Genomics 97 (2011) 235–243

Contents lists available at ScienceDirect

Genomics

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

Computational dissection of *Arabidopsis* smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites

Xiangfeng Wang [a,b,*], John D. Laurie [a], Tao Liu [b], Jacqueline Wentz [c], X. Shirley Liu [b,**]

[a] School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive Tucson, AZ 85721, USA
[b] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA
[c] Department of Bioengineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 12 October 2010
Accepted 27 January 2011
Available online 2 February 2011

Keywords:
High-throughput sequencing
Small RNAs
Principal component analysis
TSS-associated RNAs

ABSTRACT

The profiling of small RNAs by high-throughput sequencing (smRNA-Seq) has revealed the complexity of the RNA world. Here, we describe a computational scheme for dissecting the plant smRNAome by integrating smRNA-Seq datasets in *Arabidopsis thaliana*. Our analytical approach first defines *ab initio* the genomic loci that produce smRNAs as basic units, then utilizes principal component analysis (PCA) to predict novel miRNAs. Secondary structure prediction of candidates' putative precursors discovered a group of long hairpin double-stranded RNAs (lh-dsRNAs) formed by inverted duplications of decayed coding genes. These gene remnants produce miRNA-like small RNAs which are predominantly 21- and 22-nt long, dependent of DCL1 but independent of RDR2 and DCL2/3/4, and associated with AGO1. Additionally, we found two classes of transcription start site associated (TSSa) RNAs located at sense (+) and antisense (−) approximately 100–200 bp downstream of TSSs, but are differentially incorporated into AGO1 and AGO4, respectively.

Published by Elsevier Inc.

1. Introduction

Plant genomes produce a variety of small RNA (smRNA) families to mediate either post-transcriptional or transcriptional gene silencing (PTGS or TGS). In *Arabidopsis*, three known classes of small RNAs functioning in PTGS comprise microRNAs (miRNAs), *trans*-acting siRNAs (tasiRNAs) and natural antisense transcript-derived siRNAs (natsiRNAs) that guide the cleavage of mRNAs [1–4]. The fourth class of endogenous siRNAs acting in TGS arises from the transposable elements (TEs) to mediate the epigenetic silencing of cognate TEs [5–8]. Those small RNAs are recently uniformly defined as *cis*-acting siRNAs (casiRNAs) [9]. Functional categorization of those small RNAs is based on their distinct mechanisms of biogenesis by a combination of different members of RNAi components encoded in *Arabidopsis* genome, which include four Dicer-like endonuleases (DCL1–4) [10], Pol II and other two plant-specific DNA-dependent RNA polymerases, Pol IV and Pol V [11,12], six RNA-dependent RNA polymerases (RDR1–6) and ten Argonautes (AGO1–10) [13].

Transcription of a miRNA gene (*MIR*) is dependent on Pol II. The primary transcript of a *MIR* gene is a long single-stranded RNA called pri-miRNA that contains an imperfect inverted repeat and is further cleaved into precursor miRNA (pre-miRNA) with a stem–loop structure. In plants, the two steps of processing from the pri-miRNAs to pre-miRNAs, and to mature miRNA duplexes are catalyzed by DCL1 [14]. While the guide strands of the miRNA duplexes are incorporated into AGO1 of the RNA-induced silencing complex (RISC), the passenger strands called miRNA star (miRNA*) are mostly degraded. Plant miRNAs are typically 21-nt long, preferentially started with a uracil at 5′ end. Unlike the animal miRNAs that target mRNA's 3′ UTR by the "seed regions (the 2nd to 8th nucleotide from a miRNA's 5′ end)", plant miRNAs are usually complementary to their targets' coding regions with near-perfect match to induce the cleavage [14].

In plants, tasiRNAs are discovered to have the similar function with miRNAs to regulate the gene silencing at posttranscriptional level, but in a manner of imperfect matching with their targets [15]. The genomic loci encoding tasiRNAs are known as *TAS* genes transcribed by Pol II, and the mature tasiRNA products are uniformly 21-nt long started with a U at 5′ ends. The third class of siRNAs in PTGS is natsiRNA whose long dsRNA precursors are formed by the hybridization of overlapping sense and antisense RNA transcripts caused by convergently transcribed genes or TEs [16].

In plants, casiRNAs are the most predominant class of small RNAs and are prevailingly produced from transposable elements, heterochromatic regions or other repetitive sequences. Therefore, casiRNAs are previously called TE-derived siRNAs, heterochromatic siRNAs (hcRNAs) or repeat-associated siRNAs (rasiRNAs) [4,7]. The functional role of casiRNAs is to direct the DNA methylation on the genomic loci where they originate from and silence the residing TEs in *cis* [17]. It also has been indicated that casiRNA pathways might influence the transcription of the neighboring protein-coding genes as they can

* Correspondence to: X. Wang, School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive Tucson, AZ 85721, USA.
** Corresponding author.
E-mail addresses: xwang1@cals.arizona.edu (X. Wang), xsliu@jimmy.harvard.edu (X.S. Liu).

0888-7543/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2011.01.006

modify the epigenetic states of upstream sequences [18,19]. The casiRNAs possess two signatures, 24-nt long and preferential A at 5′ end, which can be recognized by AGO4, a component of RNA-directed DNA methylation (RdDM) complex.

The high-throughput profiling of small RNAs by sequencing (smRNA-Seq) has revealed the complexity of the RNA population. Those exponentially accumulating smRNA-Seq datasets have created urgent challenges for quantitative interpretation of the results and *in silico* identification of new smRNA classes and pathways. In addition to those known miRNAs, tasiRNAs, natsiRNAs and casiRNAs, many functionally uncharacterized small RNAs have been observed to arise from structured genomic sites such as long inverted repeats, short hairpin repeats, and convergent genes, whose biogenesis pathways may differ from canonical mechanisms. Recently, several software packages and pipelines have been developed to cope with the large-scale analysis of smRNA-Seq datasets mainly aiming at two purposes: first, to process raw smRNA-Seq data and annotate the small RNAs in the genome; second, to build the expression profiles of known miRNAs and discover the new miRNAs [20–25].

The first way to identify new miRNAs from smRNA-Seq data is based on cross-species comparison, which is to directly align the reads with known miRNAs in other species such as adopted by miRExpress and DSAP [20,21]. The other way is to find new miRNAs according to the miRNA biogenesis pattern which is the features of how miRNA mature products are processed from pre-miRNA hairpin precursors. The original algorithm was developed by Friedländer et al. and was implemented as a software package called miRDeep [22]. miRDeep first extracts putative miRNA precursors with uniquely mapped smRNA reads and then rules out those overlapped with rRNA, snoRNA, tRNA loci etc., as well as those that cannot fold into canonical hairpin structures [22]. Next, miRDeep uses Bayes' theorem to calculate the probability of a potential miRNA precursor by comparing with background hairpins [22]. The algorithm of miRDeep was also integrated by other smRNA-Seq analysis tools to identify the new miRNAs such as deepBase and mirTools [23,24]. Another *de novo* miRNA prediction tool, miRanalyer utilizes machine learning approach to score the new miRNAs based on a variety of features such as read counts, stem and loop lengths, and folding energy etc. [25].

As miRNA is the predominant type of small RNAs in animals, most available smRNA-Seq tools focus on miRNA analysis. Although the basic concepts of miRNA prediction from smRNA-Seq are essentially the same for animals and plants, notable differences still exist. For example, while the animal miRNA precursors have more canonical hairpin structures with relatively fixed size of stem and loop regions, plants pre-miRNAs sometimes have longer hairpin stem regions and even multiple branches. Additionally, plant genomes contain a great number of inverted repeats formed by transposable elements that produce miRNA-like siRNAs, which are usually the source of false positive results from *de novo* miRNA prediction. Furthermore, as the majority of plant small RNAs are various types of siRNAs, a more comprehensive pipeline needs to be developed to annotate existing siRNAs and discover the new species. By integrating six smRNA-Seq datasets in different developmental stages and RNAi pathway mutations [26–30] (Supplementary Table 1 and Supplementary Fig. 1), we developed an analytical framework to dissect the *Arabidopsis* smRNAome and computationally discover previously uncharacterized miRNAs and other smRNA classes.

2. Materials and methods

2.1. Define smRNA-deriving loci as primary transcription units (Pri-TU)

We obtained the four libraries of processed Argonaute-associated (AGO1, AGO2, AGO4 and AGO5) smRNA-Seq dataset from Dr. Yijun Qi's group, in which the 5′ and 3′ adaptor sequences had been trimmed off from both ends of the sequencing reads. This dataset contains totally 2,840,770 high-quality reads that represent 599,449 unique small RNA sequences.

To determine the genomic locations of small RNA reads, we employed Bowtie [31] to map the ~600,000 unique smRNA sequences to *Arabidopsis* reference genome TAIR8 (http://www.arabidopsis.org/), and kept all locations that a read was perfectly aligned to. By bowtie, 599,397 of them were mapped to 2,654,309 locations without any mismatch. Thus, each unique small RNA sequence has two layers of information: (1) the *repetitiveness*, the number of the locations it was mapped to the genome without any mismatches, and (2) the *abundance*, the number of the reads for a unique small RNA being sequenced.

We developed a tool to *de novo* scan the genomic mapping result of smRNA-Seq reads to define the primary transcription units (Pri-TUs) that give rise to small RNAs. As Fig. S2 shows, for a putative Pri-TU, it was composed of a set of small RNAs that are overlapped or next to each other with a small gap (Supplementary Fig. 2). The initial *de novo* scanning identified 108,350 Pri-TUs with maximum 50 bp gap allowed, and at least 2 reads per Pri-TUs. Since most of the Pri-TUs containing very few reads might be resulted from the wrong mapping or background noise, we only used 23,516 Pri-TUs containing more than 20 reads for the further statistics.

During the identification of Pri-TUs, we also collected following information for each Pri-TU: (1) *SeqFreq* (sequencing frequency), which is the sum of the reads that a small RNA were being sequenced, to represent the expression abundance of a small RNA; (2) *RepFreq* (repetitive frequency), which is the sum of all the locations for a small RNA whose sequence was mapped in the genome, to represent the repetitiveness of a small RNA; (3) *UniqFreq* (unique frequency), which is the sum of the number of unique smRNA sequences within a Pri-TU, to represent the excision mode; (4) *AvgSeq* is the ratio of *SeqFreq*/*RepFreq*, which is the adjusted value of small RNA abundance by repetitive frequency; (5) *size* and (6) *5′ terminal-nt* is the most prevalent length and the type of 5′ terminal nucleotides of the small RNAs inside a Pri-TU, respectively. After the Pri-TUs were identified, we also calculated the following features including the frequency of the cutting sites of di-nucelotide where small RNAs were processed from Pri-TU, the proportions of 5′A, 5′G, 5′C and 5′U, the strand-bias that small RNA derived from plus and minus strand within a Pri-TU (Supplementary Table S2).

2.2. Computational selection of candidate Pri-TUs for new miRNA prediction by principal component analysis (PCA)

Computational selection of candidate Pri-TUs was based on the facts that miRNAs tend to be sequenced more (higher *SeqFreq*), but more accurately excised from pre-miRNA hairpins, and uniquely mapped in the genome (lower *RepFreq* and *UniqFreq*). We employed principal component analysis (PCA) on *SeqFreq*, *RepFreq* and *UniqFreq* to discriminate the Pri-TUs of producing miRNAs from the ones producing siRNAs [32]. The nature of PCA algorithm is to identify the direction (first principal component, PC1) with the largest variation, and the direction of the second and third principal components (PC2, PC3) uncorrelated to PC1. The three PCs were standardized to be centered at zero, and we used $PC1>0$, $PC2<0$ and $PC3<0$ to classify the miRNA-deriving Pri-TUs and siRNA-deriving Pri-TUs. After removing Pri-TUs associated with known miRNA genes, the rest candidate Pri-TUs will be used for further new miRNAs prediction. We next searched the candidate Pri-TU sequences against *Arabidopsis* TAIR8 annotation to further exclude the false positive candidates which were tasiRNAs, snRNAs, snoRNAs, tRNAs and rRNAs etc. whose secondary structure may contain hairpins. The second round screening narrowed the candidates down to those were absolutely located in the intergenic regions based on TAIR8's annotation. We then extracted the precursor sequences by extending at 35 bp on both end of a Pri-TUs to predict their secondary structures by RNAfold.