Genomics Data 11 (2017) 9–16

Contents lists available at ScienceDirect

Genomics Data

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

Traditional *versus* 3′ RNA-seq in a non-model species

Sophie Tandonnet [a], Tatiana Teixeira Torres [a,b,*]

[a] Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, SP, Brazil
[b] Research Center on Biodiversity and Computing (BioComp-USP), Universidade de São Paulo (USP), São Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 23 August 2016
Received in revised form 1 November 2016
Accepted 2 November 2016
Available online 18 November 2016

Keywords:
Gene expression profiling
Length bias
Transcriptome
Cochliomyia hominivorax
Calliphoridae

ABSTRACT

One limitation of the widely used RNA-seq method is that long transcripts are represented by more reads than shorter transcripts, resulting in a biased estimation of expression levels. The 3′ RNA-seq method, which yields only one sequence per transcript, bypasses this limitation. Here, RNA was extracted from two samples, in which we expected to find differentially expressed genes. Each was processed by both traditional and 3′ RNA-seq protocols. Both methods yielded similar differentially expressed genes and estimated expression levels in a comparable way, confirming they both represent valid tools for RNA-seq analysis. Notably, however, we identified more differentially expressed transcripts with the 3′ RNA-seq method, suggesting a greater power to detect expression variation using this method. Hence, when little genomic information is available for the species studied, the standard RNA-seq presents a better cost-benefit compromise, whereas for model species, the 3′ RNA-seq method might more accurately detect differential expression.

1. Introduction

The RNA-seq method is a powerful tool allowing for functional genomic studies at the transcriptional level. It consists of the deep sequencing of the RNA (total or fractionated) of an individual or tissue at a certain time and condition. This approach enables the comparative assessment of the level of expression for each gene between different samples. By comparing the RNA expression profiles among samples, it is possible to identify differentially expressed (DE) genes that might explain the phenotypic differences observed between the samples.

In this study, we compared two RNA-seq methods: the standard RNA-seq and the 3′ RNA-seq that is expected to give more accurate levels of expression by solving some of the biases inherent in the classic RNA-seq method. With the standard RNA-seq method, the extracted mRNA is randomly sheared and the fragments are converted into a cDNA library. The cDNA fragments are then sequenced by one of the next-generation sequencing technologies. The total number of reads (cDNA fragments sequenced) corresponding to a given transcript is proportional to the level of expression of the corresponding gene [11]. However, one of the limitations of the standard RNA-seq strategy lies in the fact that longer transcripts are broken into more fragments than are shorter ones. This creates a statistical bias, as longer transcripts will be represented by more reads than those produced by the shorter transcripts. Consequently, the detection of DE is more likely to be over-represented for long transcripts and under-represented for shorter ones, which are at a statistical disadvantage [11]. To minimize this bias, the levels of expression (number of reads corresponding to a certain transcript) can be corrected by the size of the transcript. However, in the case of non-model species, this information is most likely to be unavailable. The correction can then be done by using the contig size from the *de novo* reconstruction of the transcript (based on the reads) or by employing the transcript sizes of a closely related model species. Nevertheless, this correction does not entirely solve the problem owing to the transcript size, as the sampling is higher for longer transcripts [11].

The 3′ RNA-seq method [15] was conceived to bypass these limitations. This method consists of sequencing only one fragment per transcript in the 3′ region. By using this strategy, regardless of the transcript length, the levels of expression can be estimated directly by the number of reads corresponding to a certain transcript, as a single fragment per mRNA molecule is sampled (Fig. 1).

In this paper, we compare both RNA-seq methods at the different steps of an RNA-seq analysis to clarify their advantages, disadvantages, and complementarities for a non-model species, *Cochliomyia hominivorax*, the New World screw-worm fly. This species is one of the most important myiasis-causing fly of the neotropical region and is responsible for severe economic losses. During the last decades, *C. hominivorax* populations were mainly controlled by applying

* Corresponding author at: Department of Genetics and Evolutionary Biology, IB-USP, Rua do Matão, 277, 05508-090 São Paulo, SP, Brazil.
E-mail address: tttorres@ib.usp.br (T.T. Torres).

http://dx.doi.org/10.1016/j.gdata.2016.11.002

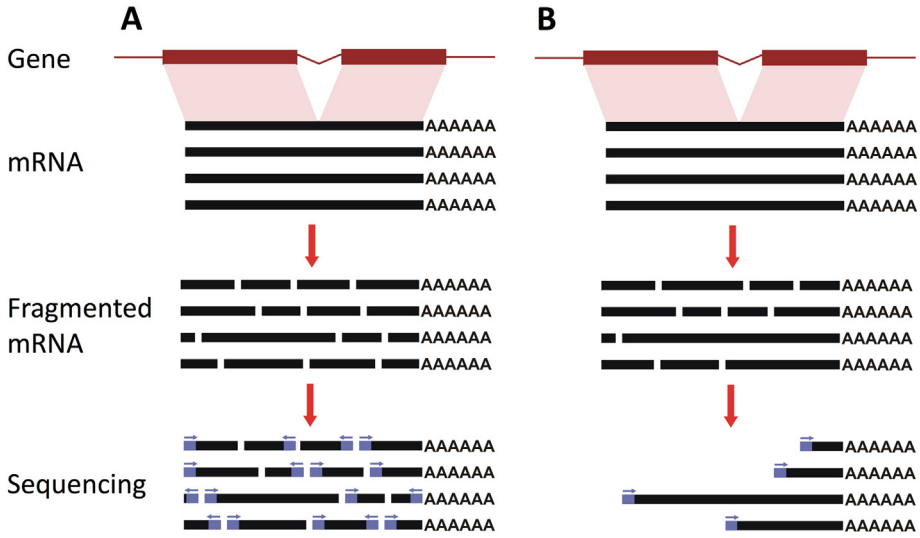

Fig. 1. Overview of the methods used to generate the RNA-seq libraries. (A) In the classic RNA-seq procedure, the RNA is fragmented and converted into cDNA using small primers of random sequence. (B) In the 3′ RNA-seq library, the mRNA molecules are randomly fragmented, generating fragments of different lengths. After fragmentation, only the 3′ portion of an mRNA molecule is selected using poly-T oligonucleotide baits attached to magnetic beads. The selected fragments (one per molecule) are then directionally sequenced.

organophosphate (OP) insecticides, but because of this constant selective pressure, resistant lineages have been strongly selected, complicating the management of this species [3,4]. In this context, the RNA-seq methods were used to discover the genes possibly involved in OP resistance.

2. Materials and methods

2.1. C. hominivorax populations

We used a laboratory colony of *C. hominivorax* composed of susceptible and known OP resistant individuals (Gly137Asp and/or Trp251Ser mutations in the esterase E3 gene), collected in Caiapônia, GO, Brazil. The colony was maintained according to standard protocols [2]. For the resistant condition, a sample from the laboratory population was treated with the OP insecticide dimethyl 2,2-dichlorovinyl phosphate; $C_4H_7Cl_2O_4P$ (dichlorvos) at 20 mg/l, a concentration lethal for 90% of the population (LC90). The insecticide was directly mixed into the medium consisting of fresh ground beef supplemented with blood and water (2:1). A total of 500 L2 instar larvae were fed on the insecticide-containing medium for 24 h. The surviving individuals (Resistant sample) were collected for the RNA extractions. The individuals of the control condition were simply sampled from this laboratory population and fed on the medium without the insecticide.

2.2. RNA extraction

RNA extractions followed previously utilized procedures [2]. Total RNA of resistant and control *C. hominivorax* larvae were extracted separately using TRIzol (Invitrogen) from the whole bodies of 87 larvae, 42 from the resistant and treated group and 45 from the control group. DNase I (Invitrogen) was used to remove genomic DNA contamination and the mRNA-enriched samples were further purified using Nucleospin RNA Clean-up columns (Macherey Nagel). RNA quantification was performed using the Qubit Quantitation Platform fluorometer (Invitrogen).

2.3. RNA-seq experiments

The extracted RNA was processed separately according to the two RNA-seq protocols. In the classic RNA-seq procedure, the RNA fragments resulting from the random breakage of the transcripts were converted into a cDNA library using the mRNA-Seq Sample Prep Kit (Illumina). Small primers (6 nt) of random sequence were used to produce the cDNA fragments. Specific adapter sequences (ACGTT and TGCAT for the control and resistant conditions, respectively) were prefixed to the cDNA fragments. These barcoded control and resistant cDNA sequences were then pooled prior to sequencing. Library preparation was performed independently twice on the same samples (technical replicates).

The 3′ RNA-seq library was constructed by Fasteris (Switzerland) using the procedure adapted from a previous study [15]. In this method, 4 μg total extracted RNA for each sample (control and resistant) was used to create the 3′ RNA libraries. A 3′ RNA library contains only those RNA fragments possessing a polyA tail. For its construction, the mRNA- Seq Sample Prep Kit (Illumina) was modified to select the 3′ RNA fragments. Briefly, the mRNA molecules were fragmented at a high temperature (80 °C) by divalent cations using the fragmentation buffer. The polyA mRNA fragments were purified using poly-T oligonucleotide baits attached to magnetic beads. After the selection of polyA fragments, the mRNA-seq Sample prep Kit (Illumina) protocol was followed according to the manufacturer's instructions. Consequently, we obtained one polyA-fragment per transcript molecule, which allowed us to directly estimate the expression level of the transcripts. Resistant and control samples were pooled prior to sequencing by Fasteris using the Illumina HiSeq100 system (single-reads of 100 bp).

2.4. Preprocessing of the reads

For a thorough comparison between the RNA-seq methods, we sampled the same number of raw reads obtained by both methods for each condition. Since the sequencing based on the cDNA obtained by the standard RNA-seq method yielded fewer reads (15,427,065 for control, 17,021,595 for resistant), we sampled those numbers of raw reads from both 3′ RNA-seq read populations (35,574,183 for control, 46,322,457 for resistant). The sampling was performed using the function “FastqSampler” from the R package “ShortRead” [10].

To eliminate poor quality regions of the sequences, we used the program fastq_quality_trimmer from the fastx toolkit suite (http://hannonlab.cshl.edu/fastx_toolkit/). We used the default quality score threshold of 20 and removed the sequences shorter than 20 bases after the trimming had being completed.