A bias-reducing strategy in profiling small RNAs using Solexa

GUIHUA SUN,1, XIWEI WU,2 JINHUI WANG,3 HAIQING LI,2 XUEJUN LI,2 HANLIN GAO,3 JOHN ROSSI,4 and YUN YEN1,5
1Department of Molecular Pharmacology, Beckman Research Institute of the City of Hope, Duarte, California 91010-3000, USA
2Bioinformatics Core, Beckman Research Institute of the City of Hope, Duarte, California 91010-3000, USA
3Solexa Sequencing Core, Beckman Research Institute of the City of Hope, Duarte, California 91010-3000, USA
4Department of Molecular and Cellular Biology, Beckman Research Institute of the City of Hope, Duarte, California 91010-3000, USA

ABSTRACT
Small RNAs (smRNAs) encompass several different classes of short noncoding RNAs. Progress in smRNA research and applications has coincided with the advance of techniques to detect them. Next-generation sequencing technologies are becoming the preferred smRNA profiling method because of their high-throughput capacity and digitized results. In our small RNA profiling study using Solexa, we observed serious biases introduced by the 5’ adaptors in small RNA species coverage and abundance; therefore, the results cannot reveal the accurate composition of the small RNAome. We found that the profiling results can be significantly optimized by using an index pool of 64 customized 5’ adaptors. This pool of 64 adaptors can be further reduced to four smaller index pools, each containing 16 adaptors, to minimize profiling bias and facilitate multiplexing. It is plausible that this type of bias exists in other deep-sequencing technologies, and adaptor pooling could be an easy workaround solution to reveal the “true” small RNAome.

Keywords: microRNA; deep sequencing; Solexa; small RNA

INTRODUCTION
In 2001, eight years after the initial discovery of the first miRNA Lin-4 in Caenorhabditis elegans (Lee et al. 1993; Wightman et al. 1993) and one year after the milestone discovery of the second miRNA Let-7, which is conserved across many species (Pasquinelli et al. 2000; Reinhart et al. 2000), three papers simultaneously revealed that the stealth world of small RNAs (smRNAs), called microRNAs (miRNAs), probably exists in all species (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Since then, smRNA research and applications have spread to almost every corner of the biological sciences. To date, in addition to in-depth investigations of their biogenesis and biological functions, miRNAs have been developed as therapeutic tools, as diagnosis and prognosis markers, and as markers for predicting the response to treatments (Bartel 2009; Kim et al. 2009; Galasso et al. 2010; Nagpal et al. 2010). The progress in our understanding of miRNA biogenesis, function, and mechanism corresponds with the profilling of smRNAs in many organisms, novel species of smRNAs, and novel smRNAs within each species.

smRNA profiling has traditionally relied on cloning and sequencing of individual RNAs using standard molecular biological methods. In the most commonly used approach, adaptor oligonucleotides are joined to the 3’ and 5’ termini of smRNAs, called microRNAs (miRNAs), probably exists in all species (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Since then, smRNA research and applications have spread to almost every corner of the biological sciences. To date, in addition to in-depth investigations of their biogenesis and biological functions, miRNAs have been developed as therapeutic tools, as diagnosis and prognosis markers, and as markers for predicting the response to treatments (Bartel 2009; Kim et al. 2009; Galasso et al. 2010; Nagpal et al. 2010). The progress in our understanding of miRNA biogenesis, function, and mechanism corresponds with the profilling of smRNAs in many organisms, novel species of smRNAs, and novel smRNAs within each species.

smRNA profiling has traditionally relied on cloning and sequencing of individual RNAs using standard molecular biological methods. In the most commonly used approach, adaptor oligonucleotides are joined to the 3’ and 5’ termini of smRNAs, and the ligation products are reverse-transcribed and PCR-amplified to generate a cDNA library. This procedure represents a significant technical challenge because it requires three gel purification steps. In addition, thousands of clones have to be individually sequenced to identify the smRNA population. This standard protocol is labor-intensive, painstaking, and requires large amounts of starting materials, and therefore is not practical for many research or clinical settings (Pfeffer et al. 2005). Although cost can be substantially reduced through concatenating smRNA fragments containing both adaptors to sequence multiple clones simultaneously, the expense remains a major obstacle in thoroughly surveying any smRNA population. Next-generation sequencing (NGS) technology was first used to discover smRNA by applying massive parallel signature sequencing to survey the smRNA library of Arabidopsis thaliana (Lu et al. 2005). Since then, many modified smRNA profiling procedures based on NGS have been developed and tested on various platforms (Lu et al. 2007; Hafner et al.

2008; Tang et al. 2010). To date, human miRNAs alone have 1048 unique sequence entries in miRBase 16 (Griffiths-Jones et al. 2010). The NGS technology also helped to discover other smRNAs and has become one of the invaluable tools for smRNA research (Lu et al. 2005). The release of Illumina’s smRNA 1-d cloning protocol v1.5 further sped up and simplified the procedure. This protocol drastically reduces the difficulties in smRNA cDNA library construction by eliminating the need for gel purification of smRNA/smRNA-adaptors—the most challenging and critical steps in smRNA discovery—and by merely requiring ~1 µg of total RNA (Fig. 1A).

Because of the bridge amplification in Solexa, both 5’ adaptors and 3’ adaptors need to be complementary to the oligo bars on the flowcell surface during cluster generation and subsequent sequencing-by-synthesis (SBS). This creates a problem using bases within the sequences of either adaptor for barcoding samples to sequence in the same flowcell. Extra nucleotides at the 3’ end of the 5’ adaptor, 5’ end of the 3’ adaptor, or both are possible choices for multiplexing in the same flowcell. There are some disadvantages to using the 3’ adaptor for indexing. First, the 3’ ends of smRNAs are heterogeneous in bases or in length produced by biogenesis processing or end base modification, which will increase the difficulty in separating the smRNA sequence from the linker. Second, SBS more likely generates errors close to the end of the short reads, which will reduce the accuracy of sequencing (Wu et al. 2007; Liu et al. 2010). Last, the v1.5 protocol specifically requires the pre-adenylated 5’ end of the 3’ adaptor, which is expensive to make. Therefore, the 3’ end of the 5’ adaptor is the best choice for multiplexing in the same flowcell for sequencing smRNAs of ~17–32 nt, such as piRNAs, tRNA halves, snRNAs, and snoRNA-derived smRNAs (Fig. 1A; Ender et al. 2008; Taft et al. 2009; Thomson and Lin 2009; Pederson 2010; Riedmann and Schwenert 2010). In this study, the addition of 3 nt at the 3’ end of the default 5’ adaptor was tested for multiplexing. Serious biases in small RNA species coverage and abundance occurred by the choice of 5’ adaptors (including the default 5’ adaptor). This kind of artifact, consistent with the previous observation by Linsen et al. (2009) is systemic, reproducible, and cannot be eliminated by simply repeating the run. A solution to this bias is crucial to the accuracy of small RNA digital profiling. In this study, we demonstrate that by using a pool of 64 customized 5’ adaptors, the profiling bias can be minimized, and the sequencing results can be significantly optimized. We further show that this index pool of 64 adaptors can be further grouped into four smaller index pools each containing 16 adaptors to minimize profiling biases and facilitate multiplexing.

RESULTS

It is reportedly possible to add several nucleotides to the 5’ adaptor’s tail to barcode for miRNA profiling using Solexa (Cronn et al. 2008; Witten et al. 2010). Our initial testing of adding 3 nt to the Illumina default 5’-adaptor tail resulted in substantial biases in the profiling results (data not shown). Therefore, we tested 64 adaptors using all three base combinations and expected to find at least several adaptors that can work similarly for multiplexing under the assumption that there may be some ligation biases (Supplemental Table S1A).

Small RNA sequencing of human pool RNAs using customized 5’ adaptors

To balance smRNA population and mimic the realistic situation of smRNA profiling such as the abundance and the 3’ and 5’ heterogeneity of the smRNA sequences, we used total RNAs pooled from 20 human tissues (Supplemental Table S1B). Each customized adaptor was individually ligated with smRNA 3’-adaptor molecules and smRNA libraries from eight adaptors pooled together for cluster generation and SBS (pooled individual ligations; we refer to these runs as “Test 1”). To our surprise, no two adaptors worked identically. We found that serious biases in results were introduced by the customized 5’ adaptors being tested as well as by the default 5’ adaptor of Solexa (Illumina) (Fig. 1B). There were biases in the number of total reads, percentage of reads that were aligned to the genomes, distribution of smRNAs to different categories, coverage of different species in the same smRNA population, and abundance of individual smRNAs. Read counts of some individual miRNAs were even different by ~200-fold (Supplemental Fig. S1A; Supplemental Tables S2, S7). To eliminate the possibility that our smRNA reads processing software produced this type of bias, we tested other miRNA profiling software implementing different data analysis algorithms. Although the ranks of miRNAs by read counts were not identical, the results analyzed by miRExpress were closely comparable to the results analyzed by Novoalign using the same data sets from lane 1 in Test 1 (Supplemental Fig. S1B; Wang et al. 2009). Looking at the top 20 miRNAs in Figure 1C, it seemed that the default Solexa 5’ adaptor was incapable of sequencing some miRNAs such as miR-1, miR-145, miR-215, miR-126, and miR-451, but favored some miRNAs, such as miR-20a, miR-17, miR-155, miR-1246, and miR-146a. After checking the distribution of the first three nucleotides in all human miRNAs as well as the secondary structures and thermal dynamic properties of all 64 adaptors (Supplemental Fig. S1C; Supplemental Table S2, lanes 2,3,9), we found that none of them had a reasonable correlation with the result bias. Interestingly, the clustered Test 1 results revealed that all three added nucleotides influenced the results. This clustering result was roughly divided into eight groups (Fig. 1B). Based on these eight groups, we divided the 64 adaptors into eight pools each with eight adaptors that have balanced bases at each position. We further merged them into four pools, two
pools, and one pool as mixtures of all 64 adaptors (Supplementary Table S1C). Correlation efficiency showed that the eight-pool adaptors (pooled data of Test 1) were adequate enough to provide significantly more compact results when compared with the random groups (Fig. 2B). Pooled results from Test 1 indicate that index pooling reduced variation in the miRNA expression. Larger index pools have less miRNA expression bias among same-size pools, and larger index pools have less variation than smaller pools (Supplemental Fig. S2).

**FIGURE 1.** (A) A diagram of Illumina v1.5 protocol using customized 5’ adaptors. (B) Heatmap of clustered 64 customized 5’-adaptor results (partial shown; see Supplemental Material for the whole heatmap). Total reads of each adaptor from Test 1 were normalized to the max of the sum of all miRNA reads in each index by scaling up total reads in each adaptor to this max. Each miRNA in an adaptor was scaled up according to the ratio of the total reads in an index to the max. Then, the normalized reads were log₂-transformed and clustered as complete linkage and used to generate the heatmap. (C) Result comparison of using customized 5’ adaptor versus Illumina default 5’ adaptor. The top 20 miRNAs are shown.
Small RNA sequencing of human pool RNAs using pool of customized 5′ adaptors

To reduce the work load and cost, we used the above adaptor pool for a 5′-adaptor ligation to an smRNA-3′ adaptor for further testing (pooled adaptors during ligation; we refer to these runs as “Test 2”). This approach will allow us to use one group of eight-, 16-, 32-, or all 64-adaptors to barcode one sample if the adaptors will not interfere with each other during ligation. Lane-to-lane comparison (pooled data of Test 2) in Test 2 showed consistent results among lanes using the adaptor pools (Fig. 2C; Supplemental Fig. S3A; Supplemental Tables S4, S7).

Test 2 results showed that the four pools of 16-adaptors were enough to minimize the results bias (Supplemental Fig. S3B). The miRNA expression measured between four pools, two pools, and the 64-adaptor pools showed ~99% correlation (Supplemental Fig. S3C). The correlation efficiency using the four-pool barcoding to run four samples per flowcell in comparison with that of running one sample is ~0.99. Only ~10% of miRNAs were undetectable in the four-samples run when compared with the one sample run per lane (Supplemental Table S5). The 10% missed were typically less abundant miRNAs. Therefore, four-pool barcoding to run four samples per flowcell is expected to identify all highly expressed miRNAs with an expression measurement that is comparable to running one sample per lane.

Comparison of Test 2 data showed that the profiling coverage of miRNAs increased by almost 100% using the pooled adaptors compared with that of using the default 5′ adaptor (Table 1). Reads from all runs in both Test 1 and Test 2 showed that 893 miRNAs were detected (Table 2). This is ~81% of the 1100 human mature miRNA/miRNA* sequences documented in miRBase 15. The miRNA* family is classified as low-abundance strands from the same precursor miRNA. The low coverage and bias of the Illumina default adaptor may also play a role in detecting low-abundance miRNA and miRNA*.

However, we observed that the abundance was reversed in 21 pairs of miRNA/miRNA* in our results, and that there were 28 miRNAs with less than 10 times the abundance of their corresponding miRNA* (Supplemental Table S6). About 20% of miRBase 15 documented human miRNAs were absent in our study (Supplemental Table S6). One reason is that our pooled RNA did not cover all human tissues (Supplemental Table S1B). It has been reported that methods based on hybridization, rather than ligation, may be superior for quantitative measurement of miRNA abundance (Tian et al. 2010). We noticed that most of the 42xx and 43xx human miRNAs were missing in our results. These miRNAs were recently predicted from sequences cloned from human embryonic stem cells and neural precursors using the SOLiD platform (Goff et al. 2009). We also noticed that most of the 55x to 66x human miRNAs that were cloned with miRAGE were either absent or present at low abundance in our results (Supplemental Table S6; Cummins et al. 2006).

Validation of adaptor pooling strategy by Northern blot analysis

To validate our pooling strategy, we sequenced total RNA isolated from a breast tissue (#AM6952) (Supplemental...
and strongly concentration-dependent, coverage and abundance are crucial to miRNAs’ biological function and their utilization as biomarkers in clinical settings. Extensive smRNA profiling data have already been produced by NGS in the past four years. Data integrity of these profiling results is critical to the conclusions drawn by these studies. We suspect that the result bias may also be present in other data, such as whole-genome sequencing, Chip-Seq, and RNA-Seq, which are conducted by Solexa and other NGS platforms. The adaptor pooling strategy is an easy work-around solution to get “bona fide” deep-sequencing results. However, we speculate that our work-around solution cannot completely address this pitfall because we chose 3 nt for barcoding to maximize our reads for all short noncoding RNAs.

The choice of 3 nt is merely to simplify the experimental testing. It is certainly possible to add >3 nt to the tail of the 5′ adaptor, add nucleotides to the head of the 3′ adaptor, or add a combination of both. But, more extra nucleotides may not be a smart choice. The use of more nucleotides as index could introduce adaptor/adaptor, adaptor/smRNA hybridization problems in addition to wasting of some sequencing results. The error rate in NGS is a concern for multiplexing using the index pool because indexes could switch from one to another. The observed NGS error rate in our GAX II platform is ∼1%. At this rate, we expected that the influence on the results is minor because only highly expressed smRNAs are functionally important to the studied system. We pulled out all reads that perfectly match to the miR-143 mature sequence in lane 1 of Test 1 and confirmed the error rate (data not shown). It is also possible to use the linker pool at one adaptor and barcoding in the other adaptor. Theoretically, three nucleotides are enough to multiplex several samples with the sacrifice of tolerable sequencing depth on the assumption that adapters will work similarly.

We do not yet have a plausible explanation for how our pool strategy can minimize the biases. Ligation most likely plays some role in producing the observed results bias since the initial step of the protocol is based on ligation. The 3′

Table S4) using the eight-pool adaptors. The results showed consistent performance among the eight pools (Supplementary Figs. S4, S7). Next, we performed Northern blot analysis to assess miR-21, miR-1, miR-126, miR-101, and miR-205 expression levels in breast RNA versus human tissue pool RNA. Comparison of the relative abundance of these miRNAs—as measured by normalized read counts—between the breast RNA sequencing results and human RNA pool sequencing results both using the eight-pool adaptors was consistent with the Northern blot results. However, the breast RNA sequencing results obtained using the eight-pool adaptors versus the human RNA pool sequencing results with the default 5′ adaptor did not agree with the Northern blot analysis results (Fig. 3).

**DISCUSSION**

In summary, our study revealed a pitfall in the Solexa platform that sequencing results are biased by the short nucleotide fragments being sequenced. Therefore, the results obtained may not reflect the “true” biological composition. This confirms some of the bias in small RNA digital gene expression profiling using high-throughput sequencing that Linsen et al. (2009) previously reported. This kind of artifact cannot be resolved through repeated sequencing of an identical sample. (Test 1 data can be considered as 64 repeated runs with one-eighth of the sequencing depth or eight repeated runs with the normal sequencing depth. Lanes 2–6 in Test 2 can be considered as five repeated runs with the normal sequencing depth that showed dramatic repeatable results over using the default adaptor run in Lane 1 of the same test.)

Many smRNA research projects rely on the initial profiling of smRNAs. It is proper to profile similar samples using the same protocol, but our results implied that different samples may have problems when comparing the “absolute” abundance and coverage of small RNAs, making it difficult to know the real rank of small RNAs in one sample. Since miRNA-to-target interaction is directly

**TABLE 2.** Detection coverage of annotated miRNAs in miRBase 15

<table>
<thead>
<tr>
<th>miRNA raw read counts</th>
<th>The whole study</th>
<th>By %</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥1</td>
<td>893</td>
<td>81.18</td>
</tr>
<tr>
<td>≥5</td>
<td>821</td>
<td>74.64</td>
</tr>
<tr>
<td>≥10</td>
<td>776</td>
<td>70.55</td>
</tr>
<tr>
<td>≥100</td>
<td>603</td>
<td>54.82</td>
</tr>
<tr>
<td>≥1000</td>
<td>407</td>
<td>37.00</td>
</tr>
</tbody>
</table>

- Many smRNA research projects rely on the initial profiling of smRNAs. It is proper to profile similar samples using the same protocol, but our results implied that different samples may have problems when comparing the “absolute” abundance and coverage of small RNAs, making it difficult to know the real rank of small RNAs in one sample. Since miRNA-to-target interaction is directly and strongly concentration-dependent, coverage and abundance are crucial to miRNAs’ biological function and their utilization as biomarkers in clinical settings. Extensive smRNA profiling data have already been produced by NGS in the past four years. Data integrity of these profiling results is critical to the conclusions drawn by these studies. We suspect that the result bias may also be present in other data, such as whole-genome sequencing, Chip-Seq, and RNA-Seq, which are conducted by Solexa and other NGS platforms. The adaptor pooling strategy is an easy work-around solution to get “bona fide” deep-sequencing results. However, we speculate that our work-around solution cannot completely address this pitfall because we chose 3 nt for barcoding to maximize our reads for all short noncoding RNAs.

The choice of 3 nt is merely to simplify the experimental testing. It is certainly possible to add >3 nt to the tail of the 5′ adaptor, add nucleotides to the head of the 3′ adaptor, or add a combination of both. But, more extra nucleotides may not be a smart choice. The use of more nucleotides as index could introduce adaptor/adaptor, adaptor/smRNA hybridization problems in addition to wasting of some sequencing results. The error rate in NGS is a concern for multiplexing using the index pool because indexes could switch from one to another. The observed NGS error rate in our GAX II platform is ∼1%. At this rate, we expected that the influence on the results is minor because only highly expressed smRNAs are functionally important to the studied system. We pulled out all reads that perfectly match to the miR-143 mature sequence in lane 1 of Test 1 and confirmed the error rate (data not shown). It is also possible to use the linker pool at one adaptor and barcoding in the other adaptor. Theoretically, three nucleotides are enough to multiplex several samples with the sacrifice of tolerable sequencing depth on the assumption that adapters will work similarly.

We do not yet have a plausible explanation for how our pool strategy can minimize the biases. Ligation most likely plays some role in producing the observed results bias since the initial step of the protocol is based on ligation. The 3′

**DISCUSSION**

In summary, our study revealed a pitfall in the Solexa platform that sequencing results are biased by the short nucleotide fragments being sequenced. Therefore, the results obtained may not reflect the “true” biological composition. This confirms some of the bias in small RNA digital gene expression profiling using high-throughput sequencing that Linsen et al. (2009) previously reported. This kind of artifact cannot be resolved through repeated sequencing of an identical sample. (Test 1 data can be considered as 64 repeated runs with one-eighth of the sequencing depth or eight repeated runs with the normal sequencing depth. Lanes 2–6 in Test 2 can be considered as five repeated runs with the normal sequencing depth that showed dramatic repeatable results over using the default 5′ adaptor did not agree with the Northern blot analysis results (Fig. 3).

**TABLE 2.** Detection coverage of annotated miRNAs in miRBase 15

<table>
<thead>
<tr>
<th>miRNA raw read counts</th>
<th>The whole study</th>
<th>By %</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥1</td>
<td>893</td>
<td>81.18</td>
</tr>
<tr>
<td>≥5</td>
<td>821</td>
<td>74.64</td>
</tr>
<tr>
<td>≥10</td>
<td>776</td>
<td>70.55</td>
</tr>
<tr>
<td>≥100</td>
<td>603</td>
<td>54.82</td>
</tr>
<tr>
<td>≥1000</td>
<td>407</td>
<td>37.00</td>
</tr>
</tbody>
</table>

- Many smRNA research projects rely on the initial profiling of smRNAs. It is proper to profile similar samples using the same protocol, but our results implied that different samples may have problems when comparing the “absolute” abundance and coverage of small RNAs, making it difficult to know the real rank of small RNAs in one sample. Since miRNA-to-target interaction is directly and strongly concentration-dependent, coverage and abundance are crucial to miRNAs’ biological function and their utilization as biomarkers in clinical settings. Extensive smRNA profiling data have already been produced by NGS in the past four years. Data integrity of these profiling results is critical to the conclusions drawn by these studies. We suspect that the result bias may also be present in other data, such as whole-genome sequencing, Chip-Seq, and RNA-Seq, which are conducted by Solexa and other NGS platforms. The adaptor pooling strategy is an easy work-around solution to get “bona fide” deep-sequencing results. However, we speculate that our work-around solution cannot completely address this pitfall because we chose 3 nt for barcoding to maximize our reads for all short noncoding RNAs.

The choice of 3 nt is merely to simplify the experimental testing. It is certainly possible to add >3 nt to the tail of the 5′ adaptor, add nucleotides to the head of the 3′ adaptor, or add a combination of both. But, more extra nucleotides may not be a smart choice. The use of more nucleotides as index could introduce adaptor/adaptor, adaptor/smRNA hybridization problems in addition to wasting of some sequencing results. The error rate in NGS is a concern for multiplexing using the index pool because indexes could switch from one to another. The observed NGS error rate in our GAX II platform is ∼1%. At this rate, we expected that the influence on the results is minor because only highly expressed smRNAs are functionally important to the studied system. We pulled out all reads that perfectly match to the miR-143 mature sequence in lane 1 of Test 1 and confirmed the error rate (data not shown). It is also possible to use the linker pool at one adaptor and barcoding in the other adaptor. Theoretically, three nucleotides are enough to multiplex several samples with the sacrifice of tolerable sequencing depth on the assumption that adapters will work similarly.

We do not yet have a plausible explanation for how our pool strategy can minimize the biases. Ligation most likely plays some role in producing the observed results bias since the initial step of the protocol is based on ligation. The 3′

**TABLE 2.** Detection coverage of annotated miRNAs in miRBase 15

<table>
<thead>
<tr>
<th>miRNA raw read counts</th>
<th>The whole study</th>
<th>By %</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥1</td>
<td>893</td>
<td>81.18</td>
</tr>
<tr>
<td>≥5</td>
<td>821</td>
<td>74.64</td>
</tr>
<tr>
<td>≥10</td>
<td>776</td>
<td>70.55</td>
</tr>
<tr>
<td>≥100</td>
<td>603</td>
<td>54.82</td>
</tr>
<tr>
<td>≥1000</td>
<td>407</td>
<td>37.00</td>
</tr>
</tbody>
</table>
Small RNA deep sequencing was performed using Illumina GAII at the COH Sequencing/SeqXaCore following the manufacturer’s new protocols (Preparing Samples for Small RNA Sequencing Using the Alternative v1.5 Protocol, Small RNA sample prep Kit #FC-102-1009, Illumina, Inc.), with minor optimization. Equal amounts of 20 human tissue RNAs from Ambion human total RNA survey panel (AM6000) were mixed as the human total RNA pool and was used for smRNA library construction, cluster generation, and deep sequencing. Briefly, 1.0 μg of pooled total RNA was ligated to the v1.5 sRNA 3’ adaptor (UCUACGUAGCGCUCUCUGCUUG) with T4 RNA Ligase 2, truncated (New England BioLabs) for 1 h at 20°C. The constructed smRNA library was first reverse-transcribed using GX1 (5’-CAAGCAGAA GACGCCATACGA) as the RT primer then subjected to PCR amplification for 12 cycles, using the primers GX1 and GX2 (5’-AATGATACGCGCCACACCAGATTCAGTTCTACAGTC GA) then followed by 6% TBE PAGE gel purification with size selection (for targeted smRNAs of 17–35 nt). The purified library was quantified using qPCR with a forward primer (5’-CAAGCAG AAGCCGGTACAGCG) and a reverse primer (5’-AATGATACGCG CCACCGA). The quantified denatured miRNA library was loaded in 1 mL of hybridization buffer to a final DNA concentration of 8 PM then used for single read flow cell cluster generation and 42 cycle (42 nt) sequencing performed using the Illumina Genome Analyzer II (GAI).

**Solexa data processing using the Novocraft suite**

The image and base-calling analyses were performed using the Illumina Genome Analysis Package (OLB v1.6 and CASAVA v1.6) with the Illumina default chastity filter off and generated 42-nt-long reads in FASTQ format. The filter discards ~30–40% of the reads based on bases with low-quality scores at the first 25 bases of a read. The reads quality control occurs during alignment by Novoalign software, which filters out low-quality reads based on the base quality score. (Comparison study by our deep sequencing core shows that the Novoalign filter and Illumina default chastity filter work similarly [data not shown].) Reads were separated using the Novobasecode software (http://www.novocraft.com). Reads were aligned to human genome hg18 (NCBI build 36.1), using the Novoalign software (http://www.novocraft.com) with default settings except for the following modifications: An aligned read requires a minimum of 16 nt in length with good base quality after 3’-adapter trimming and a perfect match to the reference genome. A mapping table was created using the human miRNA mature sequences from miBase 15 and aligned back to the hg18 genome afterward. If a read could be aligned to multiple loci, it was randomly assigned to one locus. To summarize the expression levels of smRNAs, reads of an smRNA falling into the mapped smRNA regions within 5-base extensions were treated as valid reads and counted. If an smRNA could be mapped to multiple genome loci, the counts at each locus were summed as the total number of reads for that smRNA. For statistical analysis of the miRNA expression levels, the total counts of miRNAs in each barcoded sample were scaled to a constant and log2-transformed with an offset of one. A normalized miRNA count was calculated as

\[
E_{\text{miR}(i),b} = \log_2 \left( 1 + n_{\text{miR}(i),b} \right) \times \max \left\{ \sum_{i=1}^{k} n_{\text{miR}(i),b} \right\}
\]

where \(E_{\text{miR}(i),b}\) is the log2 value of an miRNA \(\text{miR}(i)\) expression level in a barcoded sample \(b\), \(n_{\text{miR}(i),b}\) is the miRNA \(\text{miR}(i)\) read count in barcoded sample \(b\), and \(\sum_{i=1}^{k} n_{\text{miR}(i),b}\) is the sum of all miRNA \(k\) in total) reads in barcoded sample \(b\). miRNAs with more than six normalized reads in at least 10% of the samples were kept for subsequent expression correlation efficiency analysis. Hierarchical clustering with one Pearson correlation as distance measurement and complete linkage was conducted with Cluster 3.0 and viewed in Java TreeView (de Hoon et al. 2004; Saldanha 2004).
Analyses of Solexa reads using miRExpress

miRExpress software analyzed miRNAs by matching reads directly to pre-miRNA or mature miRNA sequences (Wang et al. 2009). We used the default settings to count reads that can perfectly match to pre-miRNAs.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Charles Kim, Spencer Yeh, and Mansze Kong for reading and correcting the manuscript.

Authors’ contributions: G.S., X.W., J.R., H.G., and Y.Y. conceived and designed the experiments. J.W. and H.G. constructed the smRNA cDNA libraries and performed the Solexa runs. X.W., G.S., H.L., and X.L. performed bioinformatics and statistical analysis. G.S. and Y.Y. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Received June 4, 2011; accepted September 15, 2011.

REFERENCES


Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen de Hoon MJ, Imoto S, Nolan J, Miyano S, 2004. Open source bioinformatics and statistical analysis. G.S. and Y.Y. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Received June 4, 2011; accepted September 15, 2011.

REFERENCES


Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen de Hoon MJ, Imoto S, Nolan J, Miyano S, 2004. Open source bioinformatics and statistical analysis. G.S. and Y.Y. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Received June 4, 2011; accepted September 15, 2011.

REFERENCES


Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen de Hoon MJ, Imoto S, Nolan J, Miyano S, 2004. Open source bioinformatics and statistical analysis. G.S. and Y.Y. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Received June 4, 2011; accepted September 15, 2011.

REFERENCES


Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen de Hoon MJ, Imoto S, Nolan J, Miyano S, 2004. Open source bioinformatics and statistical analysis. G.S. and Y.Y. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Received June 4, 2011; accepted September 15, 2011.

REFERENCES


Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen de Hoon MJ, Imo...
A bias-reducing strategy in profiling small RNAs using Solexa

Guihua Sun, Xiwei Wu, Jinhui Wang, et al.

RNA published online October 20, 2011

Supplemental Material
http://rnajournal.cshlp.org/content/suppl/2011/10/05/rna.028621.111.DC1

Published online October 20, 2011 in advance of the print journal.

License

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.