The Journal of Molecular Diagnostics, Vol. 24, No. 4, April 2022

the Journal of Molecular Diagnostics
jmdjournal.org

Small RNA-Sequencing for Analysis of Circulating miRNAs

Benchmark Study

Peter Androvic,*† Sarka Benesova,*‡ Eva Rohlova,*§ Mikael Kubista,*¶ and Lukas Valihrach*

From the Laboratory of Gene Expression, Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Vestec, Czech Republic; Laboratory of Growth Regulators,† Faculty of Science, Palacky University, Olomouc, Czech Republic; Department of Informatics and Chemistry,‡ Faculty of Chemical Technology, University of Chemistry and Technology, Prague, Czech Republic; Department of Anthropology and Human Genetics,§ Faculty of Science, Charles University, Prague, Czech Republic; and the TATAA Biocenter AB,¶ Gothenburg, Sweden

Accepted for publication December 20, 2021.

Address correspondence to Lukas Valihrach, Ph.D., Laboratory of Gene Expression, Institute of Biotechnology of the Czech Academy of Sciences, Prumyslova 595, Vestec 252 50, Czech Republic. E-mail: lukas.valihrach@ibt.cas.cz.

Small RNA-sequencing (RNA-Seq) is being increasingly used for profiling of circulating microRNAs (miRNAs), a new group of promising biomarkers. Unfortunately, small RNA-Seq protocols are prone to biases limiting quantification accuracy, which motivated development of several novel methods. Here, we present comparison of all small RNA-Seq library preparation approaches that are commercially available for quantification of miRNAs in biofluids. Using synthetic and human plasma samples, we compared performance of traditional two-adaptor ligation protocols (Lexogen, Norgen), as well as methods using randomized adaptors (NEXTflex), polyadenylation (SMARTer), circularization (RealSeq), capture probes (EdgeSeq), or unique molecular identifiers (QIAseq). There was no single protocol outperforming others across all metrics. Limited overlap of measured miRNA profiles was documented between methods largely owing to protocol-specific biases. Methods designed to minimize bias largely differ in their performance, and contributing factors were identified. Usage of unique molecular identifiers has rather negligible effect and, if designed incorrectly, can even introduce spurious results. Together, these results identify strengths and weaknesses of all current methods and provide guidelines for applications of small RNA-Seq in biomarker research. *(J Mol Diagn 2022, 24: 386—394; https://doi.org/10.1016/j.jmoldx.2021.12.006)*

Circulating microRNAs (miRNAs) found in various body fluids are attractive candidates for clinical biomarkers.[1] To identify disease-specific miRNAs, small RNA-sequencing (RNA-Seq) has become a method of choice for its high screening capacity, specificity, sensitivity, and ability to quantify isomiRs or detect novel miRNAs.[2,3] Despite many advantages, small RNA-Seq protocols suffer from several limitations that obscure quantification. The classic protocol for small RNA library preparation employs two sequential ligations of adaptors to the 3′ and 5′ ends of the miRNAs (in this study, represented by Norgen, Lexogen, and QIAseq; protocol versions and companies are listed in Supplemental Table S1). However, serious quantification bias is introduced in this process due to unequal ligation efficiencies, leading to systematic over- and underestimation of true miRNA.[4] The effect is particularly pronounced in biofluids, where miRNA concentration and complexity are rather low.[5] Recently, three alternative approaches have been developed to improve quantification accuracy. The first approach uses adaptors with randomized nucleotides increasing the chance of effective ligation (NEXTflex)[6]; the

Supported by Czech Science Foundation P303/18/21942S (M.K., L.V., and S.B.), Czech Health Research Council NU21-08-00286 (L.V. and S.B.), UCT research project A1-FCHT-2021-003 (S.B.). and IBT institutional support RVO 86652036 (P.A., S.B., E.R., M.K., and L.V.).

P.A. and S.B. contributed equally to this work.

Disclosures: Vendors and local distributors of small RNA-Seq preparation protocols supplied free or discounted sample kits for this study.

https://doi.org/10.1016/j.jmoldx.2021.12.006

second approach is ligation-free and employs polyadenylation and template switching during reverse transcription (SMARTer), whereas the third approach relies on ligation of a single 3′ adaptor and subsequent circularization (RealSeq).[7] Additional quantification bias may arise during PCR amplification of libraries. To mitigate PCR bias, unique molecular identifiers (UMIs) have been introduced to identify and remove PCR duplicates (employed in QIAseq protocol), but their effectiveness in small RNA-Seq applications is debated.[8,9] In addition, EdgeSeq, a platform using hybridization probes and targeted sequencing readout, specifically designed for ease-of-use in clinical setting, is available as an alternative to small RNA-Seq. Previous comparative studies performed on a subsets of available methods revealed vast differences in their performance.[7,9–16] However, how current commercial small RNA-Seq methods perform, particularly in a challenging setting such as liquid biopsy samples, is not yet established. Here, we present evaluation of seven commercial small RNA-Seq methods representing all currently available technical approaches for library preparation, with focus on their performance for miRNA quantification in human plasma.

Materials and Methods

Ethics Approval and Consent to Participate

Informed consent was obtained from all volunteers participating in the study. All procedures involving the use of human samples were performed in accordance with the ethical standards of the Institute of Biotechnology of the Czech Academy of Sciences, and with the Declaration of Helsinki.

Samples and RNA Isolation

Blood samples were collected from three healthy volunteers into K_2EDTA BD Vacutainer tubes (Becton Dickinson, Franklin Lakes, NJ) and centrifuged within 30 minutes from collection at 1500 × *g* for 15 minutes at room temperature. Plasma fraction was aspirated and transferred into 2-mL tubes (Eppendorf, Hamburg, Germany) and centrifuged again for 15 minutes at 3000 × *g*. The supernatant was transferred into new 2-mL tubes and stored at −80°C until analysis. Levels of hemolysis were assessed in each sample by measuring absorbance at 414 nm using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA). Total RNA was isolated starting from 7 to 9 plasma aliquots (250 μL) per sample (24 aliquots in total) using miRNeasy Serum/Plasma Advanced Kit (Qiagen, Venlo, the Netherlands) according to the manufacturer's instructions and eluted into 20 μL of nuclease-free water. At the lysis step, 1 μL of isolation spike-in mix and 1 μL of GlycoBlue Coprecipitant (Invitrogen, Carlsbad, CA) were added as described in Androvic et al.[17] Each RNA eluate was assessed for quality of isolation, levels of hemolysis, and presence of inhibitors by two-tailed quantitative real-time RT-PCR (RT-qPCR) panel (Supplemental Table S2), as described in Androvic et al.[17] All 24 RNA eluates were then pooled together to produce a standard plasma RNA sample used in the study. An equimolar mixture of 962 synthetic miRNAs (miRXplore Universal Reference) was purchased from Miltenyi Biotec (Bergisch Gladbach, Germany).

Library Preparation

Libraries were prepared in technical duplicates starting from 5 μL of plasma RNA pool and 5 μL of miRXplore Universal Reference (2×10^6 copies/μL) according to each manufacturer's protocol. The version of the protocol, adaptor concentrations, and number of PCR cycles for each protocol are listed in Supplemental Table S1. Libraries were quantified on the Qubit 3 fluorometer (Thermo Fisher Scientific) and Fragment Analyzer (Agilent, Santa Clara, CA). Libraries generated by the same protocol were pooled and separated on 5% TBE polyacrylamide gel electrophoresis on Mini-PROTEAN Tetra Cell (Bio-Rad Laboratories, Hercules, CA) (Supplemental Figure S1). A region representing RNA fragments of length 140 to 160 nt (ie, fragments of 22 nt ± approximately 10 nt miRNAs with adaptors) was excised from the gel, then DNA was eluted into nuclease-free water and purified with SPRIselect reagent (Beckman Coulter, Brea, CA). All libraries were sequenced in one sequencing run on NextSeq 500 high-output (Illumina, San Diego, CA) with 85-bp single-end reads. A total of 5.8 to 17.9 million reads per library were obtained, with a median of 11 million reads. EdgeSeq libraries were prepared according to the manufacturer's protocol and sequenced in the TATAA Biocenter (Gothenburg, Sweden).

RT-qPCR

Absolute quantification was performed for 35 pre-selected miRNAs using two-tailed RT-qPCR as described in Androvic et al.[18] For each miRNA, the standard curves were generated using miRXplore as the standard to calculate the absolute miRNA concentration in plasma. Both types of samples were identical to the ones used for small RNA-Seq. Briefly, 4 μL of the standard sample (miRXplore) in eight different concentrations (5 to 5×10^7 copies/μL) and 4 μL of plasma sample in two technical replicates were reverse transcribed using a qScript flex cDNA kit (Quantabio, Beverly, MA) in a 20-μL reaction containing a pool of miRNA-specific primers. After cDNA synthesis, the total volume of each cDNA sample was diluted to 200 μL, and 2 μL of the diluted cDNA were used as a template in a 10-μL qPCR reaction containing 1× SYBR Grandmaster Mix (TATAA Biocenter) and 0.4 μmol/L forward and reverse primer. qPCR was performed in two technical replicates for each diluted cDNA sample. The data were processed in Biorad CFX Manager software version 3.1 (Hercules, CA).

Cq values generated by reactions with aberrant melting curves were discarded. After quality control, only 19 miRNAs passing high confidence criteria were used for correlation analysis with RNA-Seq data.

Availability of Data and Materials

The datasets generated and analyzed during the current study are available in the Gene Expression Omnibus database (*https://www.ncbi.nlm.nih.gov/geo*; accession number GSE149513). Scripts used for processing data are available on github repository (*https://github.com/besarka16/Benchmarking-of-small-RNA-seq*, last accessed August 21, 2021).

Data Processing

Raw reads were trimmed with cutadapt tool software version 1.18[19] according to the respective library preparation manual. Reads were filtered for length between 15 and 29 bp. Artificial reads were filtered by mapping of reads with Bowtie software version 1.2.2[20] to rRNA and UniVec databases obtained from the sortmerna github repository. Reads that did not map to UniVec and rRNA sequences were further mapped to relevant references with STAR software version 2.7.3a[21] using "end-to-end" mode, and 5% of the sequence was allowed to mismatch. Reads were counted with featureCounts from the Rsubread R software package version 2.0.1 (*http://subread.sourceforge.net*), and only unique mapping reads were counted. UMI-tools software version 1.0.1 was used for deduplication before counting of mapped reads in QIAseq samples.[22] For comparability with other protocols, non-deduplicated QIAseq data were used for calculation of relevant metrics. Deduplicated QIAseq samples are referred to as QIAseq_UMI. Plasma samples were first mapped to the human genome (GRCh38.95). Reads mapping to genome were further mapped to mature human miRNA sequences in miRBase software version 22.[23] Reads that were not mapped to miRBase were further mapped in descended order to isomiRs, the tRNA database (435 mature tRNA sequences from gtRNAdb), the piRNA database (8 million sequences from piRBase version 2), and the ncRNA database (36 thousand noncoding sequences from ensemble GRCh38). isomiRs were mapped and counted using the isomiRROR tool (GitLab; *https://gitlab.lrz.de/Physio/isomiRROR*) with adjusted settings, when only longer and shorter isomiRs without mismatch in mature sequence were counted. Other small RNA references were mapped with STAR aligner with the same settings as for mapping to miRBase. MiRXplore samples were mapped to the miRXplore reference with the same settings as plasma samples to miRBase. Raw sequencing data and raw count matrices are available on Gene Expression Omnibus database (*https://www.ncbi.nlm.nih.gov/geo/*; accession number GSE149513). All scripts used for data processing and processed data in xlxs file format are available on the github repository (*https://github.com/besarka16/Benchmarking-of-small-RNA-seq*, last accessed August 13, 2021).

Evaluation Metrics

If not stated otherwise, all statistics were calculated separately for each technical replicate and their mean values are shown. All samples were normalized by the CPM method (divided by the total number of reads and multiplied by 1 million). For correlation measures, Pearson coefficients and $\log_2$-transformed values were used, if not stated otherwise. Technical bias was calculated for each miRNA as a fold change of the mean value of two technical replicates from its predicted value. The predicted value was calculated as a number of normalized counts per sample divided by the number of miRNAs in miRXplore (962 or 467 for the EdgeSeq protocol, respectively). The contribution of PCR bias and ligation bias to overall bias in the small RNA-Seq was assessed on samples processed by the QIAseq protocol with usage of the variancePartition software version 1.21.4 R package (*https://bioconductor.org/packages/release/bioc/html/variancePartition.html*), which employed linear mixed model to separate the variance of multiple variables (PCR bias, ligation bias, and technical replicates). Thermodynamic features of miRNAs were calculated by ViennaRNA package version 2.0.[24] Contribution of miRNA sequence features to overall bias was assessed using linear model in R with $\log_2$-fold deviation as a dependent variable. RNA-Seq plasma samples were computationally corrected using division of the normalized counts by a ratio of the measured and expected expression values in the miRXplore sample for the corresponding miRNA.[25] Dependence of the number of detected miRNAs on sequencing depth was assessed by down sampling the raw counts with a random generator for binomial distribution in R software. The number of miRNAs was used as a number of observations, and the number of raw counts belonging to individual miRNAs corresponded to the number of trials. The probability of success in each trial corresponded to the proportion of raw reads at a specific sequencing depth related to the number of raw reads at the original sequencing depth. False positivity was assessed in the miRXplore samples, which were remapped to human miRNAs (miRBase version 22). miRNAs with $\geq$1 count (in both replicates) and absent from the miRXplore reference were considered false hits. Sequence similarity was calculated between all pairs of false hits and the miRXplore reference using the pairwiseAlignment function from Biostrings R package version 2.56.0 (*https://bioconductor.org/packages/release/bioc/html/Biostrings.html*). Alignment scores were normalized by dividing the alignment score by miRNA length, and the miRNA with the maximal score was considered as the best match.

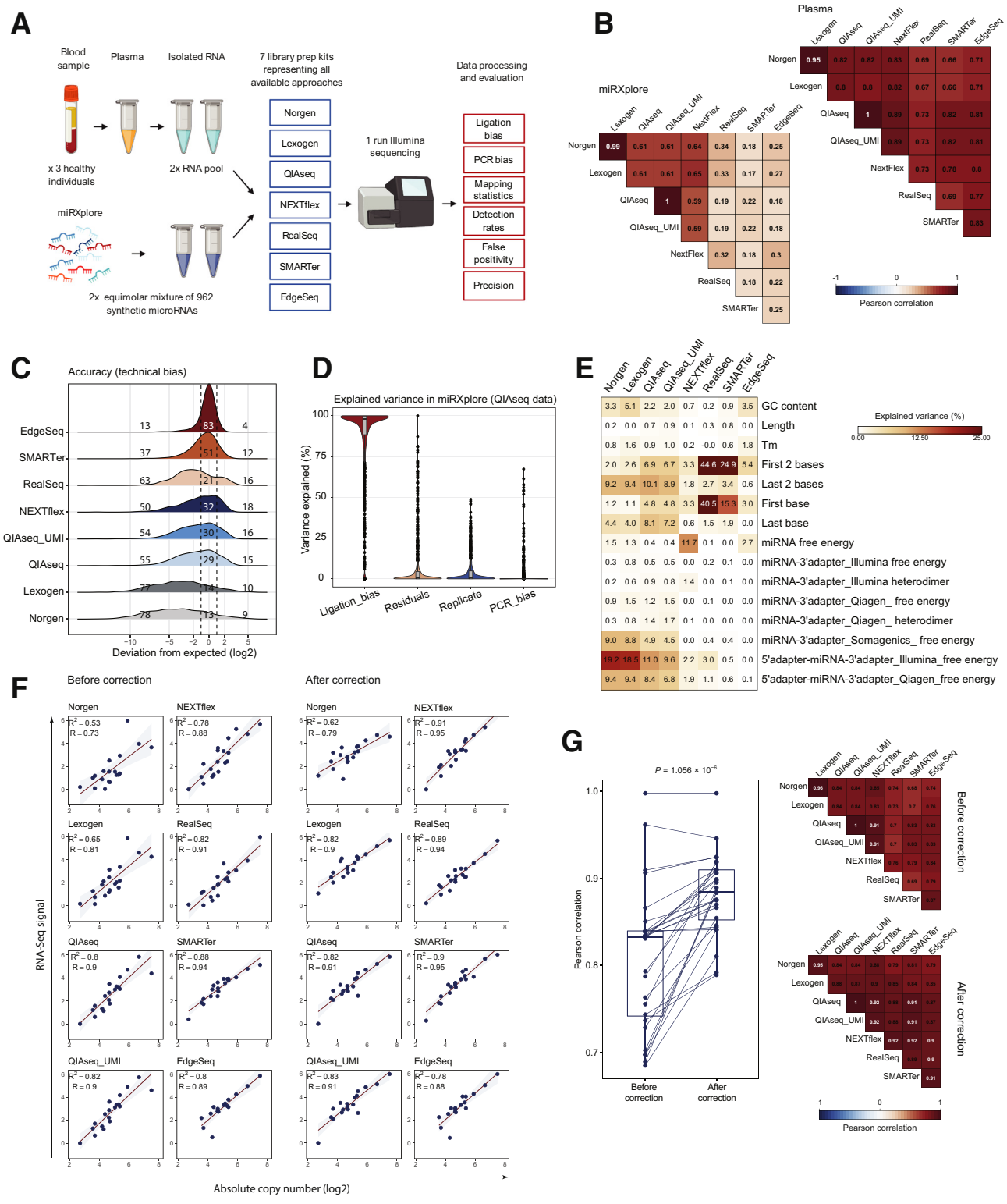

Figure 1 Experimental design, accuracy and technical biases. **A:** Schematic representation of study design. **B:** Correlation heatmaps showing between-protocol reproducibility for miRXplore and plasma samples. **C:** Accuracy determined on miRXplore sample. Density plots show distribution of $\log_2$-fold change between measured and expected value. **Dashed lines** show twofold deviation from expected value; numbers indicate percentage of microRNAs (miRNAs) within and outside of the twofold range. **D:** Percentage of variance in QIAseq data (miRXplore sample) explained by ligation bias, PCR bias, and replicates. **E:** Percentage variance in QIAseq data (miRXplore sample) explained by miRNA sequence characteristics. **F:** Correlation of small RNA-Seq with quantitative real-time RT-PCR (RT-qPCR) data measured in plasma before and after data correction with bias ratios learned from miRXplore samples. **G:** Between-protocol reproducibility for plasma samples before and after data correction using bias ratios learned on miRXplore samples. *P* value from two-tailed paired *t*-test. QIAseq unique molecular identifier (UMI) represents data after deduplication, whereas QIAseq means non-deduplicated data.

Results

Seven commercially available protocols were used to prepare small RNA-Seq libraries in technical duplicates from human plasma and an equimolar mixture of 962 synthetic miRNAs (henceforth called miRXplore) (Figure 1A). Plasma samples were controlled for isolation artifacts and hemolysis (*Materials and Methods*) (Supplemental Table S2); small-RNA library fraction corresponding to miRNAs was gel-purified (Supplemental Figure S1) and samples were sequenced in a single sequencing run to avoid batch effects (except for EdgeSeq). This design allowed for unbiased comparison of protocol performance with biofluids, as well as detailed evaluation of technical biases. All methods showed high within-protocol reproducibility (Supplemental Figure S2), in contrast to low between-protocol reproducibility (Figure 1B), demonstrating that substantial, unique technical bias is introduced by each protocol. The extent of this bias was evaluated by $\log_2$-fold deviation of the measured value from the expected value for each miRNA in the miRXplore sample, where ground truth is known (Figure 1C). EdgeSeq and SMARTer had the least bias, whereas Norgen and Lexogen were most biased, with measured miRNA levels spanning several orders of magnitude. Surprisingly, single-molecule ligation and circularization approach (RealSeq), recently claiming to significantly reduce bias,[7] showed only 21% unbiased miRNAs. In addition, the sequence bias was not reproducible between protocols (Figure 1B), showing that miRNA profiles obtained with different protocols are not comparable.

Although previous studies attributed a large proportion of the bias to adaptor ligation,[4] the contribution of PCR to overall bias is often debated, with reports of negligible[4,26] or substantial effect.[8,9] In the miRXplore sample, the contribution of various factors to overall bias was evaluated using QIAseq data, which employ UMIs and thus allow separation of the PCR contribution from other effects. Ligation bias was highly explanatory for variability in most miRNAs, whereas PCR bias was overall negligible (Figure 1D). This is in agreement with the previous result showing ligation-free protocols (EdgeSeq and SMARTer) are the least biased, whereas the ligation-based protocols are the most biased overall (Figure 1C). Of note, a short UMI length resulting in insufficient complexity of available UMIs was found to lead to erroneous overestimation of PCR bias, a likely cause for the misidentification of its contribution in the previous study[9] (Supplemental Figure S3). To provide insights into mechanisms leading to biased measurements, contributions of miRNA properties to overall variance of measured miRXplore values were evaluated (Figure 1E). First nucleotide in miRNA sequence was highly influential for RealSeq and SMARTer, explaining as much as 44% and 25% of the variability. In addition, the identity of the last nucleotides and free energy of adaptor-miRNA construct, but not the miRNA itself, had an impact with the ligation-based protocols using two defined adaptors including Lexogen, Norgen, and QIAseq. Overall, these results demonstrate that ligation, but not PCR, is a major source of quantification bias in small RNA-Seq data and is influenced by complex and technology-specific factors.

The data revealed that each miRNA is burdened by bias that is specific for each protocol. However, these results were based on a balanced mixture of concentrated synthetic miRNAs that may not be fully representative of biological samples such as biofluids, where miRNA concentrations vary broadly and sequence complexity is lower. To identify how measurements in real biofluid samples are influenced by bias, the absolute abundance of 19 miRNAs was quantified by RT-qPCR in plasma (Supplemental Figure S4) and was correlated to measured RNA-Seq values (Figure 1F). All protocols showed positive correlations with R2 values between 0.53 (Norgen) and 0.88 (SMARTer), although precision for individual miRNAs was often low. In agreement with miRXplore data, Lexogen and Norgen performed worst in this metric. The analysis demonstrates that globally, across-miRNA correlations are relatively preserved in RNA-Seq output from biofluids, that is, highly abundant miRNAs give high-count values and vice versa. However, values for individual miRNAs are biased and cannot be readily transformed to absolute abundance, making between-miRNA comparisons difficult. It was therefore explored whether protocol-specific biases learnt from the synthetic sample (miRXplore) could be leveraged to correct bias in RNA-Seq data from plasma post hoc, as was shown by Baroin-Tourancheau et al.[25] Indeed, computational correction increased both correlation of RNA-Seq values with known absolute concentrations (Figure 1F), as well as inter-protocol correlation (Figure 1G). These results suggest that protocol-specific biases are preserved (at least to a degree) even between vastly different samples such as plasma and miRXplore. Once learned on the sample with a known ground truth, they can be leveraged to both, improving precision of the RNA-Seq values and agreement between protocols, potentially facilitating comparisons across studies.

An important decision that researchers face when designing small RNA-Seq experiments is the targeted sequencing depth, which affects the detection rates and cost efficiency of the experiment. The required sequencing depth is influenced by the ability of protocol to capture molecules of interest and by the proportion of artifact reads. To assess capture efficiency, mapping statistics for each protocol were evaluated (Figure 2A). Note that adaptor-dimers were removed during library preparation in this study and therefore were not mapped (Supplemental Figure S1). Whereas the mapping statistics were comparable between protocols with miRXplore, the results revealed substantial differences with plasma samples. The most striking was the low

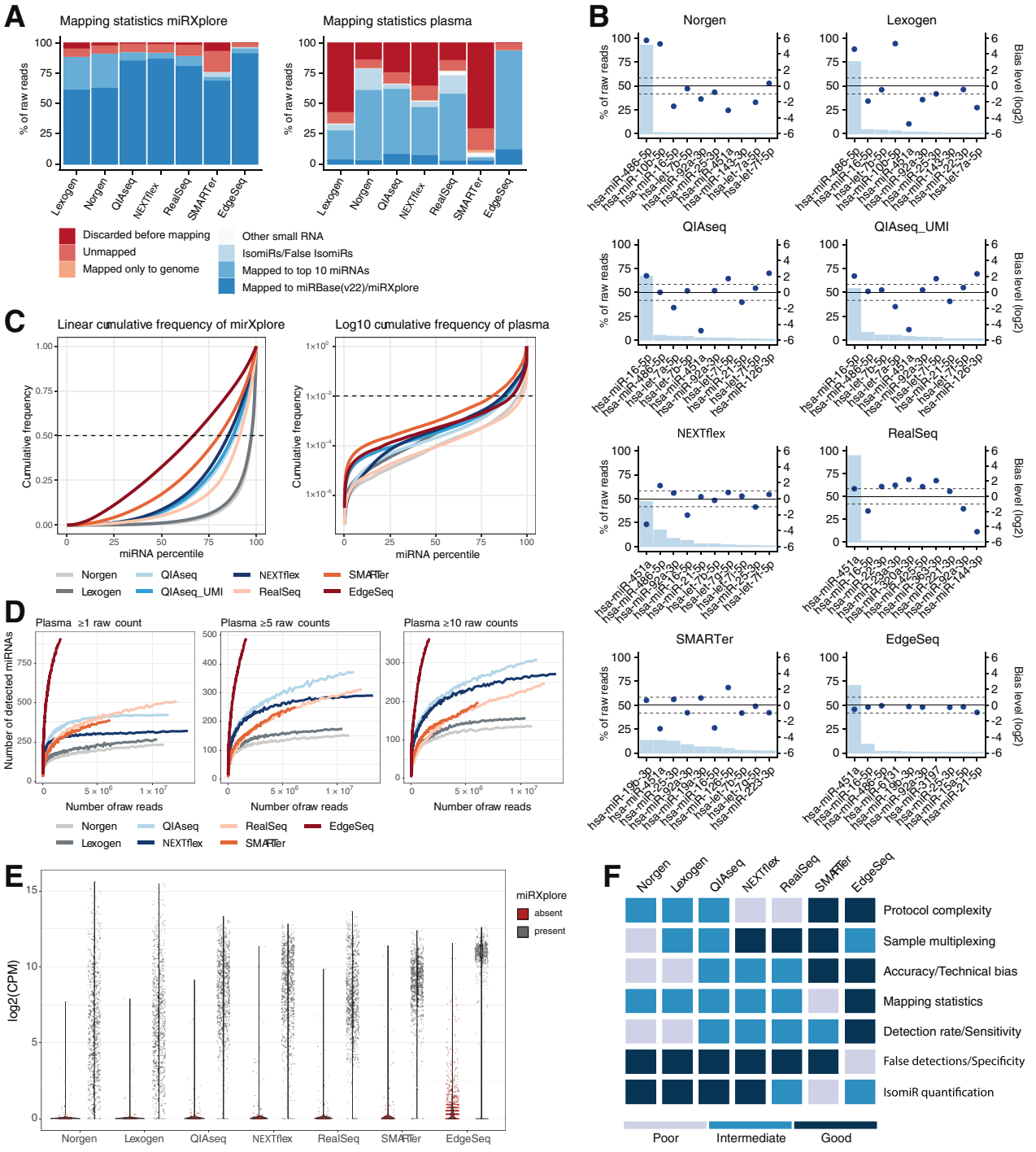

Figure 2 Mapping statistics, sensitivity, false positivity, and performance evaluation. **A:** Mapping statistics for miRXplore and plasma samples. **B:** Top 10 most abundant miRNAs in plasma. Bars (left *y* axis) show fraction of raw reads in plasma, and dots (right *y* axis) show $\log_2$-fold level of technical bias in the miRXplore sample. **Dashed lines** mark twofold deviation from the expected value. **C:** Cumulative frequency of miRXplore and plasma samples in linear and log scale, respectively. **Dashed lines** indicate cumulative frequency of 50% and 1%. **D:** Dependency of the number of detected miRNAs in plasma on sequencing depth and various detection thresholds (1, 5, 10 raw reads). **E:** Violin plots showing measured level of true- and false-positive miRNAs measured in the miRXplore samples. **F:** Final evaluation metrics; for details see Supplemental Table S3.

mapping rate to miRNAs for SMARTer, which was mostly due to inappropriate read length (Supplemental Figure S5). By contrast, targeted approach EdgeSeq showed highest mapping rate of 95%. Importantly, with all protocols the majority of miRNA-mapping reads was consumed by the few highest-ranking miRNAs (Figure 2, B and C). Because this may reflect true miRNA abundance, but also may be a consequence of bias, values of the 10 most abundant

miRNAs in plasma were plotted together with their corresponding level of bias measured in miRXplore (Figure 2B). In each protocol, except SMARTer and NEXTflex, there was always a single miRNA consuming more than 50% of all mapped reads. The rank and identity of the top 10 miRNAs differed between protocols. Although some miRNAs, such as erythrocyte-specific miR-451 and miR-16, ranked among the highest with all protocols (in agreement with their true abundance) (Supplemental Figure S4), other miRNAs, such as miR-10b with Norgen and Lexogen, appeared to be strongly overestimated (up to 64×) due to bias. To assess allocation of sequencing reads on the full miRNA spectrum, curves of cumulative frequencies were examined (Figure 2C). The fast increase in the cumulative frequency indicates that even low-ranking miRNAs contribute significantly to the total counts. In miRXplore, the number of miRNAs at cumulative frequency of 50% (CF50) would ideally be around 481 (half of 962 miRNAs consume half of the reads; lower values are better). In agreement with the percentage of unbiased miRNAs, EdgeSeq and SMARTer showed best performance, whereas Norgen and Lexogen were worst in this metric. In plasma, the shape of the ideal curve cannot be known; however, it is vastly apparent that the majority of the reads are consumed by a few miRNAs. Together, these results show that the highly skewed miRNA distribution in plasma is caused by natural miRNA abundance, as well as artificial, protocol-specific biases, and both factors need to be considered to select optimal sequencing strategy.

Considering the strong quantification bias of some miRNAs, binary evaluation of miRNA profiles (present/absent) may represent an alternative, more robust approach to identify candidate biomarkers. To characterize variables influencing such analysis, miRNA detection rates at various sequencing depths and count thresholds for each protocol were examined (Figure 2D). Although most of the untargeted protocols approached saturation at 5 million reads, SMARTer and RealSeq further benefited from increased depth. EdgeSeq, QIAseq, and NEXTflex detected highest number of miRNAs, whereas Lexogen and Norgen detected the fewest. Relative differences between protocols were most pronounced with higher detection thresholds and were retained at various sequencing depths. Interestingly, EdgeSeq detected up to hundreds more miRNAs than any other protocol (Figure 2D). This can be attributed to EdgeSeq's high mapping rate (Figure 2A), but it can be also a consequence of lower specificity of hybridization probes.[27] To investigate this, measured values for present human miRNAs (ie, true positives) versus absent human miRNAs (ie, false positives) in miRXplore sample were plotted (Figure 2E). Indeed, EdgeSeq showed a higher false-positive rate and higher false signal intensities compared with other protocols, suggesting that its higher detection rate in plasma may be partly due to false positivity. Further, enrichment of false-positive miRNAs from miRXplore was assessed between miRNAs that were uniquely detected by each protocol in plasma (ie, miRNAs not detected by any other protocol). This was indeed the case for EdgeSeq, but not for other protocols at all examined detection thresholds. Sequence similarity analysis revealed that false-positive miRNAs detected by EdgeSeq were only modestly similar to true positive miRNAs (Supplemental Figure S6), suggesting that false-positive detections may result from incomplete digestion of unbound capture probes, in addition to cross-hybridization. Because miRNA analysis on the level of miRNA variants, isomiRs, is getting more attention in miRNA biomarker studies,[28–30] the levels of false isomiR detection were evaluated using the miRXplore sample. SMARTer generated the most false isomiRs: over 4% of all raw reads, compared with <0.4% for other sequencing-based protocols (Figure 2A). Detailed analysis revealed protocol-specific bias between 3′ and 5′ isomiRs, as well as base preferences (Supplemental Figure S7). Whereas some were expected (dominance of 3′ isomiRs with added adenines in SMARTer), the prevalence of 3′ isomiRs in EdgeSeq or preference for 3′ thymine addition in RealSeq were unexpected. This raises questions on the reliability of isomiRs quantification and warrants careful validation of such data. To summarize, we observed large differences in miRNA detection rate between protocols as well as varying contribution of false positives. Although EdgeSeq captured the highest number of miRNAs, it suffered from highest false-positive rate, particularly for miRNAs with low values. Overall, the results suggest caution about spurious detections and highlight the need for data validation by independent technology.

Discussion

In this study, we compared the performance of all currently available technical approaches for RNA-Seq–based miRNA analysis in biofluids using a complex set of parameters, including, not only data-driven characteristics, but also practical features such as protocol complexity or level of multiplexing (Supplemental Table S3). There was no protocol that would stand out as the best across all metrics (Figure 2F). In agreement with other studies,[9,10,31] we show that data generated by ligation-free protocols were the least biased, suggesting they may be preferable when quantification of true miRNA abundance is of interest. Particularly, EdgeSeq outperformed others in accuracy, but also in the high mapping and detection rate. Other advantages of this platform are automatization minimizing hands-on time and the possibility to analyze crude biofluid samples. Although here we analyzed isolated RNA for consistency reasons, Godoy et al[31] found no major differences between crude and isolated samples. EdgeSeq disadvantages are represented by the higher costs of analysis, the possibility to quantify only predefined sets of miRNAs, and the lower specificity, which is in agreement with the results of Godoy et al.[31] SMARTer was the second most accurate and the least laborious method

from the wet-lab perspective. However, its performance was negatively affected by the lowest mapping rate to miRNAs and highest production of artifact reads and false isomiRs, in accordance with previous studies.[9,10] On the other hand, SMARTer may be well-suited for simultaneous analysis of various classes of small RNAs in a single experiment. Surprisingly, the most recent bias-mitigating approach RealSeq showed accuracy levels similar to NEXTflex and QIAseq, in contrast to results of Barberán-Soler et al,[7] which reported superior accuracy of over 70%. Here, we found that the circularization approach is not exempt from bias. Considering that RealSeq employs two adaptor ligation steps (one inter- and one intramolecular); our result seems to be in line with observations that ligation is the most prominent source of bias.[5,32] Random adaptors used in NEXTflex represented the third approach in our comparison developed for reduction of ligation bias. In agreement with recent studies,[11,12] NEXTflex showed good-to-average performance in most of the tested parameters and may be therefore recommended for routine applications in various experimental settings. Lastly, three representatives of traditional ligation-based methods were tested (Lexogen, Norgen, and QIAseq). As expected, Lexogen and Norgen did not perform well in the majority of tested parameters, which is in agreement with the recent literature.[11,16] Strong ligation bias leads to misbalanced miRNA profiles, low coverage of the majority of the miRNAs, and lower detection rates, and therefore to the need for higher sequencing depth. Surprisingly, QIAseq that also employs ligation of two defined adaptors ranked together with NEXTflex among the best in most metrics. Data show that this is not due to the usage of UMIs, and because details of the protocol are proprietary, it can be speculated if proper optimization or other bias-mitigating measures are responsible for improved results of QIAseq over Lexogen and Norgen.

Beside the protocol comparison, the data identified several opportunities for improvement of small RNA-Seq analysis in biofluids. First, we documented highly misbalanced miRNA profiles in plasma, where a few highly abundant miRNAs consumed most reads (partly due to biological, but also due to technical, reasons). A new generation of library preparation protocols would therefore benefit from blocking or depleting highly abundant miRNAs such as miR-451 and miR-16. A similar approach was demonstrated on tRNA-halves and improved miRNA detection in serum.[33] It might be assumed that targeted depletion of selected miRNAs might change the evaluation metrics presented in this study and lead to different conclusions. Secondly, it was demonstrated that bias can be learned on synthetic samples with known ground truth and subsequently transferred to improve precision and between-protocol correlation of values in a real biofluid sample. Development of advanced computational correction models allowing for complex cross-study comparisons would therefore dramatically increase the utility of publicly available datasets and lead to increase of current knowledge on miRNA profiles in different pathological states. Lastly, contrary to recent reports,[8,9] our results suggest that UMIs are superfluous for miRNA quantification and can even lead to serious quantification errors if designed improperly (eg, with insufficient length). However, the presented data are based on a balanced synthetic template, and sample- and protocol-specific factors may pronounce UMIs importance, which needs to be addressed in future studies. For now, we advocate for the developments primarily focused on overcoming ligation bias and improving sensitivity.

Conclusions

The presented study provides comprehensive comparison of all current approaches for high-throughput RNA-Seq–based analysis of small RNAs. The data confirmed a large bias in the data generated by traditional two-step ligation methods and highlighted superiority of the methods using capture probes (EdgeSeq) or randomized adaptors (NEXTflex). The data documented some drawbacks that still exist and offered opportunities for further development and improvement of existing workflows. Overall, this study serves as a point of reference for an informed selection of a small RNA-Seq method and provides a framework for future development of library preparation protocols and computational methods.

Acknowledgments

We thank the vendors and local distributors of small RNA-Seq preparation protocols used in the study for the generous free or discounted sample kits.

Author Contributions

P.A. and L.V. designed the study; P.A. prepared standardized material and small RNA libraries; S.B. processed data and performed the majority of the analyses, with contributions from P.A.; E.R. performed RT-qPCR measurements; P.A. and L.V. supervised data analysis; P.A. and S.B. prepared figures and wrote the manuscript; L.V. and M.K. reviewed the manuscript and wrote the final version. All authors reviewed the manuscript.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2021.12.006*.

References

1. Anfossi S, Babayan A, Pantel K, Calin GA: Clinical utility of circulating non-coding RNAs — an update. Nat Rev Clin Oncol 2018, 15:541–563
2. Chugh P, Dittmer DP: Potential pitfalls in microRNA profiling. Wiley Interdiscip Rev RNA 2012, 3:601–616

3. Valihrach L, Androvic P, Kubista M: Circulating miRNA analysis for cancer diagnostics and therapy. Mol Aspects Med 2020, 72:100825
4. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R: Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res 2011, 39:e141
5. Raabe CA, Tang T, Brosius J, Rozhdestvensky TS: Biases in small RNA deep sequencing data. Nucleic Acids Res 2014, 42:1414−1426
6. Baran-Gale J, Kurtz CL, Erdos MR, Sison C, Young A, Fannin EE, Chines PS, Sethupathy P: Addressing bias in small RNA library preparation for sequencing: a new protocol recovers microRNAs that evade capture by current methods. Front Genet 2015, 6:352
7. Barberán-Soler S, Vo JM, Hogans RE, Dallas A, Johnston BH, Kazakov SA: Decreasing miRNA sequencing bias using a single adapter and circularization approach. Genome Biol 2018, 19:105
8. Fu Y, Wu P-H, Beane T, Zamore PD, Weng Z: Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics 2018, 19:531
9. Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, Brandon NJ, Cross AJ, Jaffe AE, Weinberger DR, Shin JH: Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. BMC Genomics 2019, 20:513
10. Dard-Dascot C, Naquin D, d'Aubenton-Carafa Y, Alix K, Thermes C, van Dijk E: Systematic comparison of small RNA library preparation protocols for next-generation sequencing. BMC Genomics 2018, 19:118
11. Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, De Hoff PL, Tanriverdi K, Courtright A, Lu S, Khoory J, Rubio R, Baxter D, Driedonks TAP, Buermans HPJ, Nolte-'t Hoen ENM, Jiang H, Wang K, Ghiran I, Wang YE, Van Keuren-Jensen K, Freedman JE, Woodruff PG, Laurent LC, Erle DJ, Galas DJ, Tewari M: Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nat Biotechnol 2018, 36:746−757
12. Yeri A, Courtright A, Danielson K, Hutchins E, Alsop E, Carlson E, Hsieh M, Ziegler O, Das A, Shah RV, Rozowsky J, Das S, Van Keuren-Jensen K: Evaluation of commercially available small RNA-Seq library preparation kits using low input RNA. BMC Genomics 2018, 19:331
13. Coenen-Stass AML, Magen I, Brooks T, Ben-Dov IZ, Greensmith L, Hornstein E, Fratta P: Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. RNA Biol 2018, 15:1133−1145
14. Wong RKY, MacMahon M, Woodside JV, Simpson DA: A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. BMC Genomics 2019, 20:446
15. Herbert ZT, Thimmapuram J, Xie S, Kershner JP, Kolling FW, Ringelberg CS, Leclerc A, Alekseyev YO, Fan J, Podnar JW, Stevenson HS, Sommerville G, Gupta S, Berkeley M, Koeman J, Perera A, Scott AR, Grenier JK, Malik J, Ashton JM, Pivarski KL, Wang X, Kuffel G, Mesa TE, Smith AT, Shen J, Takata Y, Volkert TL, Love JA, Zhang Y, Wang J, Xuei X, Adams M, Levine SS: Multisite evaluation of next-generation methods for small RNA quantification. J Biomol Tech 2020, 31:47−56
16. Heinicke F, Zhong X, Zucknick M, Breidenbach J, Sundaram AYM, Flåm S T, Leithaug M, Dalland M, Farmer A, Henderson JM, Hussong MA, Moll P, Nguyen L, McNulty A, Shaffer JM, Shore S, Yip HK, Vitkovska J, Rayner S, Lie BA, Gilfillan GD: Systematic assessment of commercially available low-input miRNA library preparation kits. RNA Biol 2020, 17:75−86
17. Androvic P, Romanyuk N, Urdzikova-Machova L, Rohlova E, Kubista M, Valihrach L: Two-tailed RT-qPCR panel for quality control of circulating microRNA studies. Sci Rep 2019, 9:4255
18. Androvic P, Valihrach L, Elling J, Sjoback R, Kubista M: Two-tailed RT-qPCR: a novel method for highly accurate miRNA quantification. Nucleic Acids Res 2017, 45:e144
19. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011, 17:10
20. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10:R25
21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15−21
22. Smith T, Heger A, Sudbery I: UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 2017, 27:491−499
23. Kozomara A, Birgaoanu M, Griffiths-Jones S: MiRBase: from microRNA sequences to function. Nucleic Acids Res 2019, 47: D155−D162
24. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: ViennaRNA package 2.0. Algorithms Mol Biol 2011, 6:26
25. Baroin-Tourancheau A, Jaszczyszyn Y, Benigni X, Amar L: Evaluating and correcting inherent bias of microRNA expression in Illumina sequencing analysis. Front Mol Biosci 2019, 6:17
26. Fuchs RT, Sun Z, Zhuang F, Robb GB: Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS One 2015, 10:e0126049
27. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, Andrade PD, Demayo M, Dennis L, Derveaux S, Feng Y, Fulmer-smentek S, Gerstmayer B, Gouffon J, Grimley C, Lader E, Lee KY, Luo S, Mouritzen P, Narayanan A, Patel S, Peiffer S, Rüberg S, Schroth G, Schuster D, Shaffer JM, Shelton EJ, Silveria S, Ulmanella U, Veeramachaneni V, Staedtler F, Peters T, Guettouche T, Vandesompele J: Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods 2014, 11:809−815
28. Koppers-Lalic D, Hackenberg M, De Menezes R, Misovic B, Wachalska M, Geldof A, Zini N, de Reijke T, Wurdinger T, Vis A, van Moorselaar J, Pegtel M, Bijnsdorp I: Non-invasive prostate cancer detection by measuring miRNA variants (isomiRs) in urine extracellular vesicles. Oncotarget 2016, 7:22566−22578
29. Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I: Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res 2015, 43:9158−9175
30. Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I: Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. Nucleic Acids Res 2017, 45:2973−2985
31. Godoy PM, Barczak AJ, DeHoff P, Srinivasan S, Etheridge A, Galas D, Das S, Erle DJ, Laurent LC: Comparison of reproducibility, accuracy, sensitivity, and specificity of miRNA quantification platforms. Cell Rep 2019, 29:4212−4222.e5
32. Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JTG, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T: RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA 2011, 17:1697−1712
33. Van Goethem A, Yigit N, Everaert C, Moreno-Smith M, Mus LM, Barbieri E, Speleman F, Mestdagh P, Shohet J, Van Maerken T, Vandesompele J: Depletion of tRNA-halves enables effective small RNA sequencing of low-input murine serum samples. Sci Rep 2016, 6: 37876