Mining of public sequencing databases supports a non-dietary origin for putative foreign miRNAs: underestimated effects of contamination in NGS

JUAN PABLO TOSAR,[1,2] CARLOS ROVIRA,[3] HUGO NAYA,[1] and ALFONSO CAYOTA[1,4,5]

[1]Institut Pasteur de Montevideo, Montevideo 11400, Uruguay
[2]Faculty of Science, Universidad de la República, Montevideo 11400, Uruguay
[3]Division of Oncology, Department of Clinical Sciences, Lund University Cancer Center, Lund 223 81, Sweden
[4]Faculty of Medicine, Universidad de la República, Montevideo 11800, Uruguay

ABSTRACT

The report that exogenous plant miRNAs are able to cross the mammalian gastrointestinal tract and exert gene-regulation mechanism in mammalian tissues has yielded a lot of controversy, both in the public press and the scientific literature. Despite the initial enthusiasm, reproducibility of these results was recently questioned by several authors. To analyze the causes of this unease, we searched for diet-derived miRNAs in deep-sequencing libraries performed by ourselves and others. We found variable amounts of plant miRNAs in publicly available small RNA-seq data sets of human tissues. In human spermatozoa, exogenous RNAs reached extreme, biologically meaningless levels. On the contrary, plant miRNAs were not detected in our sequencing of human sperm cells, which was performed in the absence of any known sources of plant contamination. We designed an experiment to show that cross-contamination during library preparation is a source of exogenous RNAs. These contamination-derived exogenous sequences even resisted oxidation with sodium periodate. To test the assumption that diet-derived miRNAs were actually contamination-derived, we sought in the literature for previous sequencing reports performed by the same group which reported the initial finding. We analyzed the spectra of plant miRNAs in a small RNA sequencing study performed in amphioxus by this group in 2009 and we found a very strong correlation with the plant miRNAs which they later reported in human sera. Even though contamination with exogenous sequences may be easy to detect, cross-contamination between samples from the same organism can go completely unnoticed, possibly affecting conclusions derived from NGS transcriptomics.

Keywords: exogenous; diet-derived; microRNAs; plant MIR168a; contamination

In an exciting and widely cited article published in *Cell Research*, Chen-Yu Zhang and colleagues reported the striking finding that exogenous plant miRNAs were present in mammalian sera and tissues (Zhang et al. 2012a). These exogenous miRNAs were assumed to be orally acquired through food intake. Furthermore, the authors reported that diet-derived plant MIR168a could inhibit mammalian gene expression in the liver and interpreted this as an evidence of cross-kingdom regulation of gene activity. However, reproducibility of these results was recently questioned by Dickinson et al. (2013). The authors fed mice with a rice-rich diet but they were not able to detect any significant amounts of circulating rice miRNAs by either qPCR or deep sequencing. To increase the controversy, Chen-Yu Zhang and colleagues responded in the November issue of *Nature Biotechnology (2013)* that the data by Dickinson and colleagues suffered from sequencing bias between animal and plant miRNAs, since the authors were unable to detect significant amounts of plant miRNAs in rice-containing chow or even rice grain (Chen et al. 2013). However, Dickinson and colleagues excluded the existence of any significant bias by performing ligation-independent verification of their sequencing results by qPCR, including spiked-in methylated oligonucleotides. Remarkably, failure to efficiently deliver exogenous miRNAs after performing controlled feeding experiments was previously reported by other authors (Zhang et al. 2012b; Snow et al. 2013; Witwer et al. 2013). Due to the lack of independent confirmation and consideration of the biological barriers against

[5]**Corresponding author**
E-mail cayota@pasteur.edu.uy

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.044263.114.

the uptake and stability of diet-derived nucleic acids, the initial excitement caused by the discovery of dietary miRNAs has moved to more prudent or skeptical positions (Petrick et al. 2013; Witwer and Hirschi 2014).

In order to unravel this controversy, we searched for diet-derived miRNAs in deep-sequencing libraries of small RNAs in a variety of randomly selected studies of human tissues (SRP002347 and SRP002272: GEO: GSE21279; SRP005294 and SRP007825: GEO: GSE31037; and SRP002402: GEO: GSE21191). After careful depletion of human sequences with different sequential algorithms, we found substantial amounts of exogenous sequences (Fig. 1A). In most cases, plant MIR168a from monocots was the most abundant. Among the most interesting cases we found high amounts of exogenous sequences in three samples of human spermatozoa (GSE21191). Here, we found that 602, 15,904, and 8906 reads per million (RPM) corresponded to plant MIR168a. Plant MIR156a was also identified with 0.9, 29, and 42 RPM. Strikingly, while plant MIR168a was typically <1% of human miRNAs in our meta-analysis (Fig. 1A), a slightly lower abundance than previously reported (Zhang et al. 2012a), it was up to 12.5 times more abundant than endogenous miRNAs in human spermatozoa (1256%) (Fig. 1B). Considering that sperm-borne endogenous miRNAs are known to play a role in mammalian reproduction (Liu et al. 2012; Sendler et al. 2013), these findings, if confirmed, would imply a role for diet-derived miRNAs in early development.

To validate this finding we sequenced spermatozoa samples from three fertile donors. Taking into consideration that plant miRNAs are 2′-*O*-methylated at their 3′ end and that this biochemical singularity could hamper 3′ adapter ligation, purified RNA from the three samples was treated in parallel with sodium periodate. The periodate anion oxidizes vicinal diols (as those present in the 3′ termini of mammalian miRNAs) to aldehydes, precluding oxidized RNAs from adapter ligation and sequencing. Thus, oxidized deep sequencing should favor detection of plant miRNAs. The same procedure was followed by Zhang and colleagues to probe that their sequenced plant miRNAs were actually plant-derived (Zhang et al. 2012a).

Even though the sum of processed reads was higher in our sequencing than in GSE21191, we could not detect traces of any plant miRNAs in our study. In periodate-treated RNAs, we observed, as expected, an increase in relative abundances of human piRNAs, but we did not detect any miRNA that could be interpreted as diet-related. This suggests that the exogenous miRNAs present in the public data may have been caused by contamination from the laboratory environment. Being aware of this possibility, we prepared one of the libraries from the spermatozoa samples (with and without oxidation with sodium periodate) following the same rules for good laboratory practice but in parallel with small RNA libraries from turtle (*Trachemys scripta*). In this spermatozoa library, we found 11.6 RPM unique turtle-derived sequences and 108 RPM after periodate oxidation. Some of the reads were up to 42 nt long, and perfect blast hits from the gender *Trachemys* were retrieved (Supplemental Table S1). On the contrary, in the spermatozoa libraries prepared in the absence of turtle RNA, we were unable to detect a

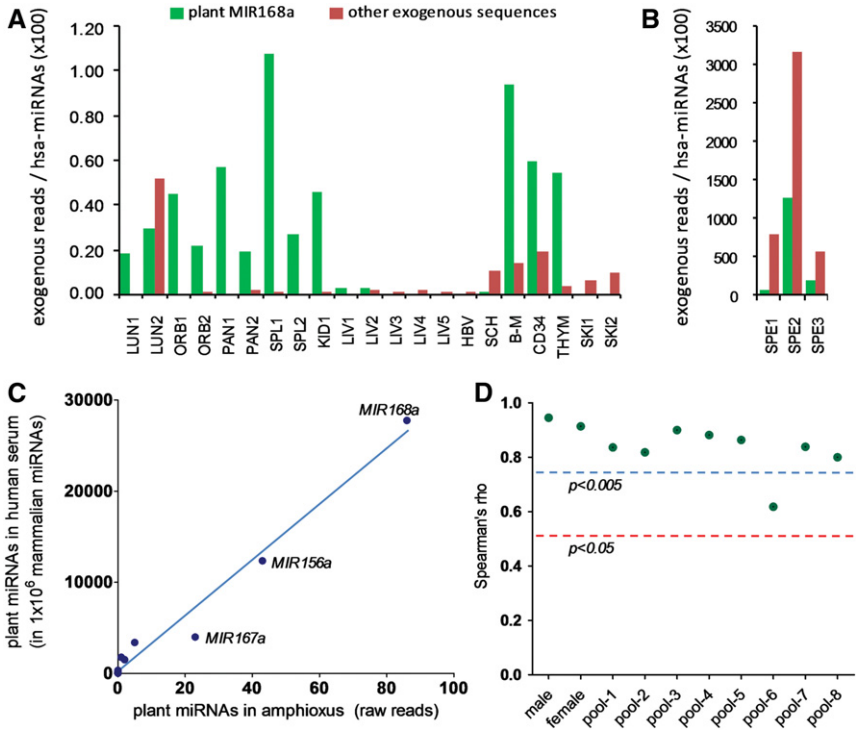

FIGURE 1. Relative abundance of putative exogenous sequences in sRNA data sets from human tissues (*A*,*B*), and correlation between plant miRNAs in human sera and in amphioxus (*C*,*D*). (*A*) Reads corresponding to plant MIR168a (green) or other exogenous sequences (red) were normalized to the total number of reads corresponding to human miRNAs in each sample and multiplied by 100. Abbreviations: LUN, lung; ORB, orbital gyrus; PAN, pancreas; SPL, spleen; KID, kidney; LIV, liver; HBV, HBV-infected liver; SCH, liver with severe chronic hepatitis B; B-M, bone marrow; CD34, CD34+ progenitor blood cells; THYM, thymocytes; SKI, skin; SPE, spermatozoa. (*B*) For reasons of scale, spermatozoa samples were presented separately. (*C*) Correlation between the levels of plant miRNAs (normalized to 1 million mammalian miRNAs) found in human serum ("male" library) according to Zhang et al. (2012a) and their corresponding levels (raw reads) found in previous data from the same group (GSE16859; small RNA sequencing in amphioxus). Pearson $r = 0.9874$; $P < 0.0001$; Spearman rho = 0.9439; $P < 0.0001$ ($n = 11$; two-tailed *P*-values). (*D*) Spearman's rank correlation coefficient (rho) values were calculated for the correlation between the 10 human samples sequenced by Zhang and colleagues and their corresponding levels in amphioxus. The red and blue lines show the critical values ($n = 11$) for $P < 0.05$ and $P < 0.005$, respectively.

single turtle-derived read. This strongly suggests that precautions should be taken to prevent cross-contamination of the samples due to the high sensitivity of next-generation sequencing methods.

Being done on a complete different sample set, our test, of course, does not prove that the detection of exogenous miRNAs in mammalian samples by Zhang and colleagues or in our analysis of meta-data was the consequence of contamination. However, if cross-contamination from the laboratory environment is the reason for the unexplained lack of reproducibility of Zhang and colleagues' data, we could expect to also find traces of plant contamination in other data sets produced by the same laboratory at approximately the same time as the *Cell Research* data were created. To test this assumption, we looked in the literature for previous sequencing reports performed by the same group, focusing on studies where the presence of exogenous miRNAs could not be explained by the diet. We found an article published in 2009 where the expressed miRNA complement in lancelets (*Branchiostoma lanceolatum* or amphioxus) was characterized using Illumina sequencing (SRP001014: GSE16859) (Chen et al. 2009). We analyzed these data and found 86 out of 1,387,630 raw reads with a 100% identity to plant MIR168a. This miRNA is specific for plants from the family Poaceae. Blast alignments to the *Branchiostoma* genome did not retrieve any high scoring hits that would indicate that the sequence was amphioxus-derived. Lancelets do not eat grasses but were fed with sea algae according to the methods section from the same report. Thus, delivery of plant miRNAs through the diet of filter feeder amphioxus is definitely not expected. We did not find significant similarity between MIR168a and any algae miRNAs.

Together with MIR168a, another abundant plant miRNA detected in human sera by Zhang and colleagues was MIR156a. Coincidently, we also found 43 reads corresponding to plant MIR156a in the amphioxus data. On top of this, plant MIR167a, MIR166a, MIR172a, and MIR169b were also detected with 23, 5, 2, and 1 reads, respectively. This is remarkable since the relative abundances of all the detected plant miRNAs in amphioxus resembled the relative abundances in the libraries from human sera analyzed in Zhang and colleagues. The correlation of the data is significant, with a $P < 0.0001$ of obtaining equal or higher correlation coefficients by chance (either Pearson's r or Spearman's rho), when comparing the relative levels of plant miRNAs in human serum (male library) with their corresponding reads in amphioxus (Fig. 1C). In addition, a significant ($P < 0.05$) correlation was found with any of the samples and pools of human serum sequenced by the authors (Fig. 1D). Relative levels of plant miRNAs in human data sets were obtained from the supplementary materials sections of the paper by Zhang et al. (2012a). Absolute read count numbers were not used because, regrettably, raw sequencing data were not submitted to public repositories.

In summary, we think that our observations render the conclusions achieved by Zhang and colleagues questionable. Contamination was already postulated as a possible explanation for the eventual detection of plant miRNAs in human samples (Zhang et al. 2012b; Dickinson et al. 2013; Witwer et al. 2013). Herein, we provide experimental evidences supporting the idea that contamination is, indeed, the underlying cause of these findings. Even if not, future confirmatory studies will need to be able to demonstrate lack of contamination in the sequencing data.

Next-generation sequencing technologies have become a widespread and indispensable tool in many research fields, and clinical use of a deep-sequencing platform has recently been approved by FDA (Collins and Hamburg 2013). These technologies have reduced the costs and increased the speed of DNA sequencing by four orders of magnitude (Shendure and Lieberman 2012), and have enough sensitivity to produce whole-transcriptome data from single cells (Tang et al. 2010). However, the other side of the coin for such a high sensitivity is that even low amounts of contaminant nucleic acids can be detected in biological samples. Leaving aside the discussion whether or not exogenous RNAs are artifacts, our main concern is what seems to be a widespread underestimation of the effects of contamination on deep-sequencing data. The case discussed here is simply a paradigmatic example but plant-derived miRNAs have been identified even in transcriptome analysis of cultured cells (Zhang et al. 2012b). Furthermore, we have analyzed sequencing data generated from all over the world, and detection of contaminant sequences was found to be ubiquitous (Fig. 1A) and in some cases reached extreme, biologically meaningless levels (Fig. 1B). Contamination with nucleic acids from unrelated organisms may be easy to spot and a bioinformatic method was recently designed for de novo detection and removal of contaminating reads (Zhou et al. 2013). However, cross-contamination between samples from the same organisms can go completely unnoticed, which may foster concern about the validity of clinical data. To solve these possible technical pitfalls the scientific community should agree on adequate standard controls that assure the proper interpretation and reproducibility of the data.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was partially funded by ANII (Agencia Nacional de Investigación e Innovación, Uruguay) and FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11. We thank Dr. José María Montes and Dr. Mariel Cánepa (Fertilab, Uruguay) for the preparation of the sperm samples, and Gonzalo Greif and Natalia Rego (Institut Pasteur de Montevideo) for their assistance with the small RNA sequencing.

Received January 11, 2014; accepted February 25, 2014.

REFERENCES

Chen X, Li Q, Wang J, Guo X, Jiang X, Ren Z, Weng C, Sun G, Wang X, Liu Y, et al. 2009. Identification and characterization of novel amphioxus microRNAs by Solexa sequencing. *Genome Biol* **10:** R78.

Chen X, Zen K, Zhang CY. 2013. Reply to Lack of detectable oral bioavailability of plant microRNAs after feeding in mice. *Nat Biotechnol* **31:** 967–969.

Collins FS, Hamburg MA. 2013. First FDA authorization for next-generation sequencer. *N Engl J Med* **369:** 2369–2371.

Dickinson B, Zhang Y, Petrick JS, Heck G, Ivashuta S, Marshall WS. 2013. Lack of detectable oral bioavailability of plant microRNAs after feeding in mice. *Nat Biotechnol* **31:** 965–967.

Liu WM, Pang RT, Chiu PC, Wong BP, Lao K, Lee KF, Yeung WS. 2012. Sperm-borne microRNA-34c is required for the first cleavage division in mouse. *Proc Natl Acad Sci* **109:** 490–494.

Petrick JS, Brower-Toland B, Jackson AL, Kier LD. 2013. Safety assessment of food and feed from biotechnology-derived crops employing RNA-mediated gene regulation to achieve desired traits: a scientific review. *Regul Toxicol Pharmacol* **66:** 167–176.

Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. 2013. Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res* **41:** 4104–4117.

Shendure J, Lieberman Aiden E. 2012. The expanding scope of DNA sequencing. *Nat Biotechnol* **30:** 1084–1094.

Snow JW, Hale AE, Isaacs SK, Baggish AL, Chan SY. 2013. Ineffective delivery of diet-derived microRNAs to recipient animal organisms. *RNA Biol* **10:** 1107–1116.

Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, Lao K, Surani MA. 2010. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* **5:** 516–535.

Witwer KW, Hirschi KD. 2014. Transfer and functional consequences of dietary microRNAs in vertebrates: concepts in search of corroboration. *Bioessays* **36:** 394–406.

Witwer KW, McAlexander MA, Queen SE, Adams RJ. 2013. Real-time quantitative PCR and droplet digital PCR for plant miRNAs in mammalian blood provide little evidence for general uptake of dietary miRNAs: limited evidence for general uptake of dietary plant xenomiRs. *RNA Biol* **10:** 1080–1086.

Zhang L, Hou D, Chen X, Li D, Zhu L, Zhang Y, Li J, Bian Z, Liang X, Cai X, et al. 2012a. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res* **22:** 107–126.

Zhang Y, Wiggins BE, Lawrence C, Petrick J, Ivashuta S, Heck G. 2012b. Analysis of plant-derived miRNAs in animal small RNA datasets. *BMC Genomics* **13:** 381.

Zhou Q, Su X, Wang A, Xu J, Ning K. 2013. QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One* **8:** e60234.