Identifying expression of new small RNAs by microarrays

James Q. Yin a,1, Robert C. Zhao b,*

a Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing 100101, China
b Institute of Basic Medical Sciences & School of Basic Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 5 Dong Dan San Tiao, Chaoyang District, Beijing 100005, China

Accepted 13 April 2007

Abstract

Although a large number of small RNAs (sRNAs) have been discovered, it is very likely that the screens conducted so far have not reached saturation. Recently, many methods for predicting and identifying new sRNAs have been developed. However, it remains unclear what the total number of sRNAs within a genome is and how many types of sRNAs exist in plants and animals. In this article, combined methods of dynamic programming prediction, enrichment of sRNAs, and microarray analysis are developed to screen and evaluate a new class of sRNAs from introns of human, protein-encoding genes. The methods used by our laboratories to design capture probes and label enriched small RNAs are thoroughly described here. The microarray results show that our modified technologies are useful to enhance sensitivity and specificity of arrays, identify expression patterns within different cells, and discover differential expression of sRNAs during the differentiation process of bone marrow stem cells. Accordingly, the combination of computational prediction and microarray analysis may be a feasible and practical approach for profiling studies of both known and predicted small RNAs.

Keywords: miRNA; Small RNA; Microarray; Gene expression; Disease; Cancer

1. Introduction

Small RNAs (sRNAs) are a growing class of newly identified endogenous nonprotein-coding RNAs (ncRNAs) with lengths of 21–35 nt. They can be divided into at least six types including microRNAs (miRNAs), tiny noncoding RNAs (tncRNAs), short interfering RNAs (siRNAs), repeat-associated small interfering RNAs (rasiRNAs), small modulatory RNAs (smRNAs) and Piwi-interacting RNAs (piRNA) [1,2]. Of them, microRNAs (miRNAs) have been widely studied [3]. They are derived from introns or exons of protein-coding genes and noncoding genes. Hundreds of miRNAs have been cloned and sequenced [4,5]. Bioinformatics and genetic analyses suggest that these unique miRNA genes are present in the genomes of mouse, human [3,6], Drosophila, Caenorhabditis elegans, Arabidopsis, virus and fungus and are expressed in a regulated manner. A number of investigations have demonstrated that small RNAs serve as guide molecules, by base pairing with partially or fully complementary sequences in target mRNAs leading to translational repression and/or mRNA cleavage [6,7]. Moreover, small RNAs have been shown to induce transcriptional silencing, resulting in the modification of DNA and/or chromatin in yeast, plants and animals [8–13]. Each miRNA has the potential to regulate hundreds of mRNAs. With the understanding of miRNA mechanisms of action, miRNAs have been shown to affect stem cell differentiation, organ development, cell death, phase change of the cell cycle, signaling, diseases including cancer, and respond to biotic and abiotic environmental stresses [3,14–16]. Furthermore, many miRNAs are highly conserved in animals and plants and have unique tissue-specific, development-specific, and disease-specific expression patterns. This implies that each tissue is characterized by a specific set of miRNAs that may form a defining feature of that tissue [3].
Historically, most small RNAs were identified by experiments. The founding members of the miRNA family, \textit{lin-4} and \textit{let-7}, were first discovered in 1993 and 2000, respectively, as key post-transcriptional modulators for developmental transitions in early larval development of \textit{C. elegans} [17,18]. To date, many small RNAs with the potential to arise from fold-back structures characteristic of the \textit{lin-4} and \textit{let-7} hairpins were identified by cloning and sequencing [19,20]. Furthermore, by using pyrosequencing technology, Ruby et al. [21] identified thousands of endogenous small RNAs in \textit{C. elegans}, including miRNAs, siRNAs, and 21U-RNAs. Many of these sRNAs appear completely complementary to target regions of mRNAs. They might be processed from long double-stranded RNA (dsRNA) and might direct the silencing of complementary mRNAs [12]. Although cloning and sequencing are the most useful approaches for identification of new sRNAs, sRNAs of low abundance may very often be missed. Additionally, it has been suggested there may be partiality in the sequence composition of cloned miRNAs [22]. Finally, these procedures require a large amount of total RNA which is not practical in cases where sample tissue is limiting. To circumvent these limitations, many other methods for predicting sRNA genes have been developed [23–26]. They include computational screens, microarray detection, comparative genomic analysis, and prediction from the existence of consensus sRNA gene structures with promoter and terminator sequences. Similarly, using a new RNAMotif algorithm, combined with a thermodynamic scoring system to predict noncoding RNAs, Lesnik et al. [27] predicted the existence of about 6635 terminator structures in the \textit{Escherichia coli} genome. Because microarrays with a high-density of oligonucleotides have been shown to be useful and powerful for detecting mRNA expression, they may be used successfully to validate expression of putative sRNAs. For this purpose, some researchers have developed combination methodology to ascertain new miRNAs [26,28]. For example, Bentwich et al. [26] discovered new miRNAs by coupling computational prediction and microarray analysis. Similarly, several independent groups also reported that microarrays are an effective, sensitive, and specific means for high-throughput detection of miRNA expression [28–37].

In the present study, we have developed a strategy that integrates computational prediction, sRNA enrichment, and microarray analysis to identify candidate sRNAs derived from introns of human, protein-coding genes with emphasis on use of sRNA microarrays. Furthermore, we discuss the design of capture probes, sRNA enrichment, and sRNA labeling to enhance specificity and sensitivity of microarrays for detection of low abundance sRNAs. Finally, we describe results obtained by these modified approaches. Thus, we conclude that more additional sRNAs will be discovered and identified in the future with the advent of new technologies and the improvement of old methods.

2. Description of methods

In this article, we pay central attention to a custom sRNA array by which global and high-throughput analysis of predicted sRNA expression can be performed. The roadmap of this study includes two major pathways (Fig. 1). The first pathway is connected to the design of probe sequences and the second one is related to the detection of mature sRNAs. The flowchart simplifies the complex steps including small RNA prediction, sample enrichment and purification, sRNA labeling and array hybridization, and sRNA expression profile analysis. In order to focus on the main issue of sRNA microarrays, we shall describe the computational screen of sRNAs in another article to be published elsewhere.

2.1. Design of probe oligos

We developed a glass-slide microarray with probe spots corresponding to 536 new sRNAs, predicted by a modified dynamic programming algorithm, from 21,000 introns of human, protein-encoding genes. Two oligonucleotides covering sense and antisense strands were designed for each sRNA precursor. We reasoned that this design would allow us to distinguish between sRNAs originating from sense or antisense transcripts. These probes were printed in triplicate throughout the slide for a total of 4080 spots per microarray, including 864 spots for different control probes. Control probes included exogenous and endogenous positive controls and negative controls. We used eight miRNA-like sequences as ‘spike ins’. These sequences have no complementary targets in the human genome. Synthetic DNA oligonucleotides were introduced in many steps from small RNA enrichment to hybridization. Moreover, we employed tRNA, \textit{let-7}, and U6 and U2 snRNAs as endogenous positive controls, and also designed different negative controls. These control probes were used to assist
normalization and to provide absolute reference points for quality control and quantitative comparison of different microarrays. DNA probe sequences were on average 25 nt and were coupled with a nonsense sequence of 15 nt. Thus, each whole oligonucleotide had a maximum size of 40 nt. Each oligonucleotide had a 5′ amino-linker as well. This modification allows selective binding of the amino-containing DNA to silylated slides through a Schiff’s base reaction with aldehyde groups on the chip surface. Another advantage of this modification is that once bound to the chip surface, the covalent amino-modification is stable to a wide range of temperatures and solvents. Moreover, 5′-end attachment of DNA to the chip via the amino group permits steric accessibility of the bound molecules during the hybridization reaction. All probe sequences with 5′ amino-modification were synthesized by MWG-BIOTECH AG (Germany).

For fabrication of sRNA arrays, amino-modified DNA samples were triplicate printed on Corning GAPS-2 coated slides using a SmartArray spotter in a 15·17 spot configuration of each sub-array. The spot diameter was 150 μm, and distance from center to center was 240 μm. Following printing, slides were left at room temperature for 24 h to permit thorough drying of DNA onto the surface of silylated slides. Then, slides were processed to remove unbound DNA and Spotting buffer as follows: transfer slides to a beaker containing a stir bar and wash them twice in 0.2% SDS at 25 °C for 5 min each, twice in dH₂O at 50 °C for 5 min each, once in sodium borohydride solution (1.3 g Na₂BH₄ dissolved in 375 ml phosphate–buffered saline, then add 125 ml pure ethanol) at 25 °C for 5 min, three times in 0.2% SDS for 1 min each, twice in dH₂O at 25 °C for 1 min each. Air dry the slides to completion. Slides are ready for hybridization.

2.2. Purification and labeling of RNA

To detect sRNAs predicted by our modified dynamic programming algorithm, we analyzed sRNAs isolated from three different cell types including bone marrow stem cells, HLF (human hepatoma cell line), and A549 (human alveolar basal epithelial cell line). According to protocols provided by the manufacturer, about 2 × 10⁷ cells were washed with PBS and total RNA was extracted using Trizol Reagent (Life-Technologies); genomic DNA potentially present in the RNA fraction was removed using RNase-free DNase I (Ambion). RNA yield was determined on a spectrophotometer such as a NanoDropND-1000 by measuring optical density at 260 nm wavelength. The A260/A280 ratio should be 1.8–2.0 for high quality RNA preparations. Integrity of the RNA was assessed by gel visualization of intact 28S and 18S ribosomal RNA (rRNA) bands which migrate at approximately 5 and 1.9 kb, respectively, with respect to a single-stranded RNA ladder (Fig. 2). Appearance of sharp 28S and 18S rRNA indicates high integrity of isolated RNA; diffuse and smeared bands are an indication of serious degradation. In order to avoid interference due to mRNA degradation, we firstly separated nucleus from cytoplasm of cells using commercial isolation kits (e.g., CelLytic NuCLEAR™ extraction kit from Sigma). Total RNA can be extracted from nucleus or cytoplasm separately. Owing to sRNA features of low molecular weight and low abundance, standard protocols for isolation of total RNA and mRNA are not optimal for recovery of sRNA molecules and may lead to major loss of miRNAs and other sRNAs. Thus, sRNA enrichment may be an important step for measurement of their expression in different tissues and developmental stages as well as physiological and pathological states. Using the mirVana miRNA Isolation Kit (Cat# 1560, Ambion, Austin, TX), we could enrich low molecular weight (LMW) RNA that is 200 nt and smaller. RNA species smaller than 200 nts contained in the flow-through were ethanol-precipitated and resuspended in RNase-free water for further processing. In order to enrich sRNA species smaller than 40 nt, RNA less than 200 nt can be fractionated using the flashPAGE Fractionator, a specialized electrophoresis instrument, according to the manufacturer’s protocols.

![Fig. 2. Electrophoresis analysis of RNAs isolated by the mirVana miRNA Isolation Kit. High molecular weight RNAs greater than 200 nt (HMW) and low molecular weight RNAs less than 200 nt (LMW) were fractionated from total RNA (TOT) using the mirVana miRNA Isolation Kit. One-microgram of RNA from each sample was resolved on a 1.2% denaturing agarose gel (a) or 15% denaturing polyacrylamide gel (b).](image)
formamide, and 1 M HEPES/100 mM MgCl₂, 100 mg/ml PEG, 10 mg/ml BSA, 10% DMSO, and 1000 ng 5'-phosphate-cytidyl-uridyl-Cy3 donor to quantitatively link the Cy3 donor to the 3'-hydroxyl of sRNA acceptor molecules [33,38]. All target sRNAs could be specifically labeled because this method is very specific for labeling of 3'-OH termini on sRNAs. RNA labeling reactions contained 2 μg of LMW RNA (<200 nt) or 200 ng sRNA (<40 nt), 0.5 mM ATP, 50 mM HEPES (pH 7.8), 5 mM DTT, 20 mM MgCl₂, 150 mg/ml PEG, 10 mg/ml BSA, 10% DMSO, and 1000 ng 5'-phosphate-cytidyl-uridyl-Cy3-3' (Dharmacon) with 3 Weiss units of T4 RNA ligase (MBI Fermentas) in 15 μl. The labeling reaction was incubated at 25 °C for 2 h. Although this method requires labeled donor ribodinucleotide and T4 RNA ligase that have a poor reputation for both reliability and differential ligation efficiencies toward the acceptor nucleotide on RNA, this method has been successfully used to directly label miRNAs on a microarray platform for analysis of mRNA gene expression. In this study, we present optimized conditions for RNA ligation and obtained reproducible results.

2.3. Array hybridization and data analysis

Labeled RNA was precipitated with 0.3M sodium acetate and 3 volumes ethanol, resuspended in 25 μl hybridization solution containing 3× SSC, 1% BSA, 0.2% SDS, 15% formamide, and 1× Denhardt’s solution, and denatured by heating for 3 min at 95 °C, and snap-cooled on ice prior to hybridization. After placing array slides into a hybridization chamber (e.g., Corning Product #2551), 30 μl of hybridization mixture was injected into the chamber of a mSeries LifterSlips™ (Thermo Fisher Scientific, part number 22X251-M-5226), which has an innovative, raised-edge design providing separation and allowing even dispersal of hybridization solution between the DNA chip and coverslip. Subsequently, 1× hybridization buffer was added into the designated places in the chamber to maintain humidity during hybridization. Then, the sealed hybridization chamber was placed into a 45 °C water bath and swayed with a 3-D swirling hybridization instrument (CapitalBio Corporation, Beijing) in order to facilitate flow of solution between the DNA chip and coverslip. The total time for hybridization is about 16 h. When hybridization was finished, slides were washed once in 2× SSC, 0.2% SDS and three times in 0.2× SSC at 42 °C. Slides were read using a LuxScan 10K-A scanner with scanning parameters set by visual detection of coincidental U6 signal across the chip. After obtaining images of chips as a TIF file, raw pixel intensities were quantified using GenePix Pro 4.0 software (Axon Instruments Inc.). Flagged spots corresponding to absent or low-quality signals were removed from the analysis before global median normalization by both visual- and software-guided flags. Image intensities were measured as a function of the median of foreground minus background. Negative values were normalized to 10. The median intensities of foreground and of the corresponding background were measured. The median intensity of the background was subtracted from the median intensity of the foreground. The resulting signal intensity values were normalized to per-chip median values. These signal intensity values were then used to obtain geometric means and standard errors for each sRNA. Each sRNA signal was transformed to log₂ and 1-sample t-test was conducted. Finally, Significance Analysis of Microarrays (SAM, version 2.1) was used for identifying sRNA genes on a microarray with statistically significant changes in expression and for analyzing differential expression of sRNAs [39,40]. Criteria for selection were set as the false discovery rate (FDR) being less than 5% and fold change being more than 2.

3. Results

To evaluate whether the sRNA array could detect expression of putative sRNAs in human stem cells, normal cells, and tumor cells, we cultured bone marrow stem cells (BMSC), hematopoietic stem cells (HSC), HLF (human hepatoma cell line), and A549 (human alveolar basal epithelial cell line). To minimize the likelihood of cross-hybridization, we used only samples consisting of RNAs either less than 200 nt or 40 nt in our microchip hybridizations. The specificity and sensitivity of this type of array can be seen in the work by Luo et al. [41]. Compared with the let-7a sequence, let-7c and let-7f contain one mismatch. Let-7a signals were clearly distinguishable from let-7c and let-7f with more than 2.1-fold differences [41]. Generally, our microarray data correlate very well with the results reported by Castoldi et al. [42] and Sun et al. [43].

As noted earlier, due to the low molecular weight and low abundance of sRNAs, standard protocols for isolation of total RNA and mRNA are not optimal for recovery of sRNA molecules and may lead to major loss of miRNAs and other sRNAs [28,44]. Thus, we employed a combination of methods that integrates the mirVana miRNA Isolation Kit and the flashPAGE Fractionator to purify sRNA less than 40 nt. Hybridization results illustrate the differences between the mirVana miRNA Isolation Kit alone and the combination method (Fig. 3a and b). Using the former technology, only 66 sRNAs from A549 cells were detected with significant signal intensity. On the other hand, the combination method greatly enhanced hybridization signals of sRNAs from several fold to more than 10-fold (Fig. 3c) in which about 203 sRNAs emitted strong signal. By contrast, a constant amount of exogenous sequence did not show any obvious change in signal intensity as noted earlier.
intensity. Owing to a remarkable decrease in the signal of endogenous U2 snRNA, the combination method was very effective in removing RNAs greater than 40 nt in length. Moreover, the combination method also increased the number of detectable sRNAs whose signals were too low using standard protocols while minimizing nonspecific hybridization to control spots. These results suggest that the enrichment of sRNAs by the combination method significantly improves sensitivity of sRNA microchips compared with the amplification of miRNAs, which does not result in increased sensitivity for low abundance miRNAs [45].

Microarray experiments were used to measure temporal changes in global gene expression profile between BMSC and HSC. Fig. 4a shows array images for hybridization of LMW RNA isolated from human bone marrow stem cells before and after differentiation. Careful observation indicated that the sRNA array technology could distinguish differential expression of sRNAs from BMSC and HSC. Quantitative analyses of data from arrays confirmed that 19 miRNAs (Fig. 4b) were sharply down-regulated during BMSC differentiation, and five sRNAs were significantly up-regulated after differentiation to HSCs. Because we observed mainly down-regulation of most miRNAs during hematopoiesis, we hypothesize that most sRNAs may suppress expression of target genes involved in differentiation. The temporal variation of expression of these sRNAs suggests a developmental stage-specific function.

Tissue-specific expression patterns for some miRNAs have been revealed using different methods including microarray approaches. To examine expression patterns of putative sRNAs, we profiled sRNA from three human cell types. Direct comparison between all probes producing high signals detected by microarray for the different cell types were concisely summarized in Fig. 5. Array sRNA expression profiling identified differentially expressed sRNAs in these three cases. Two-chip replicates and quadruple spots for each miRNA provided adequate signal intensity information for data comparison across cells. Fig. 5a shows that the total signal intensity of all probes on BMSC and HLF arrays were similar compared to the notable reduction in signal intensities on the A549 array, suggesting that this class of sRNAs is reduced in A549 cells. Comparison of expression patterns of these sRNAs indicated that 13 sRNAs were shared by all cells while different cells can also express sRNAs uniquely (Fig. 5b).
Signal intensity corresponding to each of these sRNAs is depicted in Fig. 5c. These results indicate that the type and level of these sRNAs are different in diverse cells, suggesting that they may be biomarkers for discriminating different cells and tissues.

4. Common problems and troubleshooting

There are four major technological issues to be addressed for sRNA arrays. They are determination of RNA sample quality for obtaining the “best” sRNAs, estimation of hybridization quality, maximization of signal-to-noise ratio, and correct analysis of miRNA microarray data.

4.1. Determination of RNA sample quality

For high integrity of RNA samples, most laboratories use Trizol reagent for isolation which is recommended. However, each sample may result in different quality, and further evaluation of those samples is necessary. Both the Ambion and Invitrogen isolation kits allow for isolation of sRNAs (<200 nt) from tissue and from cell lines. Ambion and Invitrogen kits can even be adapted to isolate miRNA from total RNA samples. According to our experiences, enrichment of sRNA can greatly enhance the sensitivity of microarrays to detect those sRNAs with low abundance. Quality of enriched sRNA can be checked by agarose gel or the Agilent 2100 Bioanalyser to ensure that no carry-over of larger RNA molecules has occurred. For quality control of total RNA, the 28S ribosomal RNA band should be brighter than the 18S band on an agarose gel. Tailing of these bands down the gel, or a background smear behind these bands, indicates degradation of the RNA. Degraded RNA will produce high backgrounds and low signal intensities on arrays. The presence of very sharp bands larger than the 28S ribosomal RNA band indicates excess DNA in the sample, which can be removed by treatment with RNase-free DNase I. The spectrophotometric A$_{260}$/A$_{280}$ ratio can also be used for evaluating the purity of RNA. This ratio should be as close to 2.0 as possible. Generally, ratios less than 1.7 indicate that the RNA may be contaminated with other material and should be re-purified.

4.2. Estimation of hybridization quality

To ensure highly sensitive detection and specific hybridization, we have used optimized hybridization conditions and washing protocols: (1) To avoid uneven distribution
of the hybridization solution, it is of utmost importance to employ a swirling instrument for facilitating movement of hybridization solution in the chamber; (2) To increase the humidity, we recommend using a water bath or using filter paper saturated with 2× hybridization buffer; (3) To enhance specific hybridization, a higher hybridization temperature and stringent washing conditions can be used; (4) For images, the quantification method used should be the same for all arrays comprising a project, whereas image acquisition parameters, such as laser power and/or photomultiplier, can be optimized from slide to slide. Furthermore, selection of image quantification parameters (e.g. spot size, spot distance, and signal intensity) should be carefully assessed and decided upon for each experiment as a whole as this depends on the array design, slide type, and spot morphology.

4.3. Ratios of true positive signals and noise signals

The third major problem is to maximize the number of true positives at any threshold, while minimizing the number of false positives. As for true positive signals, we suggest to first check whether signals from the spike-in controls used in the labeling can be seen. If not, the labeling procedure probably has failed. The problems may be due to the quality of sRNA and labeling material. If the spike-in controls can be seen, then it is important to check that the total RNA sample is of good quality by gel electrophoresis and optical density analysis, as noted above. If RNA quality is good, then the problem may be owing to low amounts of sRNA used in the labeling. On the other hand, it is also critical to avoid photobleaching of Cy3 or Cy5. For this issue, solutions and arrays containing Cy3 or Cy5 should be kept away from light or a small quantity of dithiothreitol (DTT) might be added to the first two wash buffers used after the hybridization. In addition, careful attention should be paid to background signal. In order to enhance the ratio of true positive and noise signal, the quality and quantity of sample, as well as the hybridization conditions, should be carefully considered.

4.4. Analysis of sRNA microarray data

A fourth major problem involves research into preprocessing methods for gene expression and sRNA arrays. In order to obtain reliable results, microarray data must be background-corrected and normalized previous to testing for significant differential expression and clustering/classifying genes based on differential expression. Normalization is performed to remove dye labeling bias and differences in hybridization and scanning. It may be possible to use signals from a set of spiked-in synthetic sRNAs as common references. Once the normalization has taken place, the log2 ratios between sample and reference for each sRNA can be calculated, allowing the direct comparison of all log2 ratios from all arrays. In fact, all sRNA signals are expressed as a ratio to a reference, which should be the same.
5. Concluding remarks

The microarray described here is a modified, selective, and sensitive approach to monitor sRNA expression in human stem cells and cancer cells. By performing comparative signal analysis, our results provide strong evidence that enrichment of sRNAs less than 40 nt may be a useful method to increase sensitivity and specificity of sRNA microarrays even though some previous reports about miRNA arrays indicated that enrichment might be unnecessary [31,33]. It may be reasonable to say that enrichment may enhance the amount of sRNAs with low abundances, reduce the amount of large RNAs with similar sequences to sRNAs, and finally, improve the signal-to-noise ratio over background. Based on our modified microarray technology, we have investigated the expression profiles of new sRNAs from introns of protein-encoding genes. These sRNAs should be of importance because they can be differentially detected in different cell types and different developmental stages under different conditions. They can display developmental stage-specific and cell-specific patterns as well. Selective expression of sRNAs in different cells and varying conditions may play important roles in controlling their target genes involved in diverse physiological and pathological activities. As for the physiological function of individual sRNA, further studies are required even though several sRNAs among them have been identified by Northern blots (data not shown). Nevertheless, it is anticipated that future use of this array-based approach will lead to greater understanding of this new class of gene-regulating RNAs and other sRNAs including miRNAs.

Acknowledgments

This work was supported by CAS “Hundred Talents Program” and the 863 program (Grant No: 2006AA02Z131).

References