Computational Biology and Chemistry 34 (2010) 313–319

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Brief communication

# In silico identification of conserved microRNAs and their target transcripts from expressed sequence tags of three earthworm species

Ping Gong [a,*], Fuliang Xie [b], Baohong Zhang [b], Edward J. Perkins [c]

[a] SpecPro Inc., Environmental Services, 3909 Halls Ferry Road, Vicksburg, MS 39180, USA
[b] Department of Biology, East Carolina University, Greenville, NC 27858, USA
[c] Environmental Laboratory, U.S. Army Engineer Research and Development Center, Vicksburg, MS 39180, USA

ARTICLE INFO

*Article history:*
Received 13 May 2010
Received in revised form
28 September 2010
Accepted 28 September 2010

*Keywords:*
MicroRNA
Expressed sequence tag (EST)
Earthworm
Algorithm
Target prediction
Gene ontology
KEGG pathway

ABSTRACT

MicroRNAs are a recently identified class of small regulatory RNAs that target more than 30% protein-coding genes. Elevating evidence shows that miRNAs play a critical role in many biological processes, including developmental timing, tissue differentiation, and response to chemical exposure. In this study, we applied a computational approach to analyze expressed sequence tags, and identified 32 miRNAs belonging to 22 miRNA families, in three earthworm species *Eisenia fetida*, *Eisenia andrei*, and *Lumbricus rubellus*. These newly identified earthworm miRNAs possess a difference of 2–4 nucleotides from their homologous counterparts in *Caenorhabditis elegans*. They also share similar features with other known animal miRNAs, for instance, the nucleotide U being dominant in both mature and pre-miRNA sequences, particularly in the first position of mature miRNA sequences at the 5′ end. The newly identified earthworm miRNAs putatively regulate mRNA genes that are involved in many important biological processes and pathways related to development, growth, locomotion, and reproduction as well as response to stresses, particularly oxidative stress. Future efforts will focus on experimental validation of their presence and target mRNA genes to further elucidate their biological functions in earthworms.



## 1. Introduction

MicroRNAs (miRNAs), an abundant new class of non-coding small RNAs that average 22 nucleotides (nt) in length, have recently been discovered playing important roles in regulating gene expression through binding by near-perfect Watson–Crick pairing to the 3′-untranslated region (UTR) of target messenger RNAs (mRNAs) (Ambros, 2004; Bartel, 2004). In association with their protein effector components, miRNAs mediate sequence-specific posttranscriptional and transcriptional gene regulation, and hence control mRNA translation, stability and localization (Parker and Sheth, 2007; Pillai et al., 2007) and feed into processes that control transposons (O'Donnell and Boeke, 2007; Vastenhouw and Plasterk, 2004) and heterochromatin structure (Grewal and Elgin, 2007; Henderson and Jacobsen, 2007). One recent comparative genomics study has shown that miRNA promoter regions are twice as conserved as mRNA promoters (Mahony et al., 2007), underscoring the importance of miRNAs in gene regulation.

It is now clear that miRNAs play a fundamental role in normal development and in disease pathology (Taylor and Gant, 2008). Greater than one third of the approximately 30,000 human genes have been predicted to be miRNA targets (Lewis et al., 2005). For instance, miRNAs have been associated with tissue development and differentiation, learning and memory, maintenance of germline stem cells, cell cycle progression, apoptosis, as well as such diverse disease processes as tumorigenesis, cardiac hypertrophy, Alzheimer's disease, and diabetes (Hudder and Novak, 2008).

As key representatives of soil fauna, earthworms form the foundation of a healthy soil ecosystem and can serve as early bioindicators of environmental stressors such as toxicants. Earthworms are essential in maintaining soil fertility through their burrowing, ingestion and excretion activities (Edwards, 2004). There are over 8000 described species worldwide, existing everywhere but in Polar and arid climates (Reynolds, 2004). They are increasingly recognized as indicators of agroecosystem health and ecotoxicological sentinel species because they are constantly exposed to contaminants in soil. Three earthworm species *Eisenia fetida*, *Eisenia andrei*, and *Lumbricus rubellus* are widely used as "model" organisms by many researchers and others working on the biology, ecology and ecotoxicology of the soil. They all belong to the *Lumbricidae* family. *E. fetida* and *E. andrei* are two sibling species commonly found in North American composters and are sold commercially for fish bait. Earthworms have a life span of 4–5 years and are obligatorily amphimictic even though

* Corresponding author. Tel.: +1 601 6343521; fax: +1 601 6343838.
*E-mail addresses:* ping.gong@usace.army.mil (P. Gong), xief@ecu.edu (F. Xie), zhangb@ecu.edu (B. Zhang), edward.j.perkins@usace.army.mil (E.J. Perkins).

1476-9271/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2010.09.004

each has both male and female reproductive organs (Reynolds, 1995). The ecological importance and roles in soil ecotoxicology have made earthworms the subject of extensive expressed sequence tag (EST) projects (Lee et al., 2005; Pirooznia et al., 2007; Stürzenbaum et al., 2003), which have generated over twenty thousand Sanger sequences. Recently, we obtained over half a million raw sequence reads from a normalized *E. fetida* nerve tissue-specific EST collection using the 454 GS20 system, a next-generation high throughput sequencing technology (Gong et al., 2010). Currently, there is also a consorted effort undertaking to sequence the whole genome of *L. rubellus* (Stürzenbaum et al., 2009; also see http://xyala.cap.ed.ac.uk/Lumbribase/index.shtml).

According to the public miRNA database miRBase (Release 14, September 2009), there are a total of 10,883 miRNAs identified from 66 animal species, 29 plant species, and 19 viruses. However, no miRNA has been reported in any earthworm species. This study was motivated and conceived to discover conserved earthworm miRNAs by taking advantage of the rich earthworm EST sequence information as well as the computational method we previously developed (Zhang et al., 2005). We also attempted to annotate the identified earthworm miRNAs by in silico prediction of their target protein-coding ESTs and bioinformatic data mining of biological functions of these mRNAs.

## 2. Materials and methods

### 2.1. Sources of earthworm EST sequences

We searched all publicly available EST sequences for two earthworm species, *L. rubellus* and *E. andrei*, and found the website www.earthworms.org to download 17,225 EST sequences for *L. rubellus* (LumbriBASE) and 1108 EST sequences for *E. andrei* (EandreiBASE). The majority of the *L. rubellus* sequences (ca. 10,000) were picked from cDNA libraries produced from earthworms of different developmental stages maintained in a controlled environment. The remainder was four targeted sets of ca. 2000 sequences generated from subtractive libraries for each of the four pollutants Cu, Cd, fluoranthene and atrazine (Stürzenbaum et al., 2003). The *E. andrei* ESTs were derived from a midgut cDNA library (Lee et al., 2005). For *E. fetida*, we used the 3144 good quality EST sequences with an average length of 310 bases that were obtained from our previous studies (Pirooznia et al., 2007). All these ESTs were sequenced using the traditional Sanger capillary array electrophoresis technology and have been deposited into the GenBank EST database. Recently, 562,327 quality filtered sequence reads with an average length of 104 bases were generated using the 454 GS20 system for a normalized, non-clonal full-length cDNA collection produced from *E. fetida* nerve tissues. The high throughput sequencing data have been submitted to NCBI's Short Read Archive with a submission number SRA009433. This dataset has been assembled into 31,114 contigs and 157,070 singletons using Newbler, which were used in this study along with the 3144 Sanger sequences for *E. fetida*.

### 2.2. Reference miRNA sequences

Genes encoding miRNAs are highly conserved through metazoan phylogeny with new families being added to bilaterian lineages through evolutionary time (Prochnik et al., 2007). Once integrated into the genomic regulatory circuitry, miRNAs are rarely secondarily lost (Wheeler et al., 2009). Therefore, we picked *Caenorhabditis elegans* as the "reference" species that has the highest number of miRNA counts among invertebrates in the miRBase (www.mirbase.org). We downloaded a total of 174 known *C. elegans* mature miRNAs and their corresponding precursor sequences (miRBase release 14.0, September 2009), which we defined as the reference set of miRNA sequences. Although some of these *C. elegans* miRNAs were initially identified by computational approaches, a majority of them have been validated by experimental approaches including direct cloning, PCR, Northern blotting, and/or deep sequencing technologies (Griffiths-Jones et al., 2008).

### 2.3. In silico identification of conserved earthworm miRNAs from ESTs

Briefly, the mature sequences of known *C. elegans* miRNAs were aligned using the BLASTN algorithm (Johnson et al., 2008) against all of the collected earthworm EST sequences (Fig. 1). To improve the search efficiency, the BLASTN parameter settings were adjusted according to our previous report (Zhang et al., 2008) as follows: (1) expect values were set at 1000 to increase the number of hits; (2) the default word-match size between the query and database sequences was set at four; and (3) the number of descriptions and alignments was raised to 1000. If the search revealed partial sequence similarity to a *C. elegans* mature miRNA sequence, the non-aligned regions were manually inspected and compared to determine the number of matching nucleotides in order to assess their potential as miRNA candidates. After removing redundant earthworm ESTs, those having no more than 4 mismatches (including insertion, mutation, and/or deletion nucleotides) against the known *C. elegans* mature miRNAs were selected to predict their secondary structures using MFOLD 3.2 (Zuker, 2003; available at www.bioinfo.rpi.edu/applications/mfold/). The default parameters for MFOLD 3.2 were used, and all MFOLD outputs including the number of each nucleotide (A, G, C and U), location of the matching regions, the number of arms per structure, and minimal folding free energy (MFE, $\Delta$G kcal/mol) were recorded. We also calculated the minimal folding free energy index (MFEI) as follows: (MFE $\times$ 100/sequence length)/(G+C)% (Zhang et al., 2006).

### 2.4. Criteria for miRNA candidates

In this study, an RNA (EST) sequence was considered a miRNA candidate only if it met all of the following requirements (Zhang et al., 2008): (1) predicted mature miRNAs had four or fewer nucleotide substitutions compared with *C. elegans* mature miRNAs; (2) the RNA sequence could fold into an appropriate stem-loop hairpin secondary structure; (3) the mature miRNA could be localized in one arm of the hairpin structure; (4) no more than 6 mismatches between the predicted mature miRNA sequence and its opposite miRNA* sequence in the secondary structure; (5) no loop or break in the miRNA or miRNA* sequences; and (6) predicted secondary structure had a high MFEI and a high negative MFE. Overall, the application of these criteria for inclusion of RNAs as miRNAs reduced the number of RNAs analyzed, minimized the likelihood that non-miRNAs would be included in subsequent analyses, and hence significantly reduced the number of false positive predications.

### 2.5. Target prediction and annotation for identified miRNAs

We chose to use miRanda (version 1.0b) for predicting miRNA targets among the available ESTs for each of the three earthworm species (Enright et al., 2003). We used the default setting for all the parameters except for the score threshold which was set at 80. To annotate the predicted target ESTs, the EST sequences were queried using BLASTX against the NCBI's non-redundant protein database with the *E*-value set at $10^{-15}$, and against the Gene Ontology (GO) protein database (Ashburner and Bergman, 2005). Pathway enrichment analysis was performed by using a combined