Supporting Information for

Are herbal sRNAs really novel precision medicines?

Shanfa Lu [1,*]

[1] Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, No.151 Malianwa North Road, Haidian District, Beijing 100193, China

* Corresponding author (email: sflu@implad.ac.cn)

Materials and methods

Blast analysis of sRNAs against high-throughput *R. crenulata* genomic DNA and transcriptome sequence data and was performed using blastn with the default algorithm parameters and the word size to be 7 (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&BLAST_SPEC=SRA&SHOW_DEFAULTS=on&DB_GROUP=Exp&view=search_blast). Blast analysis of sRNAs against NCBI Nr database was performed using blastn with the search parameters adjusted to search for a short input sequence (Altschul et al., 1997). Blast analysis of sRNAs against miRBase (http://www.mirbase.org/) was performed with the default algorithm parameters (Kozomara et al., 2019).

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller,W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res 47(D1), D155–D162.

Table S1. HJT-sRNA-m1–HJT-sRNA-m8 contain conserved animal miRNA sequences.

sRNA Name	Sequence (5′→3′) [1]	Length (nt)	Animal miRNA Name
HJT-sRNA-m1	UGUCUCGUACCGUGAGUAAUAAUGCG	26	mir-126a
HJT-sRNA-m2	GCUGAGAUGAAGCACUGUAGCUC	23	mir-143
HJT-sRNA-m3	GUUAUUCAAGUAAUCCAGGAUAGGCU	26	mir-26a
HJT-sRNA-m4	UCUGAGGUAGUAGGUUGUAUGGUUAU	26	let-7c
HJT-sRNA-m5	GUAUGUAAACAUCCUCGACUGGAAGCU	27	mir-30a
HJT-sRNA-m6	GUUAUGAGGUAGUAGAUUGUAUAGU	25	let-7f
HJT-sRNA-m7	UGAGGUAGUAGGUUGUGUGGUUGUAAGC	28	let-7b
HJT-sRNA-m8	GACGGUCGUACCGUGAGUAAUAAUGCGA	28	mir-126a

[1] Animal miRNA sequences are underlined.

Various factors causing sRNA annotation errors

Many factors could cause sRNA annotation errors. To correctly identify the source of a sRNA, special attention should be paid to the matters listed below.

Contaminations: There are various chances to introduce contaminations from microbes, human beings, animals and other plants during the isolation and identification of herbal sRNAs. It includes but not limited to experimental material preparation, RNA extraction, and sRNA library construction. To minimize contaminations, the herb materials purchased from pharmacies should be washed thoroughly with pure water, and decoction preparation, RNA extraction and sRNA library construction should be carried out using clean utensils in a clean environment.

Sequence errors: Sequence errors might be introduced by PCR and sequencing (Clarke et al., 2001; Cummings et al., 2010). It is particularly true for PCR amplification. About 10% of all sequences contained one or more PCR errors when amplifying a 250 bp sequence (Keohavong and Thilly, 1989; Kobayashi et al., 1999). The overall rate of base misincorporation was about 1.85×10^{-5} misincorporations per base per cycle (Cummings et al., 2010). To minimize sequence artefacts introduced by PCR, the cycles of PCR can be properly reduced. In addition, using high-fidelity thermostable DNA polymerase for PCR amplification may significantly reduce sequence errors.

Adapter sequence: During sRNA library construction, adapters are usually ligated to short sequences of DNA. The adapter sequences need to be removed from sRNA sequences after sequencing. However, partial adapter sequence could be left at the ends of a sRNA during computational removal. It is particularly true when the adapter sequences were not intact after sequencing or sequence errors were introduced during PCR amplification and sequencing. To solve this problem, sRNAs without intact adapter sequences must be discarded. It means that only those with intact and correct adapter sequences may be retained and used for further analysis.

miRNA isoforms: miRNA isoforms, known as isomiRs, are variants of known miRNAs (Morin et al., 2008). They can be generated through the next four mechanisms. (1) Imprecise and alternative cleavage of Dicer and Drosha in animals and Dicer-like in plants during pre-miRNA processing (Guo and Lu, 2010); (2) exosome-mediated RNA decay (Lu et al., 2009); (3) miRNA editing (Li et al., 2018); and (4) miRNA modification (Li et al., 2005; Heo et al., 2008; Katoh et al., 2009; Lu et al., 2009). IsomiRs generated through mechanism (1) can be perfectly mapped to miRNA precursors, whereas they cannot be perfectly mapped to known miRNAs. IsomiRs generated through mechanism (2) can be perfectly mapped to both miRNA precursors and known miRNAs. However, isomiRs generated through mechanism (4) can be perfectly mapped to neither miRNA precursors nor known miRNAs.

Modifications, such as uridylation and adenylation, are important regulatory mechanisms to miRNA stability in both plants and animals and are biologically meaningful (Li et al., 2005; Heo et al., 2008; Katoh et al., 2009; Lu et al., 2009). They are usually processed by adding one or few non-genomic DNA template nucleotides to the 3’ end of a miRNA after biogenesis (Guo and Lu, 2010). The portion of modified miRNAs is often significant in a tissue and exhibits subtle variability in expression during the development of an organism (Xie et al., 2015; Fernandez-Valverde et al., 2010). Sequence modification may result in that the miRNA sequence cannot perfectly match the corresponding genomes during computational analysis. In these cases, the sRNAs analyzed need to be carefully checked.

sRNA conservation: Although many sRNAs are species or lineage-specific and show low conservativity among organisms, some miRNAs are deeply conserved and widely exist in the animal or plant kingdom. It is difficult to identify the source of

these deeply conserved miRNAs by computational sequence alignment, particularly when contaminants exist in the experimental materials. Under these circumstances, taking the abundance of sRNA sequence into account is very important.

Sequence alignment: To identify the source of sRNAs, various sequence alignments should be performed. First, clean data of sRNAs can be aligned to the whole genome sequence assemblies, high-throughput genomic DNA and transcriptome sequences of the analyzed organisms. The alignment can be performed using blastn with the word size to be 7 (Altschul et al., 1997). Second, the identified sRNAs may be further blast-analyzed against NCBI Nr database using blastn with the search parameters adjusted to search for a short input sequence (Altschul et al., 1997). Third, to check whether the sRNAs are miRNAs or contain miRNA sequence, blast analysis of sRNAs against miRBase can be carried out with the default algorithm parameters (Kozomara et al., 2019). The results from these analyses can be used to cross-check the source of a sRNAs.

There are many matters needing special attention during the analysis of herbal sRNA candidates functioning in animals and human beings. Hopefully, the proposed matters needing special attention and problem-solving solutions listed above are useful to scientists who are working on cross-kingdom regulation of sRNAs.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller,W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Clarke, L.A., Rebelo, C.S., Gonçalves. J., Boavida, M.G., and Jordan, P. (2001). PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol 54, 351–353.

Cummings, S.M., McMullan, M., Joyce, D.A., and van Oosterhout, C. (2010). Solutions for PCR, cloning and sequencing errors in population genetic analysis. Conserv Genet 11, 1095–1097.

Fernandez-Valverde, S.L., Taft, R.J., and Mattick, J.S. (2010). Dynamic isomiR regulation in *Drosophila* development. RNA 16, 1881–1888.

Guo, L., and Lu, Z. (2010). Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. Comput Biol Chem 34, 165–171.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V.N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. Mol Cell 32, 276–284.

Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Kashiwabara S, Baba T, and Suzuki T. (2009). Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. Genes Dev 23, 433–438.

Keohavong, P., and Thilly, W.G. (1989). Fidelity of DNA-polymerases in DNA amplification. Proc Natl Acad Sci USA 86, 9253–9257.

Kobayashi, N., Tamura, K., and Aotsuka, T. (1999). PCR error and molecular population genetics. Biochem Genet 37, 317–321.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res 47(D1), D155–D162.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis*. Curr Biol 15, 1501–1507.

Li, L., Song, Y., Shi, X., Liu, J., Xiong, S., Chen, W., Fu, Q., Huang, Z., Gu, N., and Zhang, R. (2018). The landscape of miRNA editing in animals and its impact on miRNA biogenesis and targeting. Genome Res 28, 132–143.

Lu, S., Sun, Y.H., and Chiang, V.L. (2009). Adenylation of plant miRNAs. Nucleic Acids Res 37, 1878–1885.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18, 610–621.

Xie, F., Wang, Q., and Zhang, B. (2015). Global microRNA modification in cotton (*Gossypium hirsutum* L.). Plant Biotechnol J 13, 492–500.

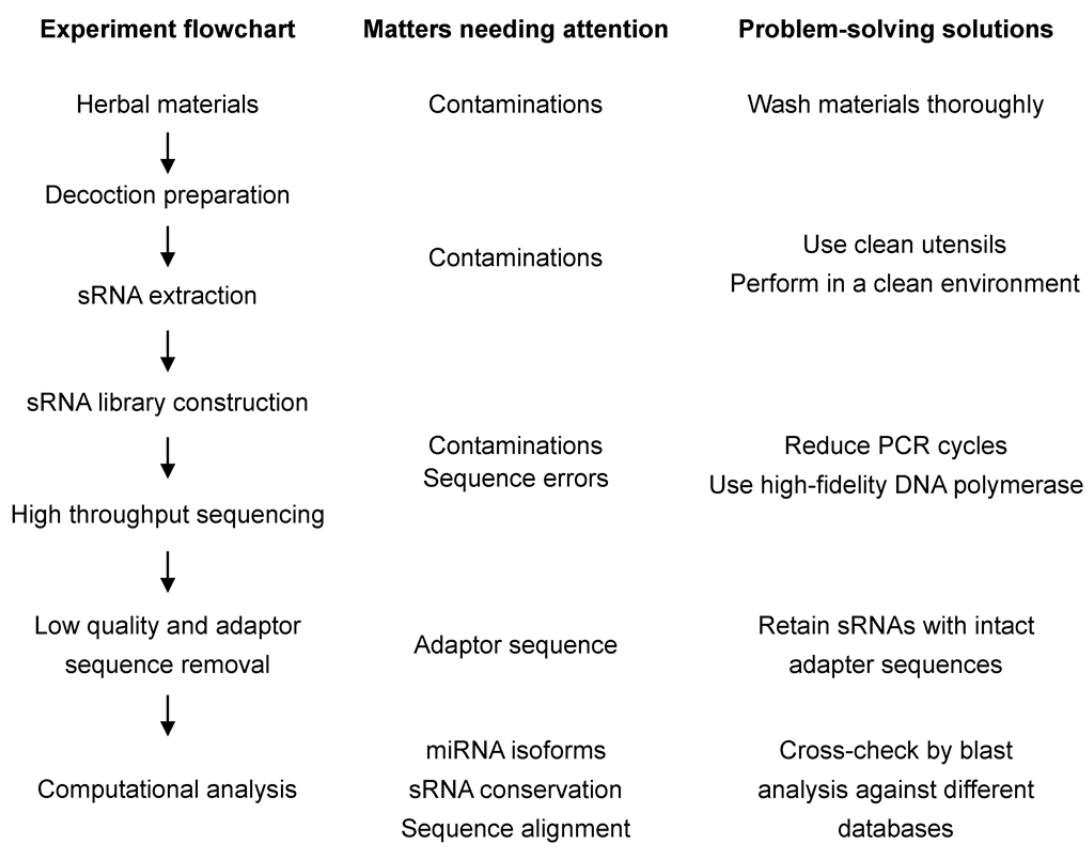

Figure S1 Experiment flowchart and summary of matters needing special attention and problem-solving solutions during the analysis of herbal sRNAs.