Laboratory Investigation (2014) 94, 350–358

© 2014 USCAP, Inc All rights reserved 0023-6837/14

Robust global microRNA expression profiling using next-generation sequencing technologies

Shirley Tam[1,2,3], Richard de Borja[4], Ming-Sound Tsao[1,2,5] and John D McPherson[1,3,5]

miRNAs are a class of regulatory molecules involved in a wide range of cellular functions, including growth, development and apoptosis. Given their widespread roles in biological processes, understanding their patterns of expression in normal and diseased states will provide insights into the consequences of aberrant expression. As such, global miRNA expression profiling of human malignancies is gaining popularity in both basic and clinically driven research. However, to date, the majority of such analyses have used microarrays and quantitative real-time PCR. With the introduction of digital count technologies, such as next-generation sequencing (NGS) and the NanoString nCounter System, we have at our disposal many more options. To make effective use of these different platforms, the strengths and pitfalls of several miRNA profiling technologies were assessed, including a microarray platform, NGS technologies and the NanoString nCounter System. Overall, NGS had the greatest detection sensitivity, largest dynamic range of detection and highest accuracy in differential expression analysis when compared with gold-standard quantitative real-time PCR. Its technical reproducibility was high, with intrasample correlations of at least 0.95 in all cases. Furthermore, miRNA analysis of formalin-fixed, paraffin-embedded (FFPE) tissue was also evaluated. Expression profiles between paired frozen and FFPE samples were similar, with Spearman's $\rho > 0.93$. These results show the superior sensitivity, accuracy and robustness of NGS for the comprehensive profiling of miRNAs in both frozen and FFPE tissues.

Laboratory Investigation (2014) **94,** 350–358; doi:10.1038/labinvest.2013.157; published online 20 January 2014

KEYWORDS: microarray; microRNAs; NanoString; next-generation sequencing; quantitative real-time PCR

Since their initial discovery in nematodes two decades ago,[1] microRNAs (miRNAs) have come to be recognized as key regulators of many biological processes and promising biomarkers for disease. miRNAs are endogenous, small noncoding nucleotides that negatively regulate gene expression post-transcriptionally by recognizing and binding to the 3′-UTR of mRNAs in a sequence-specific manner.[2] Depending on the degree of sequence complementarity, this interaction can mediate either translational inhibition or mRNA degradation.[3] Over 30% of human protein-coding genes are predicted to be conserved targets of miRNAs.[4] Compared with the ~30 000 estimated mRNAs, there are currently ~1400 miRNAs deposited in public databases;[5] a single miRNA can potentially target many hundreds of genes, causing substantial effects on gene expression networks.[6] Thus, variation in the abundance level of a few miRNAs is likely to be associated with development and progression of diseases. Understanding their patterns of expression could provide new insights into complex biological processes and the possible clinical implications of miRNA dysfunction.

Technological advances have brought about a multitude of platforms for the systematic evaluation of miRNAs. These tools are largely derived from mRNA expression analysis and array-based comparative genomic hybridization. However, compared with other nucleic acids, the analysis of miRNA is complicated by several factors:[7] their short length, highly similar sequences between family members, discrimination between mature and primary forms, and their rapid rate of discovery. Understanding the strengths and limitations of different profiling approaches can help apply these tools more effectively for specific biological applications.

Microarrays have been used extensively for the simultaneous profiling of thousands of genes in a single experiment. Along with quantitative real-time polymerase chain reaction

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada; [2]Princess Margaret Cancer Centre, University Health Networks, Toronto, ON, Canada; [3]Genome Technologies, Ontario Institute for Cancer Research, Toronto, ON, Canada; [4]Informatics and Bio-Computing, Ontario Institute for Cancer Research, Toronto, ON, Canada and [5]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada

Correspondence: Dr JD McPherson, PhD, Genome Technologies, Ontario Institute for Cancer Research, MaRS Building, South Tower, 101 College Street, Toronto, ON, Canada M5G 0A3.

E-mail: John.McPherson@oicr.on.ca

Received 4 August 2013; revised 12 November 2013; accepted 19 December 2013

(qPCR), they are the most commonly used platform to evaluate the expression of known miRNAs.[8–10] They are relatively cost-effective, quick from RNA labeling to data generation and simple to use.[11] However, the short length of these molecules does not always allow for optimal probe design, as the miRNA sequences themselves have to be used as the probe sequences.[7] The evolution of digital count technologies has provided new methods for miRNA profiling. Next-generation sequencing (NGS) allows for the simultaneous discovery of new miRNAs and confirmation of known miRNAs. It can overcome the shortcomings of microarrays, which can suffer from background signal and cross-hybridization issues; however, sample preparation involves many steps that can introduce biases and sequencing errors,[12] and the computational tools for analysis are in their infancy. A more recent innovation in expression profiling is the NanoString nCounter system, a hybridization-based technology that can detect specific nucleic acid molecules from low amounts of starting material without the need for reverse transcription or cDNA amplification. Multiplexed sequence-specific probe pairs are first hybridized in-solution to transcripts of interest, and abundance levels are determined by tabulating the number of target-specific fluorescent tag for each miRNA assayed.[13]

In this study, five pairs of non-small-cell lung cancer cell lines and their corresponding xenograft models were profiled on four platforms representative of different detection mechanisms: the Illumina Human microRNA Expression Profiling v.2 microarray (Illumina, San Diego, CA, USA), Life Technologies SOLiD™ 4 (Life Technologies, Carlsbad, CA, USA), Illumina HiSeq 2500 (Illumina) and the NanoString nCounter Human miRNA Expression Assay v.1 (NanoString, Seattle, WA, USA). The platforms were evaluated according to the following criteria: (i) interplatform concordance, (ii) concordance with qPCR, the current gold-standard assay for expression measurements, (iii) detection of differentially expressed (DE) miRNAs in a biologically relevant setting and (iv) dynamic range of detection. On the basis of these criteria, NGS platforms were the most robust for the comprehensive expression profiling of miRNAs.

MATERIALS AND METHODS

Cell Culture, Xenografts and RNA Isolation

A549, H460, H520, H1264 and RVH6849 cells were cultured in RPMI-1640, supplemented with 10% FBS and 1 × penicillin/streptomycin. Xenografts were grown by subcutaneous injection of two million trypsin-dissociated tumor cells into non-obese diabetic/severe-combined immunodeficient mice. Total RNA was isolated from confluent cell lines and fresh-frozen xenograft tissues using Trizol reagent (Life Technologies) according to the manufacturer's instructions, followed by DNAse I treatment (Life Technologies) and purification using the Qiagen RNeasy kit (Qiagen, Venlo, The Netherlands). For formalin-fixed, paraffin-embedded (FFPE) samples, 10–15 μm tissue sections were first deparaffinized using xylene and ethanol washes. Total RNA was isolated using the Norgen FFPE RNA Purification kit (Norgen BioTek, Ontario, Canada) as per the manufacturer's instructions. RNA quantity and quality was assessed using the Qubit Fluorometer (Life Technologies), Nanodrop 1000 (Nanodrop Technologies, Wilmington, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Illumina miRNA Assay

Aliquots of the RNA samples were provided to the University Health Networks (UHN) Microarray Centre for microarray analysis using the Illumina Human microRNA Expression Profiling v.2 Panels. In brief, 200 ng of total RNA from each sample was labeled using the Illumina microRNA Assay kit according to the manufacturer's protocol (Illumina). The labeled samples were hybridized to a 12-sample Universal BeadChip, incubated at 60 °C for 30 min and hybridized at 45 °C for 18 h. The BeadChips were then washed and stained as per the Illumina protocol, and scanned on the iScan (Illumina). The data files were quantified in BeadStudio v.3.3.8 (Illumina) and loaded into the R statistical environment (v.2.14.0) using the lumi package (v.2.6.0)[14] implemented in the Bioconductor libraries.[15] The probes were reannotated against miRBase v.16,[16–19] $\log_2$ transformed and normalized using the robust spline normalization algorithm. No background correction was performed.

Life Technologies SOLiD™ Sequencing

Total RNA samples were fractionated using the flashPAGE™ fractionator system (Life Technologies) and small RNAs ($\sim<40$ nt) were recovered by ethanol precipitation. Small RNA enrichment was confirmed using the small RNA Lab-on-a-Chip kit and the Bioanalyzer 2100. Fifteen nanograms of small RNAs from each sample was used to construct cDNA libraries according to the SOLiD™ Total RNA-Seq protocol (Life Technologies). The libraries were amplified by emulsion PCR, and beads were deposited on a slide for sequencing in-house on the SOLiD™ 4 System.

Short-read sequences were output in color space FASTA format with corresponding base qualities. The reads were trimmed for adapter sequences using Cutadapt (v.0.9.3),[20] and the resulting color space FASTA file contained sequences with either bases trimmed from the 3′ end corresponding to the adapters or without trimmed bases. Untrimmed reads were aligned to a tRNA and homopolymer reference to remove random sequences. The remaining reads were aligned to miRBase v.16 mature miRNA reference using SHRiMP (v.2).[21] Only perfectly and uniquely aligned reads were retained and counted to determine the abundance of each annotated miRNA. The summarized count data were loaded into the R statistical environment (v.2.14.0); samples with higher coverage were downsampled before normalization by total count scaling.

Illumina Sequencing

Libraries for sequencing on the HiSeq 2000 and 2500 systems were prepared as per Illumina TruSeq Small RNA protocol (Illumina). In brief, 3′ and 5′ adapters were sequentially ligated to the ends of small RNAs fractionated from 5 μg of total RNA, and reverse transcribed to generate cDNA. The cDNA was amplified (11 cycles of PCR) using a common primer complementary to the 3′ adapter, and a primer containing 1 of 48 index sequences. Samples were size-selected (140–160 bp fragments) on a 6% polyacrylamide gel, purified, quantified and pooled for multiplexed sequencing. The resulting pooled libraries were hybridized to oligonucleotide-coated single-read flow cells for cluster generation using the Illumina cBot or on-instrument (HiSeq 2500), and subsequently sequenced on the Illumina HiSeq 2000 or HiSeq 2500 for 50 sequencing cycles.

Base calling was performed using CASAVA (v.1.8.2) (Illumina) and short-read sequences were output in FASTQ format with corresponding base quality scores. Quality control (QC) of the raw sequences from each sequenced library was investigated using FastQC (v.0.9.1)[22] to check for homopolymers, adapters and distribution of base quality. The raw data were initially filtered for reads containing ambiguous base calls, which did not meet the Illumina chastity filter based on quality measures. The remaining reads were trimmed for adapters and mapped to the miRBase v.16 mature miRNA reference using Novocraft's Novoalign (v.2.07.14).[23] The summarized count data were loaded into the R statistical environment (v.2.14.0) and normalized by linear regression using the median count value of each miRNA across the samples as reference.

NanoString nCounter System miRNA Assay

One hundred nanograms of total RNA from each sample was provided to the UHN Microarray Centre for NanoString nCounter analysis. The samples were prepared for nCounter miRNA expression profiling according to the manufacturer's recommendations (NanoString). For each sample, a scan of 600 fields of view (FOV) was imaged.

Before data normalization, nCounter data imaging QC metrics were assessed. There was no significant discrepancy between the FOVs attempted and the FOVs counted. The binding density for the samples ranged between 0.24 and 0.72—within the typical recommended range. The raw data were loaded into the R statistical environment (v.2.14.0), and reannotated against miRBase v.16. First, probes indicated to have some level of background were corrected using probe level adjustment factors. Then, the geometric mean of the positive controls was used for code count normalization, while the background was estimated using the mean of the negative controls. Sample input amounts were normalized to the geometric mean of five housekeeping mRNA controls (ACTB, B2M, GAPDH, RPL19 and RPL10) included in the assay, and finally to total miRNA count.

Probe Reannotation

The Illumina Human microRNA Assay was designed against miRBase v.12.05, additional novel sequences derived using Illumina sequencing technology and novel miRNAs discovered in two separate published studies.[24,25] These probes were first aligned to the genome (hg19), and then to miRBase v.16 using BWA (v.0.5.9-r16).[26] Following removal of nonspecific and non-uniquely aligned probes, 812 probes were retained and reannotated. The NanoString Human miRNA Expression Assay v.1 kit profiled 734 human and human-associated viral miRNAs from miRBase v.14 (http: //www.nanostring.com/). As the probe sequences were not provided by the manufacturer, the provided target sequences were aligned to miRBase v.16. Thirteen of the 654 human miRNA probes corresponded to retracted miRBase entries, whereas four were annotated differently between miRBase v.14 and v.16. Thus, 641 probes were considered.

TaqMan® miRNA Quantitative Real-time PCR

qPCR was performed using TaqMan® microRNA assays (Life Technologies). cDNA for each miRNA of interest was synthesized from an input of 5 ng of total RNA using the TaqMan® microRNA Reverse Transcription Reagents (Life Technologies) and specific reverse transcription primers (Life Technologies). Real-time PCR with TaqMan® probes was performed on a Life Technologies ViiA™ 7 Real-Time PCR System using the following conditions: 10 min at 95 °C, followed by 40 cycles of 95 °C for 30 s and 60 °C for 1 min. All assays were performed in triplicates. C_T values were determined using the SDS software with automatic baseline and threshold settings. The data were loaded into the R statistical environment (v.2.14.0) and preprocessed. Triplicate C_T values were averaged and normalized to the geometric mean of let-7g, miR-191 and miR-335-3p, which were selected as endogenous controls based on geNorm[27] and NormFinder[28] (Supplementary Figure 1). The normalized expression was calculated as $\log_2|2^{-\Delta CT}|$. C_T values >36 were considered to be below the limit of detection.

Data Analysis

All data analyses and graphical representations were performed and generated in the R statistical environment (v.2.14.0). Agglomerative hierarchical clustering was performed using Spearman's correlation coefficients as input, Euclidean distance as the distance metric and complete linkage. Results were visualized with heatmaps using the lattice (v.0.20-0) and latticeExtra (v.0.6-19) packages.

To identify DE miRNAs between cell lines and xenografts from the microarray and qPCR data sets, linear models were fit to each individual miRNA. Each miRNA was tested for changes in abundance levels using empirical Bayes moderated t-statistics, where the standard errors were moderated across the probes.[29] For the sequencing and NanoString data, the count data was modeled as negative binomial distributed and the gene-wise dispersion was estimated by the Cox–Reid profiled-adjusted method.[30] An empirical Bayes approach was applied to moderate the variance. DE miRNAs were identified using a generalized linear model-likelihood ratio test. A paired sample design was used for all analyses, matching each cell line with its respective xenograft model. The *P*-values were adjusted for multiple testing using the false discovery rate approach.[31] Significant miRNAs were selected based on an arbitrary $|\text{fold change}| \geq 2$ and $P_{\text{adjusted}} \leq 0.05$. The performance of each platform in identifying DE and non-DE miRNAs was evaluated using a binary classification system and qPCR results as the true values. Statistical analyses were performed using the limma (v.3.10.2)[32] and edgeR (v.2.4.3)[33] packages for the R statistical environment (v.2.14.0).

RESULTS

miRNA Profiling Study Design

In comparison with previous cross-platform analyses, which used tissues of significantly different origins,[11,34–36] we have chosen to compare the miRNA expression profiles between more closely related samples—representative of a realistic application of profiling experiments. Previous gene expression profiling studies comparing tumor cell lines grown *in vitro* and *in vivo* have identified subsets of genes upregulated in cultured cells.[37–39] We hypothesize that the miRNA expression profiles are also altered between cells grown in culture and as xenografts. We have selected this as our biological model for evaluating the characteristics and applications of different miRNA profiling technologies because they are renewable sources of material, which provide reproducible results when used with the same protocol and at the same passage.[40] Total RNA was isolated from five pairs of non-small-cell lung cancer cell lines ($n = 5$) and their corresponding xenograft models ($n = 5$), and miRNA profiles were analyzed on each of the four platforms (see Figure 1a for experimental and data analysis workflow). All samples ($n = 10$) were analyzed once on each of the four platforms, except when evaluating the technical reproducibility of NGS technologies, where samples were sequenced in duplicates on the HiSeq 2500. A total of 1146 and 654 miRNAs were assayed on the microarray and NanoString platforms, respectively, whereas 1072 and 1084 miRNAs were detected by SOLiD and HiSeq 2500 sequencing, respectively (Figure 1b). All raw and preprocessed data reported in this study have been deposited in the Gene Expression Omnibus repository, under accession no. GSE51508.

Probe Reannotation and Data Filtering

Owing to the frequent update of the miRBase miRNA database,[5] probe-based miRNA detection platforms are usually designed against different versions of the database. The name and length of miRNAs can vary between versions,

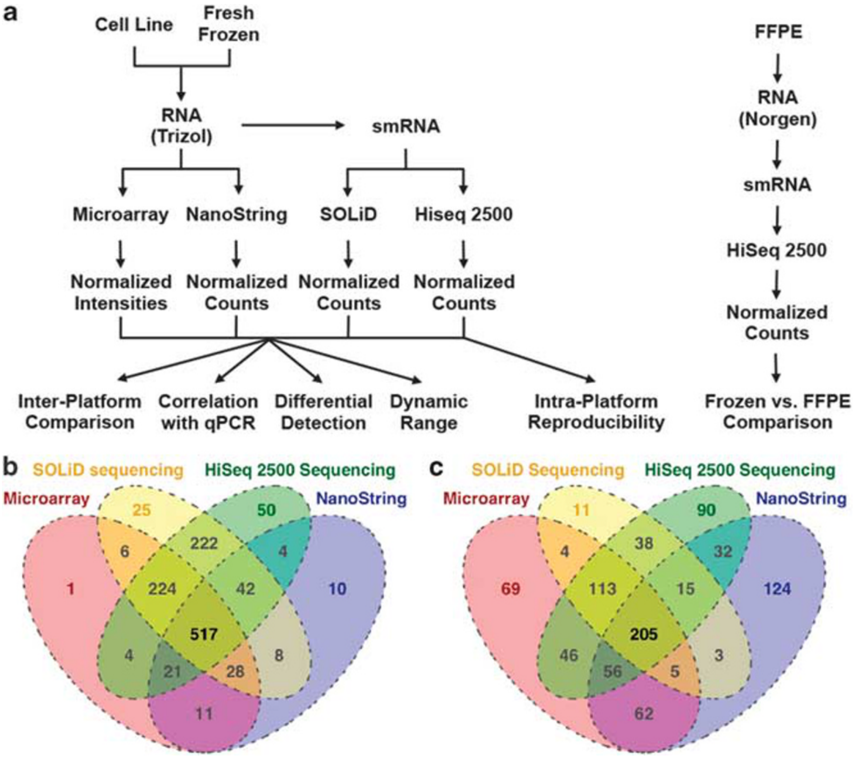

Figure 1 Experimental outline. (**a**) Experimental and data analysis workflow. (**b**) Overlap of microRNAs (miRNAs) interrogated by all platforms. (**c**) Overlap of miRNAs after filtering for detectable levels above a specified threshold across all platforms. This filtering process reduced the overlap of 517 miRNAs to 205 miRNAs.

resulting in an annotation problem for cross-platform comparisons. Furthermore, miRNA entries are sometimes retracted based on overlap with annotated snoRNAs or tRNAs or invalidated by subsequent work. Therefore, probes were remapped using the most up-to-date information from current genome sequence databases (see Materials and Methods).

We focused our analyses on 517 miRNAs that were interrogated by all three digital count platforms, and whose microarray probes had no predicted cross-hybridization (Figure 1b). Our filtering criteria for detection for the microarray was a threshold detection *P*-value of 0.05, and for the sequencing data, a threshold of 10 normalized count-per-million that mapped to mature miRNA sequences in at least 50% of the samples. Likewise, for NanoString, only miRNAs that were detected in at least 50% of the samples were considered. On the basis of these criteria, 560, 693, 595 and 502 miRNAs were detected using the microarray, SOLiD, HiSeq 2500 and NanoString, respectively. The intersection of all three platforms was 205 miRNAs (Figure 1c).

Interplatform Variability

To allow for a non-biased comparison of the platforms' performance, correlation analyses were performed on the set of 205 miRNAs detected across the four platforms. Fold changes between the cell lines and xenografts were plotted for each pairwise platform comparison, and Spearman's correlation values were calculated (Figure 2). Results from the two sequencing platforms showed good correlation to each other ($\rho = 0.75$), and to the microarray data ($\rho = 0.73$ and 0.69 for SOLiD and HiSeq 2500, respectively). By contrast, correlation with the NanoString data was only moderate for all comparisons ($\rho \sim 0.50$). A Friedman test revealed significant differences in fold change measurements across the platforms ($\chi^2 = 159.26$, $P < 2.2 \times 10^{-16}$). *Post hoc* analysis using Wilcoxon's signed-rank tests showed significant differences between all platforms (Supplementary Table 1).

Comparison with qPCR Results

Results from miRNA profiling experiments are regularly validated by qPCR because of its detection sensitivity, specificity, reproducibility and large dynamic range. Accordingly, the expression of 86 miRNAs was analyzed using TaqMan qPCR assays, including miRNAs identified as significantly and nonsignificantly DE by at least one platform, and non-DE miRNAs. Of these, 68 miRNAs were targeted by all four platforms and had C_T values <36 in at least one sample. The fold changes between xenografts and cell lines generated by each platform were plotted against qPCR results, and Spearman's correlation values were calculated (Figure 3). Strong correlation was observed for all platforms, with the relative accuracy of NGS to qPCR results being the highest ($\rho = 0.86$, $P < 2.2 \times 10^{-16}$). A lower, but highly significant, correlation was observed between NanoString and qPCR ($\rho = 0.72$).

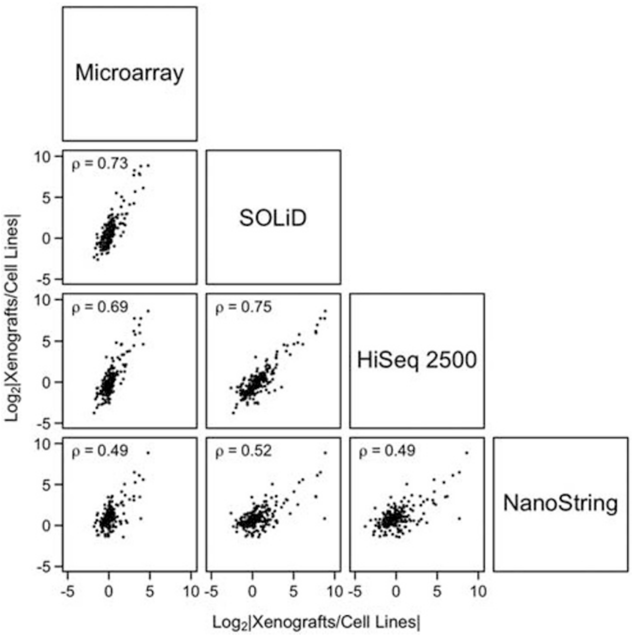

Figure 2 Variability across profiling technologies. Fold changes for microRNAs (miRNAs) detected across all platforms were plotted for each pairwise comparison. Spearman's correlation values are shown on the upper left corner of each plot. Correlation of NanoString to all platforms was lower than between microarray and next-generation sequencing (NGS) platforms.

Differential Expression Analysis: Cell line–Xenograft Comparison

The purpose of many profiling experiments is to assess the molecular changes that can alter a given physiological state, such as uncovering differences in global expression levels between a 'normal' control and a 'disease' state. A list of DE miRNAs between cell lines and xenografts was generated for each platform using all 517 common miRNAs. The overlap is displayed in a Venn diagram in Supplementary Figure 2. To assess the robustness of each platform in correctly predicting DE miRNAs, qPCR results were used as ground truth. Of the 68 miRNAs targeted by all four platforms, 23/68 were determined to be DE by qPCR ($|\text{fold change}| \geq 2$, $P_{\text{adjusted}} \leq 0.05$). Using these data, the sensitivity, specificity and accuracy of each platform for the detection of differential expression was evaluated (Table 1 and Supplementary Table 2). Although the specificity (true-negative rate) was highest for HiSeq 2500 and NanoString, the sensitivity (true-positive rate) was superior for NGS technologies, with the HiSeq 2500 system being the most accurate for differential expression analysis (ACC = 0.88) when compared with qPCR.

Dynamic Range of Detection

The dynamic range of detection affects the accurate quantification of transcripts with varying abundance between sample classes. For platforms with a small dynamic range, differences in abundance could be underestimated or even undetected; this fold change compression is characteristic of microarray technology. The fold changes corresponding to the 68 miRNAs validated by qPCR were examined for each platform (Supplementary Table 2). The magnitude of these fold changes show that NGS technologies have the largest dynamic range (at least 10 logs) as measured by $\log_2$ count values or signal intensity, followed by NanoString (~8 logs), and lastly, the microarray platform (<5 logs). As qPCR is often expected and assumed to detect a wide variety of transcripts present at very different levels, log-ratio compression or expansion was examined relative to qPCR results (Figure 3). Although the best-fitted line from linear regression analysis showed that the fold changes determined by NGS platforms is comparable to qPCR owing to the large dynamic range of detection (slope, $\beta_1 = 0.96$ and 0.87 for SOLiD and HiSeq 2500, respectively), the microarray and NanoSring suffered from strong fold-change compression ($\beta_1 \sim 0.50$).

Table 1 Performance of platforms in predicting differential expression

	Microarray	SOLiD	HiSeq 2500	NanoString
Specificity	0.84	0.76	0.93	0.93
Sensitivity	0.43	0.91	0.78	0.61
Accuracy	0.71	0.81	0.88	0.82

Technical Reproducibility of miRNA Sequencing

So far, we have shown that data generated using NGS technologies are comparable to the more established microarrays, can be validated by qPCR and is superior for differential expression analysis. To further validate the use of NGS for global miRNA expression profiling, the technical reproducibility of data generated on the HiSeq 2500 was evaluated. Unsupervised hierarchical clustering was used to visualize the similarity between the expression profiles of the technical and biological replicates (Figure 4a). All technical replicates clustered closely together, separate from biological replicates (Spearman's $\rho > 0.95$ for all cases, see Supplementary Figure 3). In addition, all cell lines were more similar to each other than they were to their xenograft counterparts. Next, the coefficient of variation (CV) of miRNAs detected in technical duplicates was examined (Figure 4b). As expected, more variability was present across biological replicates compared with the variability between technical replicates.

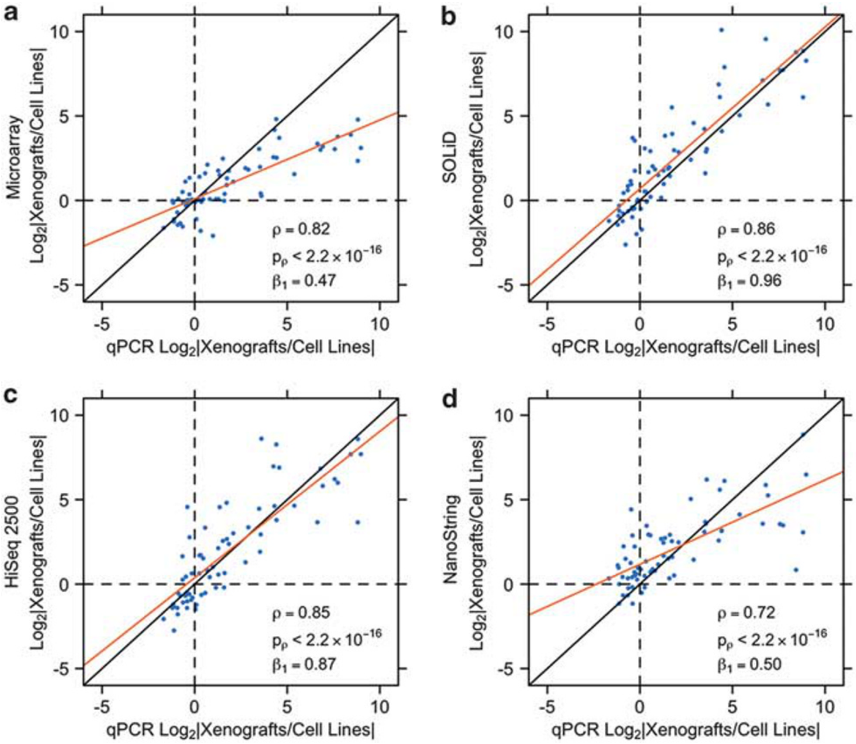

Figure 3 Comparison of microRNA (miRNA) profiling results with quantitative real-time polymerase chain reaction (qPCR). Fold changes for 68 miRNAs determined from (**a**) microarray, (**b**) SOLiD sequencing, (**c**) HiSeq 2500 sequencing and (**d**) NanoString were plotted against qPCR results. Correlation coefficients and slopes are listed at the bottom left corner of each plot. The solid black line shown is the identify function, which represents perfect accuracy relative to qPCR. The orange line represents the best-fitted line from linear regression analysis. $\beta_1 < 1$ indicates log ratio compression relative to qPCR.

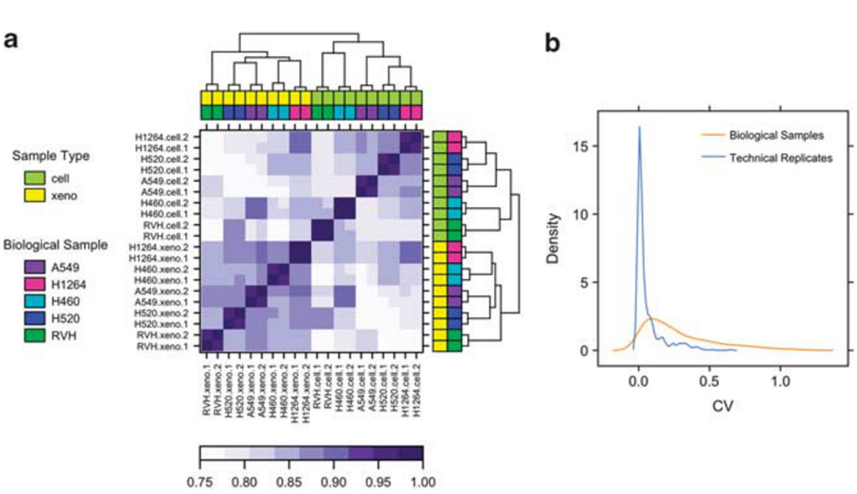

Figure 4 Technical reproducibility of microRNA (miRNA) sequencing. Reproducibility of next-generation sequencing (NGS) technologies for miRNA profiling was evaluated by profiling samples in duplicates on the HiSeq 2500 system. (**a**) Spearman's correlation coefficients between all samples were calculated and subjected to unsupervised hierarchical clustering. The resulting heatmap and dendrogram show the high similarity between all technical replicates, with $\rho > 0.95$ for all cases. Furthermore, all cell lines clustered together, separate from their xenograft counterparts. (**b**) The coefficient of variation of the expression values between a set of technical replicates was calculated for miRNAs detected in both duplicates. The distribution of the duplicate coefficient of variations (CVs) is compared with the CV between biological groups; replicate CV measures are lower.

Consistent miRNA Profiles Between Frozen and FFPE Specimen

Although cell lines and fresh-frozen tissues yield high-quality RNA suitable for expression-profiling studies, FFPE tissues are often the only tissue type available in the clinical setting. To evaluate the feasibility of sequencing FFPE tissue specimen and the ability to recover miRNA profiles similar to their snap-frozen counterparts, three pairs of matched frozen and FFPE xenografts tumors were profiled on the Illumina HiSeq 2000. The age of the FFPE blocks ranged from approximately 3 to 5 years old, and the fragmentation end point of the isolated RNA samples was around 100–150 nucleotides (Supplementary Figure 4). Degraded species, such as primary and precursor miRNAs, that contain the same sequences as their corresponding mature form can cause higher background noise. Hierarchical clustering showed similarity between the expression profiles of paired frozen and FFPE samples (Spearman's $\rho > 0.93$) (Figures 5a and b); however, samples of different biological origin were less correlated (Spearman's $\rho < 0.82$).

DISCUSSION

The growing number of studies examining the global changes in miRNA expression suggests that these molecules are associated with a variety of human diseases, such as neuropsychiatric disorders,[41] diabetes[42] and cancer.[43,44] Recently, it has also been shown that miRNA profiles of various cancer types, including chronic lymphocytic leukemia,[45] lung,[46] breast[47] and ovarian cancer[48] may potentially contain diagnostic and prognostic information.[43,44] Accurately detecting and quantifying these differences in miRNA abundance between physiologically distinct states provides insight into the role of these regulatory molecules in complex biological processes and in the pathogenesis of diseases. Although several technological options are available to analyze miRNA expression comprehensively, the detection of miRNAs and subsequent interpretation of such data can be strongly influenced by the specific platforms used. An informed perspective of the characteristics of these different profiling tools could guide the choice of a platform best suited to the biological question being investigated.

Our study examines the comparability of four miRNA profiling platforms representative of different technologies, which differ significantly in their mechanisms of detection. Although both the Illumina microarray and the NanoString system are based on the hybridization of miRNAs to a set of predetermined probes, the former relies on static hybridization and binding intensity, and the latter depends on in-solution hybridization and digital counting. Sequencing, on the other hand, detects and quantifies the number of miRNAs directly. The performance attributes of these profiling technologies were evaluated using five pairs of cells lines and their xenograft models. From this work, we conclude that (i) the concordance across miRNA profiling technologies is, at best, only moderate, (ii) the optimal usage of a given platform is dependent on the specific application and (iii) NGS technologies show superior sensitivity, accuracy and robustness for global miRNA profiling in both frozen and FFPE tissue.

A surprising observation from this study is the moderate correlation between NGS platforms and NanoString. Our expectation was that both these digital count technologies would generate highly similar expression profiles, but the resulting interplatform correlation was only moderate ($\rho \sim 0.50$) (Figure 2). Furthermore, comparison of the NanoString data to qPCR results yielded the lowest, but still

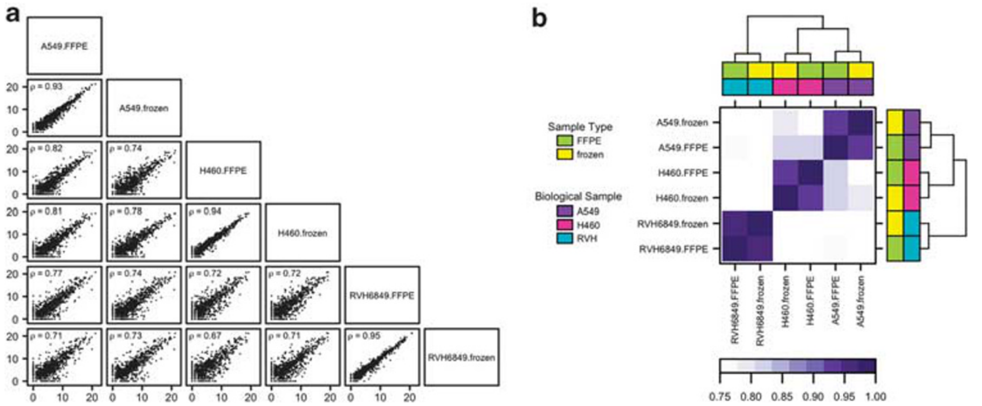

Figure 5 MicroRNA (miRNA) sequencing of formalin-fixed, paraffin-embedded tissue (FFPE) specimen. Comparison of miRNA expression profiles between matched frozen and FFPE specimen shows the validity of using next-generation sequencing NGS technologies for miRNA profiling of FFPE blocks. (**a**) The distribution of expression values between matched pairs is much tighter than with other biological samples, with Spearman's correlation values of at least 0.93. (**b**) Unsupervised hierarchical clustering also shows high similarity between paired samples.

highly significant, correlation of 0.72 ($P<2.2\times10^{-16}$). Data generated from similar platforms appear to be more reproducible than those generated on considerably different platforms—both microarray and sequencing, which include PCR-based processing steps, were more correlated to each other ($\rho=0.69$–0.73) and to qPCR ($\rho=0.82$–0.86) than to NanoString (Figures 2 and 3). Thus, results obtained using considerably different technologies are not always highly concordant, compromising their comparability.

The sensitivity, specificity and dynamic range of detection of a platform affect the identification of DE transcripts. A platform with low sensitivity will generate many false-negative calls, whereas higher sensitivity and reduced specificity would result in a higher number of false positives. The low sensitivity and small dynamic range of microarrays in comparison with other profiling tools limit their use in comparative analysis; strong fold-change compression affects the ability to identify accurate changes in transcript abundance between sample groups. Optimal use of microarrays for comparative analysis should be limited to transcripts of which differences in abundance fall within the boundary of dynamic range. Alternatively, the lower sensitivity of NanoString for miRNA detection may be attributed to the miRNA-specific protocol and assay. Although a starting RNA input of 100 ng (used to generate our data) may be sufficient for mRNA profiling,[13] this may be inadequate for the detection of miRNAs, which represent only a small fraction ($\sim$0.01%) of the mass in a total RNA sample.[49] Assays that do not involve amplification would require more starting material because of lower detection sensitivity. In addition, a single miRNA transcript can represent more than 50% of the counts in a sample, as observed in our data. Using higher amount of starting input RNA, which is reflected in the protocol change by the UHN Microarray Centre, combined with a less complex custom nCounter probe library, or a probe set with highly abundant species attenuated, could potentially enhance the signal from miRNAs of interest. Thus, the amplification-free design of the NanoString platform may be better suited for the interrogation of a smaller subset of miRNAs.

The sensitivity, large dynamic range of NGS, along with its consistent prediction of fold changes when compared with gold-standard qPCR, support its use for discovery-oriented and exploratory miRNA profiling experiments. Furthermore, high technical reproducibility (Spearman's $\rho>0.95$) eliminates the need for technical replicates when profiling samples using NGS technologies. Finally, we also show that miRNA expression profiles of samples that have been subjected to the formalin fixation and paraffin-embedding process are highly similar to their snap-frozen counterpart, even with the use of different RNA extraction protocols. This will allow the use of NGS technologies to evaluate tissues in the clinical setting where FFPE blocks may be the only available sample type, rendering these archival samples an invaluable source of readily available tissue for retrospective studies of human diseases.

Although microarrays and qPCR have been used extensively for expression profiling and are highly reproducible, they are limited to the detection of known targets identified at the time of assay development and manufacturing. Furthermore, comparison of data generated on different hybridization-based platforms is complicated by their differences in miRBase content. High-throughput sequencing provides an unbiased approach to miRNA profiling, is less prone to batch effects, has a large dynamic range of detection and avoids any cross-hybridization issues. With the rapid increase in miRNAs being discovered and deposited in public databases, NGS can offer a more comprehensive view of the miRNA transcriptome.

Supplementary Information accompanies the paper on the Laboratory Investigation website (http://www.laboratoryinvestigation.org)

ACKNOWLEDGMENTS

We thank Roya Navab and Devang Panchal for providing the xenograft samples, Michelle Sam, Ada Wong and Jeremy Johns for technical support and Melania Pintilie for helpful statistical advice and discussions. ST was supported by an award from the Ontario Graduate Scholarship (OGS) Program. The work is supported by funds from the Ontario Institute for Cancer Research (JDM), Canadian Cancer Society Research Institute Grant No. 020527 (MST) and the Ontario Ministry of Health and Long Term Care. MST is the M Qasim Choksi Chair in Lung Cancer Translational Research at the Princess Margaret Hospital.

DISCLOSURE/CONFLICT OF INTEREST

The authors declare no conflict of interest.

1. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75:843–854.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–297.
3. Zeng Y, Yi R, Cullen BR. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. Proc Natl Acad Sci USA 2003;100:9779–9784.
4. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005;120:15–20.
5. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods Mol Biol 2006;342:129–138.
6. Lim LP, Lau NC, Garrett-Engele P, *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005;433:769–773.
7. Wark AW, Lee HJ, Corn RM. Multiplexed detection methods for profiling microRNA expression in biological samples. Angew Chem Int Ed Engl 2008;47:644–652.
8. Liu C-G, Calin GA, Meloon B, *et al.* An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. Proc Natl Acad Sci USA 2004;101:9740–9744.
9. Babak T, Zhang W, Morris Q, *et al.* Probing microRNAs with microarrays: tissue specificity and functional inference. RNA 2004;10:1813–1819.
10. Castoldi M, Schmidt S, Benes V, *et al.* A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). RNA 2006;12:913–920.
11. Pradervand S, Weber J, Lemoine F, *et al.* Concordance among digital gene expression, microarrays, and qPCR when measuring differential expression of microRNAs. BioTechniques 2010;48:219–222.
12. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Brief Bioinform 2009;10:490–497.
13. Geiss GK, Bumgarner RE, Birditt B, *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 2008;26:317–325.
14. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547–1548.
15. Gentleman RC, Carey VJ, Bates DM, *et al.* Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
16. Griffiths-Jones S. The microRNA registry. Nucleic Acids Res 2004;32: 109D–111D.
17. Griffiths-Jones S, Grocock RJ, van Dongen S, *et al.* miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006;34: D140–D144.
18. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011;39:D152–D157.
19. Griffiths-Jones S, Saini HK, van Dongen S, *et al.* miRBase: tools for microRNA genomics. Nucleic Acids Res 2007;36:D154–D158.
20. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011;17:10–12.
21. David M, Dzamba M, Lister D, *et al.* SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 2011;27:1011–1012.
22. Andrews S. FASTQC. A quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc 2010; accessed on 10 July 2013.
23. Hercus C. Novocraft short read alignment package. Available at: http://www.novocraft.com 2009; accessed on 10 July 2013.
24. Berezikov E, Thuemmler F, van Laake LW, *et al.* Diversity of microRNAs in human and chimpanzee brain. Nat Genet 2006;38:1375–1377.
25. Berezikov E, van Tetering G, Verheul M, *et al.* Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res 2006;16:1289–1298.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–1760.
27. Vandesompele J, De Preter K, Pattyn F, *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 2002;3:RESEARCH0034.
28. Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 2004;64: 5245–5250.
29. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article3.
30. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881–2887.
31. Benjamini Y, Hochberg Y. J R Stat Soc Ser B (Methodological) 1995; 57:289–300.
32. Smyth GK. limma: Linear Models for Microarray Data. Statistics for Biology and Health: New York, NY, USA, 2005, pp 397–420.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140.
34. Sato F, Tsuchiya S, Terasawa K, *et al.* Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. PLoS One 2009;4:e5540.
35. Yauk CL, Rowan-Carroll A, Stead JD, *et al.* Cross-platform analysis of global microRNA expression technologies. BMC Genom 2010;11:330.
36. Ach RA, Wang H, Curry B. Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. BMC Biotechnol 2008;8:69.
37. Creighton C, Kuick R, Misek DE, *et al.* Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. Genome Biol 2003;4:R46.
38. Creighton CJ, Bromberg-White JL, Misek DE, *et al.* Analysis of tumor–host interactions by gene expression profiling of lung adenocarcinoma xenografts identifies genes involved in tumor formation. Mol Cancer Res 2005;3:119–129.
39. Gieseg MA, Man MZ, Gorski NA, *et al.* The influence of tumor size and environment on gene expression in commonly used human tumor lines. BMC Cancer 2004;4:35.
40. van Staveren WCG, Solís DYW, Hébrant A, *et al.* Human cancer cell lines: experimental models for cancer cells *in situ*? For cancer stem cells? Biochim Biophys Acta 2009;1795:92–103.
41. Abelson JF, Kwan KY, O'Roak BJ, *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 2005;310:317–320.
42. Poy MN, Eliasson L, Krutzfeldt J, *et al.* A pancreatic islet-specific microRNA regulates insulin secretion. Nature 2004;432:226–230.
43. Volinia S, Calin GA, Liu C-G, *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci USA 2006;103:2257–2261.
44. Lu J, Getz G, Miska EA, *et al.* MicroRNA expression profiles classify human cancers. Nature 2005;435:834–838.
45. Calin GA, Ferracin M, Cimmino A, *et al.* A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med 2005;353:1793–1801.
46. Yanaihara N, Caplen N, Bowman E, *et al.* Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 2006;9: 189–198.
47. Iorio MV, Ferracin M, Liu C-G, *et al.* MicroRNA gene expression deregulation in human breast cancer. Cancer Res 2005;65:7065–7070.
48. Iorio MV, Visone R, Di Leva G, *et al.* MicroRNA signatures in human ovarian cancer. Cancer Res 2007;67:8699–8707.
49. Shingara J, Keiger K, Shelton J, *et al.* An optimized isolation and labeling platform for accurate microRNA expression profiling. RNA 2005;11:1461–1470.