de Ciências Farmacêuticas, Universidade Estadual Paulista (UNESP) – Araraquara and Laboratório Nacional e Internacional de Referência em Taxonomia de Triatomíneos/FIOCRUZ/Rio de Janeiro (males: PMEF; females: PFEF) and a strain maintained at the Institute of Biology, University of Neuchâtel, Switzerland (males: PMES; females: PFES). Samples included several individuals pooled by origin and by sex (Table 1). To target genes involved in the chemosensory system, RNA was extracted from the antennae and rostrum of Brazilian strains (PMEF and PFEF), and from the antennae, rostrum and head of Swiss strains (PMES and PFES) using the TRIzol® Reagent kit. Because sequencing with 454 technology requires a lot of RNA (5–10 μg), we pooled RNA aliquots extracted from the antennae and rostrum from UNESP laboratory strain (3 μg for PMEF, 3 μg for PFEF). To complete the sample, we used RNA extracted from the head, antennae and rostrum of the University of Neuchâtel strain. Since our goal was to favor chemosensory genes, we used less RNA from the latter strain (1 μg for PMES, 1 μg for PFES). The RNA mix (8 μg) was used as template for constructing a concatamerized, normalized cDNA library prepared and sequenced using 454 Roche GS FLX Titanium (½ Pico Titer Plate) by LGC Genomics GmbH (Berlin, Germany). We obtained 508,191 reads with a median length of 300 bp (279 bp after reads cleaning). Independent Illumina libraries, which requires less RNA, were prepared for three RNA samples (PMEF, PFEF, PMES) using 0.5 μg RNA aliquot per sample. Following rRNA removal (Ribo-Zero kit, Epicentre), the RNA was used to prepare cDNA libraries using the ScriptSeq™ mRNA-Seq Library Preparation Kit (Epicentre). Libraries were sequenced in paired-end reads (2 × 100 bp) on Illumina® HiSeq 2000 platform (Genomic Platform, IBENS, Institut de Biologie de l'École Normale Supérieure, Paris, France).

2.2. Assembly of transcriptomes

- De novo assemblies

Three assemblies were compared using 454 reads alone, Illumina reads alone or mixing both 454 and Illumina reads for each library (Fig. 1). Illumina and 454 reads were first cleaned using cleaned-1 criterion outlined in Marchant et al. (2014) with the Prinseq version 0.20.4 (Schmieder and Edwards, 2011). This version cleans forward and reverse reads simultaneously. When one read from a pair was discarded, the other was automatically eliminated. 454 reads were assembled with Newbler, CAP3 (Huang and Madan, 1999) and MIRA (Chevreux et al., 2004) according to the workflow described in detail in Marchant et al. (2014, see assembly 4). Illumina reads were assembled using Trinity (Grabherr et al., 2011).

A hybrid assembly was then performed according to Marchant et al. (2014, see assembly 10). Possible homopolymer tracts were corrected by aligning Illumina reads on the 454 contigs using BWA (Li and Durbin, 2009). 454 corrected contigs and Illumina contigs were reassembled in "super-contigs" using CAP3 and repetitions were deleted using cd-hit-est (Li and Godzik, 2006).

- Reference-based assemblies

Two assemblies were generated for each of the three Illumina samples by mapping raw data on the genome assembly *R. prolixus* RproC1.25.dna.genome.fa (http://metazoa.ensembl.org) using TopHat (defaults options; Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2010, Fig. 1). The first assembly was done without specifying gene annotation and was named Cufflinks WA (without annotation; Fig. 1). In the second assembly, we added the *R. prolixus* gene annotation (version Rhodnius_prolixus.RproC1.25.gff3 of Ensembl) to Tophat (-G option) and Cufflinks (-GTF guide option) to guide the RABT (reference annotation based transcript) assembly. The output of RABT assembly included all annotated transcripts (with or without mapped reads), new isoforms and transcripts

Please cite this article in press as: Marchant, A., et al., Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*, Insect Biochemistry and Molecular Biology (2015), http://dx.doi.org/10.1016/j.ibmb.2015.05.009