*12140–12151* *Nucleic Acids Research, 2017, Vol. 45, No. 21* *Published online 24 October 2017*
*doi: 10.1093/nar/gkx999*

# sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline

**Xiaogang Wu[1], Taek-Kyun Kim[1], David Baxter[1], Kelsey Scherler[1], Aaron Gordon[1], Olivia Fong[2], Alton Etheridge[2], David J. Galas[2] and Kai Wang[1,*]**

[1]Institute for Systems Biology, Seattle, WA 98109, USA and [2]Pacific Northwest Research Institute, Seattle, WA 98122, USA

Received February 16, 2017; Revised October 06, 2017; Editorial Decision October 10, 2017; Accepted October 11, 2017

## ABSTRACT

**Although many tools have been developed to analyze small RNA sequencing (sRNA-Seq) data, it remains challenging to accurately analyze the small RNA population, mainly due to multiple sequence ID assignment caused by short read length. Additional issues in small RNA analysis include low consistency of microRNA (miRNA) measurement results across different platforms, miRNA mapping associated with miRNA sequence variation (isomiR) and RNA editing, and the origin of those unmapped reads after screening against all endogenous reference sequence databases. To address these issues, we built a comprehensive and customizable sRNA-Seq data analysis pipeline—*sRNAnalyzer*, which enables: (i) comprehensive miRNA profiling strategies to better handle isomiRs and summarization based on each nucleotide position to detect potential SNPs in miRNAs, (ii) different sequence mapping result assignment approaches to simulate results from microarray/qRT-PCR platforms and a local probabilistic model to assign mapping results to the most-likely IDs, (iii) comprehensive ribosomal RNA filtering for accurate mapping of exogenous RNAs and summarization based on taxonomy annotation. We evaluated our pipeline on both artificial samples (including synthetic miRNA and *Escherichia coli* cultures) and biological samples (human tissue and plasma). sRNAnalyzer is implemented in Perl and available at: http://srnanalyzer.systemsbiology.net/.**

## INTRODUCTION

The function and application of small non-coding RNAs (ncRNA), especially microRNA (miRNA) has gained significant interest in recent years (1). miRNAs are 17–23 nt-long evolutionarily conserved regulatory RNAs. Numerous reports have shown the existence of RNA molecules, including miRNA, in various bodily fluids (2,3). These extracellular RNAs are either encapsulated in lipid vesicles or complexed with proteins to prevent RNase degradation. The spectrum of extracellular miRNA in various body fluids correlates with physiopathological conditions, and it has been suggested that the concentration changes of specific extracellular miRNAs can be used as biomarkers for disease diagnosis (3–5). Therefore, the ability to accurately profile and quantify the small RNA content in biological samples is critical for further development of miRNA-based applications. Although many experimental and computational tools have been developed (6–9), it is still a challenge to accurately and comprehensively analyze the small RNA population (10). For miRNA, the short sequence length, high sequence similarity among family members and sequence length variation (isomiR) all contribute to the difficulty of miRNA quantification. In addition, recent studies have suggested that a fraction of the RNA in circulation is derived from exogenous species including bacteria, fungi and plants (10–14), further complicating unambiguous assignment of the origin of many RNA species.

Due to the wealth of information provided and decreased cost, the next generation sequencing (NGS) platform has gradually become one of the most common tools for studying small RNAs (15). The NGS platform can resolve closely related sequences, is not affected by sequence length variation, does not require prior knowledge of the sequence, may provide information to identify new miRNA sequences and allows the profiling of exogenous RNAs in the sample. However, most of the tools (summarized in Supplementary Table S1) for small RNA sequencing (sRNA-Seq) data analysis deliver poor sequence mapping specificity.

To address some of the small RNA analysis problems, particularly for miRNA, we have built a comprehensive and customizable pipeline—sRNAnalyzer, based on the framework published earlier (4,10–11). An installation package for the pipeline and all the reference database indexes used in the pipeline can be downloaded from http://srnanalyzer.systemsbiology.net/. We also built a web interface for the pipeline, which can be accessed at the same website, mainly

*To whom correspondence should be addressed. Tel: +1 206 732 1336; Fax: +1 206 732 1299; Email: kwang@systemsbiology.org

for human small RNA profiling. In this paper, we introduce the overall workflow with a description of key features, followed by performance evaluation using a synthetic miRNA panel and biological samples.

## MATERIALS AND METHODS

Similar to all other tools for NGS-based small RNA profiling, the tool described here can be grouped into three functional modules (shown in Figure 1), i.e. data pre-processing, sequence mapping/alignment and result summarization. The sequence mapping/alignment module is designed for two major categories, one aimed at endogenous RNA profiling and the other for identifying RNA sequences derived from exogenous species. With this modular design, sRNAnalyzer allows for rapid modifications of each module, for example, adding/removing a reference sequence database, or changing the mapping order of databases used in the pipeline.

### Data pre-processing module

In the data pre-processing module, the pipeline trims the adapter sequences and removes empty reads (adapter dimer). Adapter trimming is especially important for sRNA-Seq data analysis, since most of the sequence reads are short and may contain part or all of the adapter sequences. If the adapter sequences are not completely removed, mapping accuracy will be significantly affected. The module also assesses the overall sequence quality and removes low quality and low complexity reads, such as reads containing simple repeats—homopolymer, di- or trinucleotide repeats.

*Processes to trim adapters and remove low quality sequences.* To remove adapter sequences, the pipeline adapted Cutadapt (16), to remove both 3′-end and 5′-end adapter sequences. After adapter trimming, sRNAnalyzer removes empty reads by utilizing index sequences (i.e. a 6-mer barcode embedded in the primer sequence) used for individual samples during the library construction process. If no index/barcode information is provided, the pipeline can retrieve index information directly from raw sequence files by calculating the top frequency of potential index sequences (Supplementary File 1—Method details). sRNAnalyzer handles degenerate bases used in adapters, such as those in kits from Bioo (Bioo Scientific, Austin, TX, USA), through configuring different settings/parameters in Cutadapt. The pipeline also provides options to identify and remove other contaminant sequences commonly encountered in small RNA libraries. To reduce adapter dimer, the Illumina TruSeq small RNA library construction kit uses stop oligo in the library preparation process. The stop oligo sequences (GAATTCCACCACGTTCCCGTGG) will be incorporated into the library especially when the input RNA concentration is low, such as libraries from various body fluid samples. The level of stop oligo in the library can be used as one of the criteria for library quality assessment. Our pipeline removes stop oligo sequences based on template sequences observed from experiments from synthetic miRNA samples. For quality controls, the pipeline uses Prinseq (17), to remove low-quality reads.

*Generating unique sequence file.* To accelerate sequence alignment, the pipeline uses fastx_collapser (http://hannonlab.cshl.edu/fastx_toolkit/) to generate a file where identical reads are collapsed together. The new header of each unique FASTA sequence in the new file contains a unique sequence ID, followed by the number of reads that have the same sequence. When calculating sequence abundance, the read count for each unique sequence is added together.

### Sequence mapping and alignment module

The main function of the read sequence alignment module is to determine the sequence identity by mapping the reads against various sequence databases. Like many other sRNA-Seq data analysis tools, sRNAnalyzer uses Bowtie (18) alignment results to determine the identity of sequence reads.

*Sequential mapping strategy.* The pipeline follows the 'map and remove' strategy adapted in the original report (11) to sequentially map reads against various databases as described (Supplementary Figure S1). The order of databases depends on the research focus and can be changed by the user. The sequence mapping is based on a progressive alignment strategy—maximum mismatch allowance progressively increases from 0 to 2 for mapping against sequence databases. The error tolerance can also be adjusted by the user.

The sequence reads can map to either the positive or negative strand of the reference sequences. In the default setting, the pipeline only counts reads mapped to the positive (forward) strand for various RNA sequence databases (e.g. miRNA, piRNA, snoRNA, lncRNA, RefSeq, ncRNA etc.), but counts reads mapped to both strands (±) for DNA sequence databases (e.g. human genome, human microbiome DNA databases and bacterial genomes). For reads mapped to the reverse (−) strand of sequences in RNA-specific databases, the results will not be included as expression values, but their mapped locations will still be reported since these reads may have regulatory functions.

*miRNA mapping.* As illustrated in Supplementary Figure S2, our pipeline directly maps read sequences against miRNA precursor sequences (annotated with mature miRNA locations) by applying full alignment in Bowtie, rather than using Bowtie seed alignment option as in sRNAbench (19). Using seed alignment (only considering matches on a seed region, default setting is 20 nt in sRNAbench) cannot control mismatch tolerance outside the seed region, and will mix RNA editing events and SNPs with sequencing errors. The latest sRNAbench paper (19) also mentioned the deficiency of using seed alignment. In our pipeline, we generated composite sequence header containing precursor sequence ID, and mature miRNA names and coordinates during the construction of miRNA reference database. A mapped read will then be assigned to either one of the corresponding mature miRNAs (either at 5′ end (-5p) or 3′ end (-3p) of the precursor sequence).

With this approach, some miRNA IDs generated through sRNAnalyzer do not exist in miRBase (20). For example, hsa-miR-101–1-3p and hsa-miR-101–2-3p share the

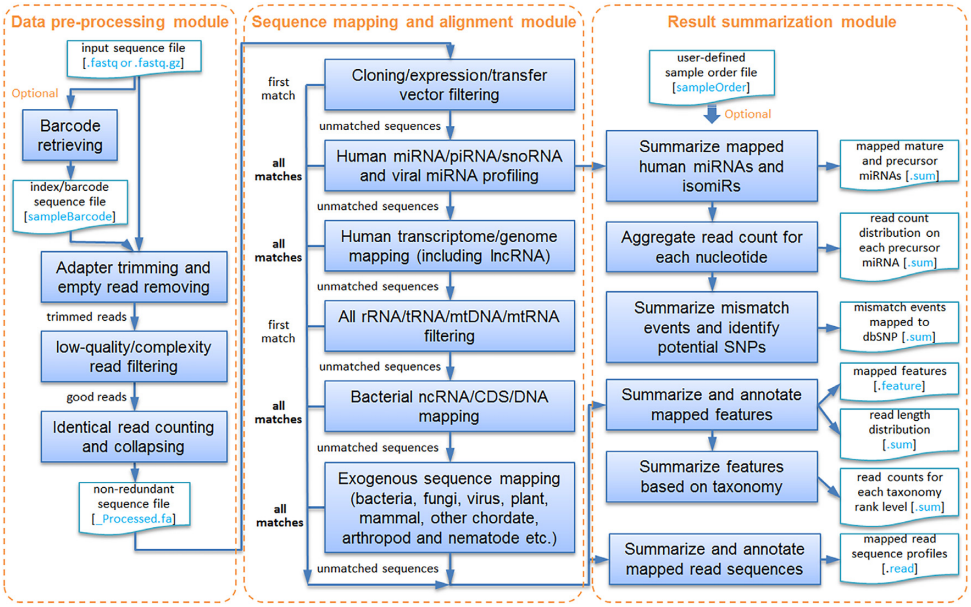


**Figure 1.** Main framework of sRNAnalyzer. The pipeline can be divided into three functional modules which are separated by dotted lines. The data format for each process is indicated in square brackets.

same mature sequence (hsa-miR-101–3p in miRBase 21) but contain different sequences beyond the mature sequence. This strategy offers the advantage of obtaining all length variations—isomiR associated with miRNA. With mismatch allowance, we can also detect potential RNA editing events and single nucleotide polymorphisms (SNPs) associated with miRNA. We demonstrated this in a case study using colorectal cancer (CRC) sequence dataset. To verify potential SNPs or RNA editing issues, the pipeline will further search each mismatch event, including both mismatch types and coordinates on genome, against dbSNP (21).

To address the concern about validity of miRNA repertoires recently raised (22–24), the pipeline also provides MirGeneDB (24) as an option for human miRNA profiling with better annotations. The pipeline has modular designs on reference database mapping, which allows for adding/removing a customized reference sequence database, or changing the mapping order of databases very conveniently (through a markup configuration file). To summarize miRNA mapping results, sRNAnalyzer uses two read summarization strategies: single assignment and multiple assignment approaches first introduced in sRNAbench (19). Under the single assignment approach, each read is counted only once even when it maps to multiple reference sequences. In the multiple assignment approach, all mapped reference sequences are counted.

*Single assignment for miRNA mapping.* One of the difficulties associated with miRNA profiling or measurement is the extreme sequence similarity of some sequences, especially for mature miRNAs among family members (sequences can sometimes be identical). Under traditional approaches, identical reads from different samples may be assigned to different IDs, since the Bowtie aligner randomly assigns all the best matching reference sequences as the top hit. To solve this mapping issue, we adapted a local probabilistic model (LPM) to assign each mapped read to the most-likely miRNA within a sample (Supplementary Figure S3). This LPM is based on feature (sequence ID) ranking order, which is assessed by the frequency of mapped reads (a higher ranking order indicates the specific sequence ID has more mapped reads). Therefore, a miRNA mapping assignment is determined by both the alignment of individual reads and the overall mapped read count distribution among all the miRNAs within a sample.

*Weighted multiple assignment approach for miRNA mapping.* Due to the nature of nucleic acid hybridization, which forms the basis of both qRT-PCR and microarray platforms, the results from qRT-PCR and microarray are less specific than NGS-based profiling results. To generate results that are in better agreement with other miRNA measurement platforms, the user can also select a weighted multiple assignment approach. For miRNA mapping, our default parameters count a perfect match (no mismatch) as 1, one mismatch as 0.8 and two mismatches count as 0.6. Users can also change these weight factors.

*Exogenous RNA mapping—ribosomal and transfer RNA filtering.* Based on prior studies (4,10), most of the exogenous RNAs map to various ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) (11,25–26). Both rRNAs and tRNAs possess high sequence similarity among different species, which creates uncertainty about the origin of exogenous RNA and a significant mapping problem due to multiple matches. In some cases, a single sequence read can map to several thousand sequences in the database. For a more accurate mapping result for exogenous RNAs, the pipeline offers an option to filter out microbial tRNA and rRNA sequences prior to exogenous sequence database mapping.

We tested the use of tRNA and rRNA for exogenous species mapping using bacteria RNA dataset and found re-

moving rRNA and tRNA sequences increased the accuracy for species assignment. However, this needs to be further tested with additional datasets. Our pipeline does allow the user to bypass the rRNA and tRNA filtering step so that user can test the mapping accuracy with or without rRNA/tRNA filtering. Even though we 'filter' out the tRNA and rRNA reads, the current pipeline still provides tRNA and rRNA fragment mapping results. The reason we call this a filtering step is to highlight the importance of 'rRNA and tRNA filtering' in the accuracy and specificity of exogenous RNA mapping as shown in Results section.

Recently, there are many studies discovered possible regulatory roles and potential applications of tRNA or rRNA fragments (27,28), especially in plasma or serum samples (29). To make this clear, we also modified our pipeline by adding specific human tRNA from GtRNAdb (30) and human rRNA databases from RDP II (31) and SILVA (32) in the mapping steps. In addition, we provide profiles for yRNA fragments from human transcriptome mapping results by NCBI/RefSeq (33).

### Result summarization module

The result summarization module provides reports for various mapping results. The results can be summarized at different levels—from individual transcripts (such as individual miRNAs) to different phyla.

*miRNA mapping results.* The pipeline can summarize read counts for each individual mature or precursor miRNA, as well as aggregate read counts for each nucleotide across individual precursor miRNAs. Moreover, sRNAnalyzer provides mismatch counts and rates for all the possible 16 mismatch types (A|T|G|C > A|T|G|C|N) at each position. This function allows researchers to review miRNA sequence variations in the sample.

*Summarization based on taxonomy.* Besides detailed small RNA expression profiles, the mapping results can also be viewed at different taxonomy levels (including Phylum, Class, Order, Family, Genus, Species and Strain/Scientific Name). The taxonomy annotation is based on two sources: (i) descriptions in the sequence header from the original reference sequence databases, and (ii) taxonomy information from gi_taxid obtained from NCBI (mainly for bacteria and nt database from NCBI). For convenience, common names for different species are also added to the summarization results.

### Datasets used in the performance evaluation study

To assess the performance and determine various parameters used in sRNAnalyzer, we used datasets generated using both artificial and biological samples, including synthetic miRNA, bacterial culture, human tissue and plasma samples.

*Synthetic miRNA.* A total of 286 synthetic miRNAs that are commonly observed in circulation were synthesized by Integrated DNA Technologies (Coralville, IA). We randomly assigned synthetic miRNAs to each of three plates (Plate 1, Plate 2 and Plate 3) respectively. Six combinations of different concentrations from these three synthetic miRNAs plates were subjected to small RNA sequencing (GEO accession number: GSE94912), qRT-PCR (Exiqon, Woburn, MA, USA) and microarray (Exiqon, Woburn, MA, USA) analyses (Supplementary File 3—Tables for synthetic miRNA pool information and measurements). The synthetic miRNA samples are labeled with 'Platform or library kit name'_ 'Plate 1 concentration' _ 'Plate 2 concentration' _ 'Plate 3 concentration'. For example, NEB_10_1_0.1 represents the synthetic miRNA sample with NEB NEBNext® Small RNA Library Prep library kit (New England Biolabs, Ipswich, MA, USA) using 10 ng of pooled synthetic miRNAs from plate 1, 1 ng of pooled miRNAs from plate 2 and 0.1 ng of pooled miRNAs from plate 3.

*Bacterial culture.* To test the capability and accuracy of sRNAnalyzer on exogenous RNA mapping, we re-analyzed the small RNAseq data from our prior report (25) on RNA in outer membrane vesicles (OMV) derived from *Escherichia coli* K-12 substrain MG1655. The bacteria were grown in Luria-Bertani (LB) broth at 37°C. The small RNA libraries from the bacterial culture, OMV-containing supernatant, OMV-free supernatant and unused fresh LB media were generated with the Illumina TruSeq Small RNA-Seq Prep Kit (San Diego, CA, USA). In the original study, processed reads (after removing adapter sequences) were mapped directly against the *E. coli* K-12 substrain MG1655 reference genome (accession number: NC_000913.2).

*Human tissue and plasma data.* To test the performance of sRNAnalyzer on human samples, we used two public domain datasets. A human CRC tissue miRNA study (GEO accession number: GSE46622 or SRA accession number: SRP022054) (34) and our prior human CRC plasma study (SRA accession number: ERP002414) (11). These datasets allow us to compare the results from sRNAnalyzer with published results. The CRC tissue study contains the primary tumor tissues, adjacent normal colon epithelium and liver metastases from eight patients. The small RNA libraries were made with the Illumina v1.5 Small RNA-Seq Prep Kit (3′ adapter: TCGTATGCCGTCTTCTGCTTG, 5′ adapter: GTTCAGAGTTCTACAGTCCGACGATC). In the original report, the adapters were trimmed using BLAT (35) and then the processed sequences were mapped directly against the human genome (hg19) with the BWA aligner. The plasma dataset contains nine different samples, including plasma samples from three CRC patients, three Ulcerative Colitis (UC) patients and three normal controls. The small RNA libraries were generated with the Illumina v1.5 Small RNA-Seq Prep Kit (3′ adapter: ATCT CGTATGCCGTCTTCTGCTTG, 5′ adapter: GTTCAG AGTTCTACAGTCCGACGATC).

## RESULTS

### Synthetic miRNAs

One of the challenges associated with miRNA profiling/measurement is the inconsistency of results from different platforms. This makes it difficult to compare

data generated from different platforms or to use different platforms for verification of results. This is especially true for results between NGS and other platforms, since NGS does not utilize hybridization of oligonucleotides/primers for its measurement. To better align the sequencing results with qRT-PCR and microarray results, we used measurement results from different platforms based on the 286 synthetic miRNAs to evaluate mapping parameters.

The comparison results between different measurement platforms under different synthetic miRNA composition are shown in Figure 2. The concentration distribution of miRNA profiles from miRNA array showed good correlation with input miRNA concentration (Figure 2A). For example, the third sample—Array_10_1_0.1, with 10 ng, 1 ng and 0.1 ng of pooled miRNAs from plate 1, 2 and 3 respectively, has Log2-transformed hybridization intensities around 9.5, 6.5 and 3.5 on average for plate 1, 2 and 3. The results from qRT-PCR (qPCR_10_10_10) correlated with the corresponding microarray results (Array_10_10_10, and Array_10_10_1) (Figure 2B). However, the correlation between microarray and qRT-PCR results dropped significantly when lower concentrations of synthetic miRNAs were involved.

The correlation coefficients between miRNA profiling results generated from miRNA array and sRNA-Seq under different mapping assignment approaches are summarized in Figure 2C and D. The NGS platform can differentiate different compositions of synthetic miRNA, and the NGS mapping results showed good correlation with miRNA array (Figure 2E) and qRT-PCR (Figure 2F) data. In general, NGS results generated using the multiple assignment mapping strategy with some mismatch allowance provided better correlation with microarray and qRT-PCR results compared to single assignment with no mismatches allowed. The weighted approach to adjust the contribution for different levels of mismatch on the overall mapping results further enhanced the correlation between NGS profiling results with either microarray or qRT-PCR. However, the optimal mismatch allowance conditions are different between microarray and qRT-PCR. For example, multiple assignment with a maximum of two mismatches (using weight adjustment factor of one for no mismatch, 0.8 for one mismatch and 0.6 for two mismatches) showed higher correlation with microarray results, while multiple assignment using a maximum of one mismatch (using a parameter of one for no mismatch and 0.8 for one mismatch) showed higher correlation with qRT-PCR. This is consistent with the idea that the qRT-PCR based measurement is more specific than the microarray-based measurement. However, this may also depend on differences in PCR primer or microarray probe design between different vendors. Users can adjust the parameters to fit their experimental platform and condition. Nevertheless, sRNAnalyzer provides options to better align the sequencing results with other platforms, which will be useful for users to verify sRNA-Seq results with qRT-PCR or other platforms.

### Analyzing small RNA in human samples with sRNAnalyzer

The miRNA mapping can be affected by some common features associated with miRNA, including length variation (isomiR), sequence polymorphism, RNA editing and high sequence similarity among family members. To solve these issues, sRNAnalyzer uses a combination of approaches, including mapping to miRNA precursors instead of mature miRNA and different levels of error tolerance in combination with the LPM. To test the performance of sRNAnalyzer on complex biological samples, we downloaded the raw sequencing data of a CRC study from NCBI/SRA database (SRP022054). The original study used qRT-PCR to verify several differentially expressed miRNAs identified by NGS (34).

We used the sRNAnalyzer, with multiple assignment and maximum of one mismatch, to test whether we could identify those validated differentially-expressed miRNAs in CRC. After running the dataset through sRNAnalyzer, we performed a differential analysis based on the mapping results. To compare data between different samples across different experiments, the mapped read counts were normalized with Log2-transformed reads per million (RPM) reads mapped to the reference database at each step. This normalization method is most similar to percentage. The top-ranked differential miRNAs identified by sRNAnalyzer (tumor versus normal) are shown in Table 1 and include all qRT-PCR verified miRNAs. This suggests that the results from sRNAnalyzer agree with the original study (34). The fold changes from the qRT-PCR verification study matched with results from sRNAnalyzer better than the results from the original report. For example, miR-1 is the miRNA showing the most concentration difference in both qRT-PCR and sRNAnalyzer based analyses.

When one mismatch in miRNA mapping was allowed (user adjustable), the sRNAnalyzer provided a summary file containing both match and mismatch profiles (read counts) for every nucleotide across the miRNA precursor sequence. An example of summarized read count distributions for both match and mismatch events for hsa-mir-1–1 is shown in Figure 3. The mismatch rate at nucleotide position 67 (first nucleotide of the precursor sequence is 0) is lower in normal tissue samples, compared to metastatic tissues. Whether this relates to disease status remains to be determined. This function in sRNAnalyzer allows the user to identify miRNA associated RNA editing events as well as sequence polymorphisms.

To show a global view of small RNAs in these tissue samples, we further performed differential analysis on piRNA, snoRNA and lncRNA profiles generated from sRNAnalyzer with single assignment and no mismatch. From Supplementary Tables S2–4, we can see that there are several interesting small RNAs differentially expressed in tumor tissues and significantly upregulated in metastases tissues.

### Exogenous RNA mapping with sRNAnalyzer

Exogenous species-derived RNAs, especially from bacteria and fungi, have been observed in human and murine samples (11,13–14,26). Even though the source, abundance and function of exogenous RNA remains controversial (36), the ability to accurately quantify and map the origin of exogenous sequences is critical to assess the diversity and possible function of these RNAs. We incorporated an exogenous RNA mapping process in sRNAnalyzer.

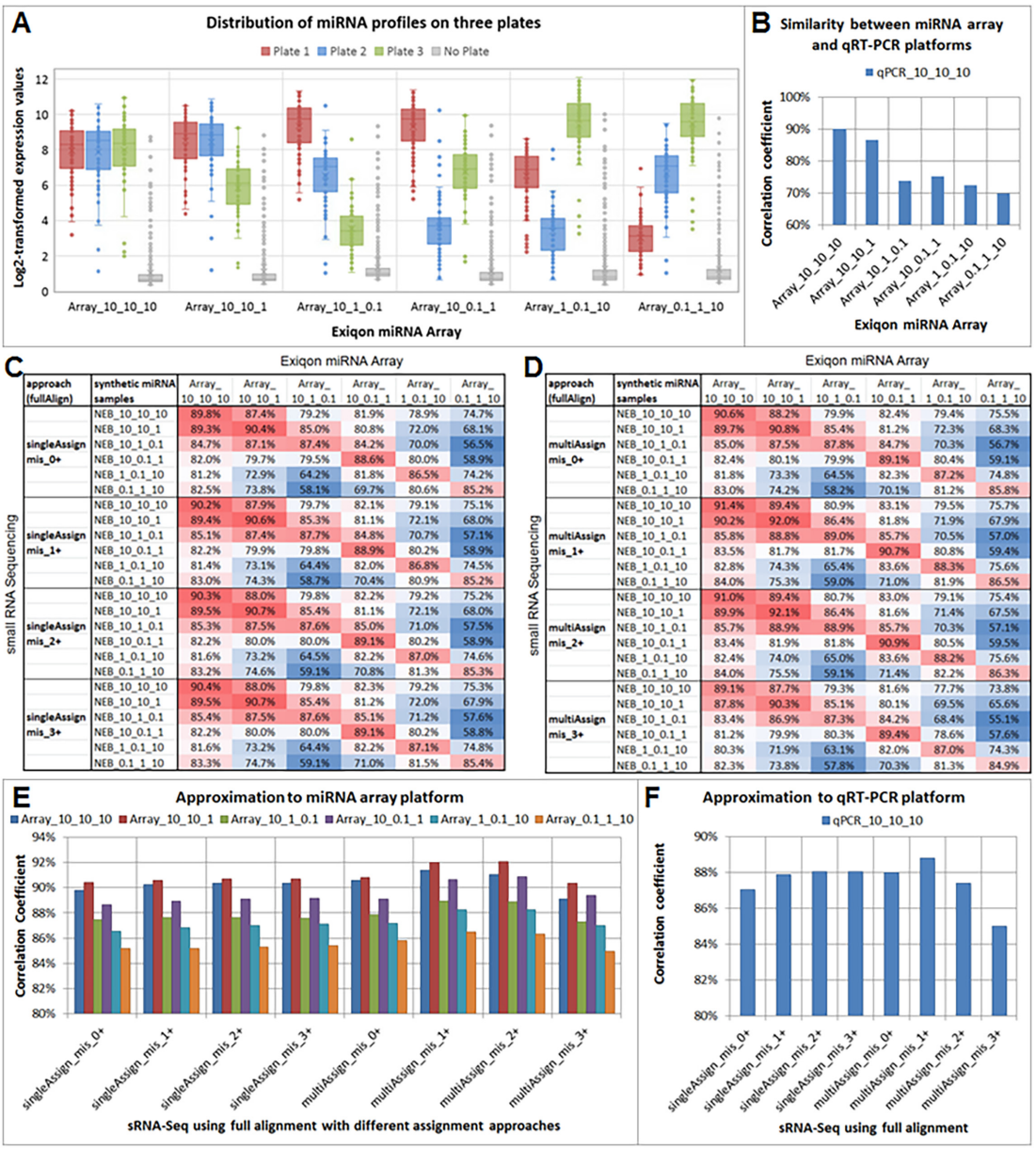


**Figure 2.** Comparison of miRNA profiles from sRNA-Seq, miRNA array and qRT-PCR platforms for synthetic miRNA samples. (**A**) Distribution of miRNA profiles from miRNA array on three plates; (**B**) Correlation between miRNA profiles from miRNA array and qRT-PCR; (**C**) Correlation between miRNA profiles from miRNA array and sRNA-Seq using single assignment approach; (**D**) Correlation between miRNA profiles from miRNA array and sRNA-Seq using multiple assignment approach; (**E**) Approximation from sRNA-Seq to miRNA array; and (**F**) Approximation from sRNA-Seq to qRT-PCR.

The conservation of rRNA and tRNA sequences among different species often causes mapping issues since a single read may assign to several hundred species. To obtain more reliable information on the origin of exogenous RNA, sRNAnalyzer includes optional rRNA and tRNA filtering steps.

To test the effect of rRNA and tRNA filtering steps on the exogenous RNA mapping accuracy, we used the small RNAseq data from a previous bacterial out membrane vesicle (OMV) study (25). In this study, *E. coli* K-12 (MG 1655) was cultured in standard LB growth medium and samples include RNA extracted from bacteria ($RNA_{intra}$), bacterial OMV particles ($RNA_{exOMV}$), OMV-depleted media ($RNA_{exOMV\text{-}free}$), cultured LB after removing bacteria ($RNA_{LB\text{-}cultured}$) and fresh LB ($RNA_{LB}$).

We first input raw read sequences into the data preprocessing module of our pipeline, including the stop oligo removing step, since these samples were made by Illumina TruSeq small RNA library kit. For the $RNA_{exOMV}$ sample, we can see that around 20% reads mapped to the stop oligo (stp) sequences have been removed (Supplementary Figure S4), which not only improved the overall mapping rate but also increased the mapping efficiency.

We then ran the trimmed/clean reads through the remaining pipeline, which allowed us to test (i) whether

**Table 1.** Differential analysis of miRNA profiles for human tissue samples from a CRC study (GEO accession number: GSE46622 or SRA accession number: SRP022054)

| miRNA | T/N_FC | T/N_pVal | M/N_FC | M/N_pVal | M/T_FC | M/T_pVal |
|---|---|---|---|---|---|---|
| **hsa-miR-1–3p** | −8.22250 | 0.00221 | −12.19746 | 0.00019 | −0.39062 | 0.25391 |
| **hsa-miR-133a-3p** | −4.76817 | 0.00251 | −8.74696 | 0.00009 | −0.59894 | 0.10591 |
| hsa-miR-194–1-3P | −4.30349 | 0.02117 | −6.83087 | 0.00324 | −0.29064 | 0.29476 |
| **hsa-miR-375** | −4.15678 | 0.01461 | −1.40532 | 0.10508 | 0.72822 | 0.22547 |
| hsa-miR-143–5p | −4.05022 | 0.00980 | −3.44819 | 0.00970 | 0.02421 | 0.41680 |
| **hsa-miR-133b** | −3.42637 | 0.00068 | −3.98052 | 0.00014 | −0.02076 | 0.35883 |
| hsa-miR-129–2-3p | −3.41944 | 0.02247 | −3.91576 | 0.00998 | −0.01681 | 0.41124 |
| hsa-miR-129–1-3p | −3.39301 | 0.01807 | −3.05203 | 0.01421 | 0.00903 | 0.42976 |
| hsa-miR-1224–5p | −3.01311 | 0.00238 | −0.49318 | 0.23264 | 1.06825 | 0.14720 |
| hsa-miR-147b | −2.74872 | 0.02911 | −4.74095 | 0.00365 | −0.26982 | 0.29074 |
| hsa-miR-124–3p | −2.70090 | 0.00639 | −1.08949 | 0.14865 | 0.35959 | 0.26224 |
| hsa-miR-490–3p | −2.68564 | 0.02562 | −3.50480 | 0.01366 | −0.05444 | 0.24816 |
| hsa-miR-215–3p | −2.64760 | 0.07514 | −5.24797 | 0.01009 | −0.44050 | 0.29178 |
| hsa-miR-133a-5p | −2.31880 | 0.00400 | −3.63812 | 0.00081 | −0.14794 | 0.13683 |
| **hsa-miR-145–3p** | −2.26530 | 0.02788 | −4.19600 | 0.00459 | −0.29520 | 0.21543 |
| **hsa-miR-96–5p** | 2.55000 | 0.01849 | 4.31795 | 0.00278 | 0.23145 | 0.25813 |
| hsa-miR-182–5p | 3.42372 | 0.03417 | 7.79898 | 0.00119 | 0.88800 | 0.17776 |
| **hsa-miR-183–5p** | 3.99934 | 0.01719 | 7.40056 | 0.00479 | 0.51921 | 0.25077 |
| hsa-miR-135a-5p | 4.57508 | 0.00792 | 7.91757 | 0.00206 | 0.45545 | 0.20791 |
| hsa-miR-122–5p | 5.42554 | 0.08238 | 34.37772 | 0.00572 | 12.48896 | 0.06981 |
| hsa-miR-31–3p | 6.04394 | 0.00683 | 2.12517 | 0.01529 | −1.00130 | 0.15600 |
| **hsa-miR-135b-5p** | 6.65842 | 0.00524 | 10.20095 | 0.00138 | 0.37638 | 0.24436 |
| **hsa-miR-31–5p** | 17.06142 | 0.00031 | 10.03755 | 0.00234 | −0.92609 | 0.17938 |

Note: N—Normal, T—Tumor, M—Metastasis, FC—Fold Change and pVal—*P*-Value. Nine miRNAs validated by qRT-PCR in the original study (34) are highlighted (bold).

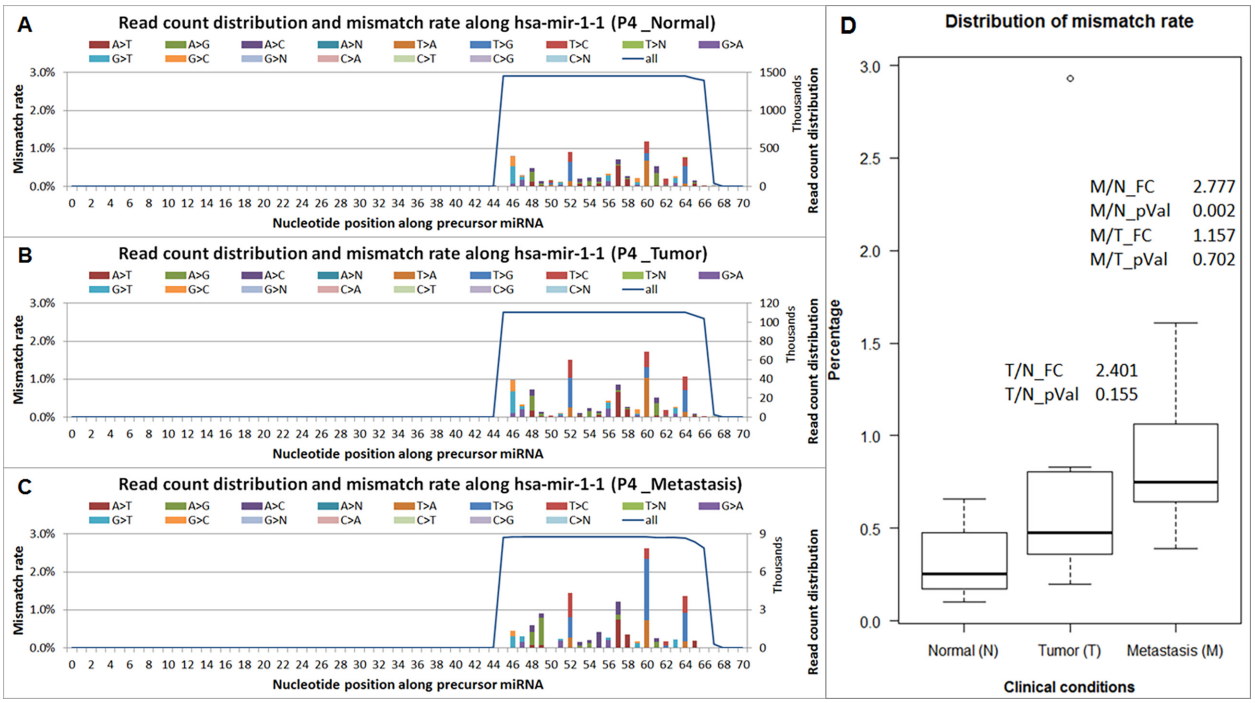


**Figure 3.** Example of summarized read count distributions for every nucleotide for both match and mismatch events across the miRNA precursor sequence (has-mir-1–1).

the parameters used for endogenous reference sequence database (mainly for human in this scenario) mapping process in sRNAnalyzer would filter out too many informative bacterial sequences and (ii) whether sRNAnalyzer can correctly identify the bacterial species. We tested various options for exogenous mapping including with/without rRNA/tRNA filtering, and with/without LPM.

Under the settings single assignment with LPM, most of the $RNA_{intra}$ reads mapped to bacteria (Figure 4A), which suggests rigorous human reference sequence mapping steps (allowing two mismatches) do not remove a significant

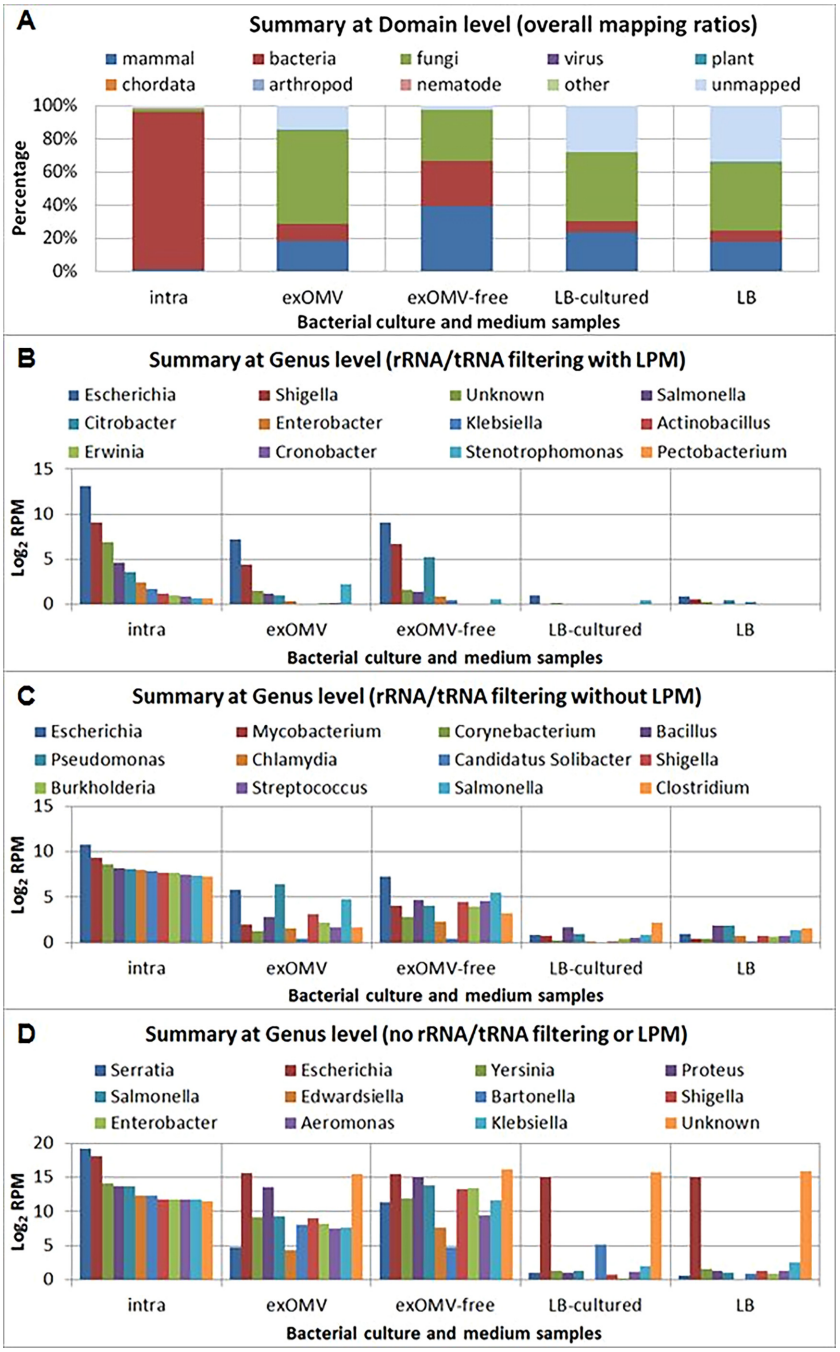


**Figure 4.** Overall mapping ratios and taxonomy summarization for *Escherichia coli* bacterial culture samples. (**A**) Overall domain mapping ratios; (**B**) Taxonomy summary at Genus level using ribosomal RNA (rRNA) filtering and applying a local probabilistic model (LPM); (**C**) Taxonomy summary at Genus level using rRNA filtering but without LPM; and (**D**) Taxonomy summary at Genus level neither using rRNA filtering nor applying LPM.

fraction of the sequence reads (over 95% of sequence reads) prior to bacterial sequence database mapping. Unlike the $RNA_{intra}$ samples, many reads from the $RNA_{exOMV}$, $RNA_{exOMV\text{-}free}$, $RNA_{LB\text{-}cultured}$ and $RNA_{LB}$ samples mapped to fungi and mammals. This is not surprising since these four samples were extracted from LB medium and the standard LB medium does contain yeast extract and tryptone (from mammalian milk).

To compare these samples under different analytical approaches, we used log2 transformed RPM mapped reads. At the genus level, *Escherichia* has the most mapped sequence reads from $RNA_{intra}$, $RNA_{exOMV}$ and $RNA_{exOMV\text{-}free}$ (Figure 4B). In results for $RNA_{intra}$, the second top-ranked genus was *Shigella*, which is evolutionarily close to *Escherichia*, followed by *Salmonella*, which shares 18% of its genome with *E. coli* substrain MG1655. It should be noted that the $RNA_{exOMV}$ and $RNA_{exOMV\text{-}free}$ samples were

more similar to the $RNA_{intra}$ sample at genus (Figure 4B), species (Supplementary Figure S5) and strain levels (Supplementary Figure S6), even though they are more similar to $RNA_{LB\text{-}cultured}$ and $RNA_{LB}$ samples at the domain level (Figure 4A). These results indicated that using default setting, sRNAnalyzer (rRNA/tRNA filtering with LPM) can successfully identify specific bacterial species from samples.

*Escherichia* is still the top ranking genus for $RNA_{intra}$ and $RNA_{exOMV\text{-}free}$ samples (Figure 4C), but not for the $RNA_{exOMV}$, when LPM was not used. This suggests that exogenous reference sequence mapping without LPM has less specificity even if rRNA and tRNA filtering step is applied. Without rRNA and tRNA filtering, *Serratia* was incorrectly identified as the top ranking genus for $RNA_{intra}$ (Figure 4D). This suggests that the rRNA/tRNA filter step can improve the accuracy of exogenous sequence mapping.

### Analyzing exogenous RNA in human samples with sRNAnalyzer

To test the ability of sRNAnalyzer to identify exogenous RNAs in samples, we used a dataset of plasma samples from CRC patients that we published earlier (11). The data contains nine human plasma samples—three from CRC patients, three from Ulcerative Colitis (UC) patients and three from healthy individuals (Norm). We used log2-transformed and RPM normalized data for comparison. Despite different clinical conditions, the overall profile of mapped sequence reads at domain level is similar among the samples (Figure 5A) which is consistent with prior findings. However, at the genus level, some CRC samples showed higher concentrations of *Pseudomonas* (crc1 and crc2) and *Ralstonia* (crc2 and crc3) derived sequences compared to UC and Norm samples (Figure 5B and C).

We further identified RNA derived from two bacterial species—*Pseudomonas fluorescens* and *Ralstonia pickettii* in the three CRC samples that were more abundant than other samples. When compared to human CRC tissue samples (34) reanalyzed with sRNAnalyzer, the abundance of RNA derived from *Pseudomonas* and *Ralstonia* were both significantly higher (>3-fold difference with $P$-value $< 0.05$) in metastatic tissue samples (Figure 6B and C). We did not see this correlation between tissue and plasma datasets without rRNA/tRNA filtering. This also suggests the potential of identifying disease-associated exogenous species-derived RNA in circulation using rRNA and tRNA filtering with LPM in sRNAnalyzer.

## DISCUSSION AND CONCLUSIONS

There are many experimental and data analysis challenges in the field of small RNA profiling, especially when dealing with body fluid samples. To improve the reliability of NGS-based small RNA profiling, sRNAnalyzer adopts a comprehensive adapter trimming strategy and sequence quality assessment in the data preprocessing module. For sequence mapping, we developed a flexible database searching approach to fit specific goals for different studies. We used several public domain datasets to evaluate the performance and determine the proper setting of various parameters in sRNAnalyzer. In the result summarization module, sRNAnalyzer provides several reporting formats for users to view and inspect the mapping results. Compared to other tools, the sRNAnalyzer has many enhanced features including (i) modular pipeline design, (ii) rigorous data preprocessing steps, (iii) endogenous and exogenous RNA mapping capability, (iv) use of precursor miRNA sequences annotated with mature sequence information for miRNA mapping, (v) use of LPM (local probability mapping) strategy to increase mapping specificity, (vi) extensive rRNA and tRNA filtering steps to enhance the accuracy of exogenous RNA mapping and (vii) various levels of result summarization. These features increase the flexibility for users to inspect the results and gain a global view of the small RNA in the sample.

For miRNA profiling, sRNAnalyzer adopted both the single and multiple assignment approaches as introduced in sRNAbench (19) to report the miRNA mapping results. Both approaches support different levels of error tolerance, from perfect match to two mismatches allowed. The single assignment approach based on LPM provides a more reliable picture of the RNA profile in a sample, while results from the multiple assignment approach provide better agreement with results from other miRNA measurement platforms such as microarray and qRT-PCR, especially when applying weighted correction factors for different levels of mismatch allowance. The final estimation of the expression values were calculated with all mapped reads and summarized as miRNA profiling results. This weight approach actually considered sequence mismatches caused by all possible reasons including SNPs, RNA editing issues, as well as sequence errors.

Since a significant fraction of circulating RNAs are the degraded products of large transcripts, sRNAnalyzer includes a number of endogenous sequence databases including various noncoding RNAs, protein coding transcripts and genomic DNA to gain a comprehensive profile of RNA in circulation. For these endogenous sequences, sRNAnalyzer uses 'map and remove' structure (i.e. only unmapped reads will be fed into next steps) with progressive alignment strategy to sequentially map reads against various databases. The mapping order of databases and the error tolerance on each mapping step depend on the research focus. This strategy can enrich the focused small RNA family, while may also bring a certain level of bias due to multiple alignments. To avoid this bias, sRNAnalyzer provides a very convenient way to change the mapping order and mismatch allowance by simply changing a text-based configuration file instead of reprogramming.

The extensive endogenous sequence mapping step is critical for exogenous sequence mapping since it removes all possible endogenous sequence reads. This may also raise the concern regarding the possibility of removing too many reads derived from exogenous species. To address this concern, we tested our pipeline using a dataset derived from RNA samples extracted from cultured bacteria. Most of the bacteria-derived reads still passed the rigorous endogenous 'filtering step' and correctly mapped to bacteria species (Figure 4A). This suggests that the rigorous human reference sequence mapping steps (allowing a maximum of two mismatches) do not remove significant fraction of the bac-

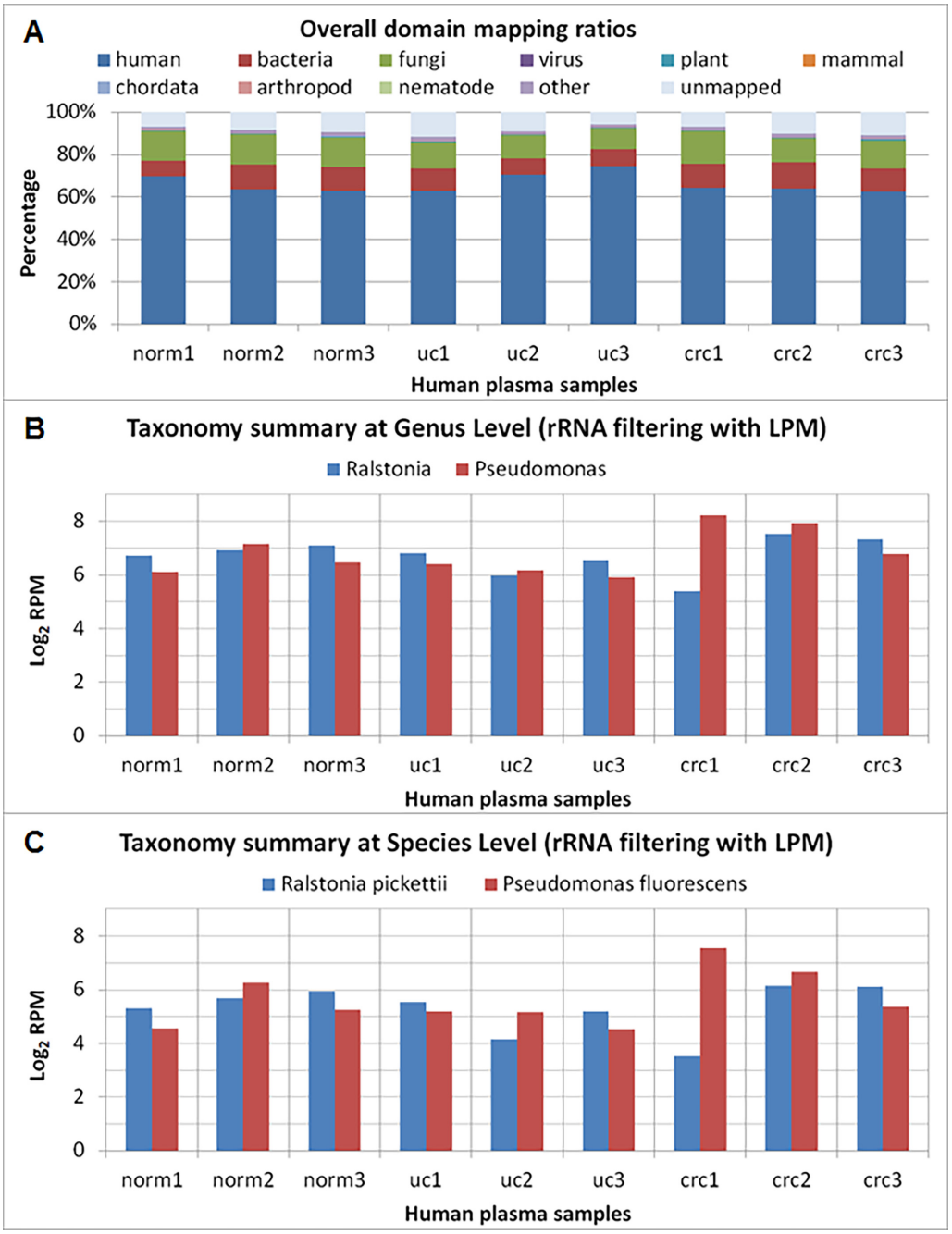


**Figure 5.** Overall domain mapping ratios and taxonomy summarization for human plasma samples from an exogenous RNA spectra study. (**A**) Overall domain mapping ratios; and (**B**) Taxonomy summary at Genus level; and (**C**) Taxonomy summary at Species level for human plasma samples from an exogenous RNA spectra study (SRA Session: ERP002414).

teria sequence reads (over 95% of sequence reads still remained).

Due to high sequence conservation of rRNA and tRNA sequences across species, sRNAnalyzer offers users an option to remove these reads prior to exogenous sequence mapping. We showed that rRNA and tRNA filtering is critical to identify the correct bacterium species in the case study of bacterial culture samples (Figure 4). To avoid information loss, sRNAnalyzer can summarize reads mapped to exogenous RNAs at different taxonomy levels with or without considering those filtered rRNAs/tRNAs, although the default setting is the latter one. In addition, the use of LPM can further enhance the specificity of exogenous RNA mapping. Our pipeline can also be used for samples from different species by making specific databases for different species and changing the order of mapping steps in the configuration files. Currently we have configuration files for human, mouse, rat, horse, macaque and plant for different research focuses.

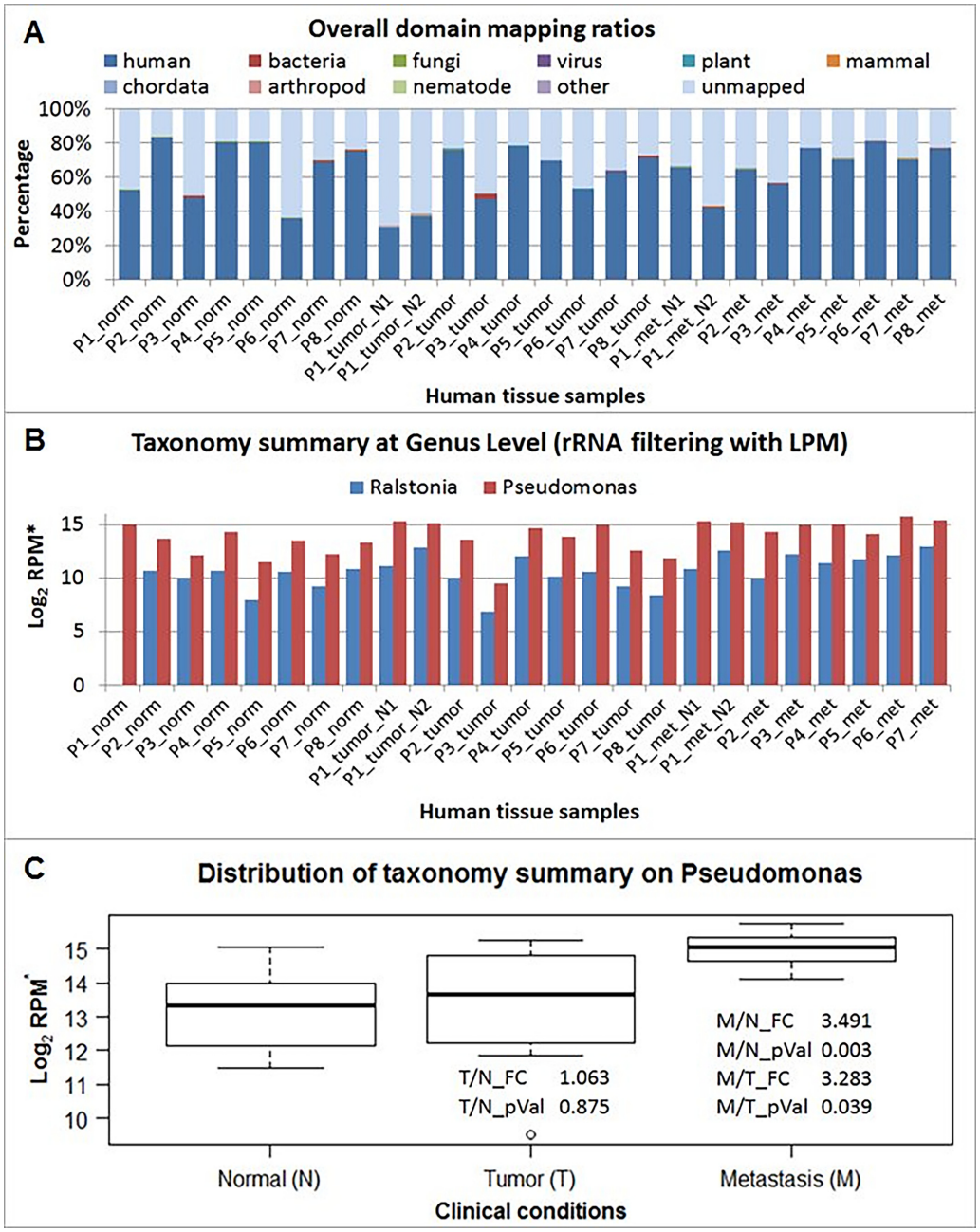


**Figure 6.** Overall mapping ratios and taxonomy summarization for human tissue samples from a CRC study. (**A**) Overall domain mapping ratios; (**B**) Taxonomy summary on Pseudomonas; and (**C**) Distribution of taxonomy summary on Pseudomonas for human tissue samples from a CRC study (SRA Session: SRP022054).

In summary, sRNAnalyzer is a full function, customizable small RNA analysis pipeline for NGS data. It can effectively map both endogenous and exogenous RNA. It also offers different levels of detail in the mapping results—from individual sequence reads, transcripts, species to domain. This allows users to gain a better picture of the complexity of small RNA in samples.

## AVAILABILITY

GEO accession number: GSE94912.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors thank the sequencing facility at Institute for Systems Biology (ISB) for sequencing the NGS libraries. The authors also thank Dr Inyoul Lee for stimulating discussions, Andrew Xue, Drs Kathie-Anne Walters and Mary Brunkow for critical reading and suggestions for the manuscript.

## FUNDING

US Department of Defense [W911NF-10–2-0111, HDTRA1–13-C-0055]; National Institutes of Health (NIH) Extracellular RNA Communication Consortium (ERCC) [1U01HL126496]. Funding for open access charge: NIH ERCP Consortium [1U01HL126496].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Krol,J., Loedige,I. and Filipowicz,W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
2. Weber,J.A., Baxter,D.H., Zhang,S., Huang,D.Y., Huang,K.H., Lee,M.J., Galas,D.J. and Wang,K. (2010) The microRNA spectrum in 12 body fluids. *Clin. Chem.*, **56**, 1733–1741.
3. Cortez,M.A., Bueso-Ramos,C., Ferdin,J., Lopez-Berestein,G., Sood,A.K. and Calin,G.A. (2011) MicroRNAs in body fluids—the mix of hormones and biomarkers. *Nat. Rev. Clin. Oncol.*, **8**, 467–477.
4. Etheridge,A., Lee,I., Hood,L., Galas,D. and Wang,K. (2011) Extracellular microRNA: a new source of biomarkers. *Mutat. Res.*, **717**, 85–90.
5. Chevillet,J.R., Lee,I., Briggs,H.A., He,Y. and Wang,K. (2014) Issues and prospects of microRNA-based biomarkers in blood and other body fluids. *Molecules*, **19**, 6080–6105.
6. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
7. Hackenberg,M., Rodríguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
8. Li,Y., Zhang,Z., Liu,F., Vongsangnak,W., Jing,Q. and Shen,B. (2012) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res.*, **40**, 4298–4305.
9. Rueda,A., Barturen,G., Lebrón,R., Gómez-Martín,C., Alganza,Á., Oliver,J.L. and Hackenberg,M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.*, **43**, W467–W473.
10. Etheridge,A., Gomes,C.P., Pereira,R.W., Galas,D. and Wang,K. (2013) The complexity, function, and applications of RNA in circulation. *Frontiers in genetics*, **4**, 115.
11. Wang,K., Li,H., Yuan,Y., Etheridge,A., Zhou,Y., Huang,D., Wilmes,P. and Galas,D. (2012) The complex exogenous RNA spectra in human plasma: an interface with human gut biota? *PLoS One*, **7**, e51009.
12. Witwer,K.W. and Hirschi,K.D. (2014) Transfer and functional consequences of dietary microRNAs in vertebrates: concepts in search of corroboration. *Bioessays*, **36**, 394–406.
13. Witwer,K.W., McAlexander,M.A., Queen,S.E. and Adams,R.J. (2013) Real-time quantitative PCR and droplet digital PCR for plant miRNAs in mammalian blood provide little evidence for general uptake of dietary miRNAs: limited evidence for general uptake of dietary plant xenomiRs. *RNA Biol.*, **10**, 1080–1086.
14. Buck,A.H., Coakley,G., Simbari,F., McSorley,H.J., Quintana,J.F., Le Bihan,T., Kumar,S., Abreu-Goodger,C., Lear,M. and Harcus,Y. (2014) Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat. Commun.*, **5**, 5488.
15. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
16. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.
17. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
18. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
19. Barturen,G., Rueda,A., Hamberg,M., Alganza,A., Lebron,R., Kotsyfakis,M., Shi,B.-J., Koppers-Lalic,D. and Hackenberg,M. (2014) sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing*, **1**, 21–31.
20. Griffiths-Jones,S., Grocock,R.J., Van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
21. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Meng,Y., Shao,C., Wang,H. and Chen,M. (2012) Are all the miRBase-registered microRNAs true? A structure-and expression-based re-examination in plants. *RNA Biol.*, **9**, 249–253.
23. Castellano,L. and Stebbing,J. (2013) Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. *Nucleic Acids Res.*, **41**, 3339–3351.
24. Fromm,B., Billipp,T., Peck,L.E., Johansen,M., Tarver,J.E., King,B.L., Newcomb,J.M., Sempere,L.F., Flatmark,K. and Hovig,E. (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.
25. Ghosal,A., Upadhyaya,B.B., Fritz,J.V., Heintz-Buschart,A., Desai,M.S., Yusuf,D., Huang,D., Baumuratov,A., Wang,K. and Galas,D. (2015) The extracellular RNA complement of Escherichia coli. *Microbiol. Open*, **4**, 252–266.
26. Wang,K., Yuan,Y., Li,H., Cho,J.-H., Huang,D., Gray,L., Qin,S. and Galas,D.J. (2013) The spectrum of circulating RNA: a window into systems toxicology. *Toxicol. Sci.*, **132**, 478–492.
27. Sobala,A. and Hutvagner,G. (2011) Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdiscip. Rev.*, **2**, 853–862.
28. Anderson,P. and Ivanov,P. (2014) tRNA fragments in human health and disease. *FEBS Lett.*, **588**, 4297–4304.
29. Dhahbi,J.M., Spindler,S.R., Atamna,H., Boffelli,D. and Martin,D.I. (2014) Deep sequencing of serum small RNAs identifies patterns of 5′ tRNA half and YRNA fragment expression associated with breast cancer. *Biomark. Cancer*, **6**, 37–47.
30. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
31. Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,C.T. Jr, Saxman,P.R., Farris,R.J., Garrity,G.M., Olsen,G.J., Schmidt,T.M. and Tiedje,J.M. (2001) The RDP-II (ribosomal database project). *Nucleic Acids Res.*, **29**, 173–174.
32. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
33. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
34. Röhr,C., Kerick,M., Fischer,A., Kühn,A., Kashofer,K., Timmermann,B., Daskalaki,A., Meinel,T., Drichel,D. and Börno,S.T. (2013) High-throughput miRNA and mRNA sequencing of paired colorectal normal, tumor and metastasis tissues and bioinformatic modeling of miRNA-1 therapeutic applications. *PLoS One*, **8**, e67461.
35. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
36. Lusk,R.W. (2014) Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*, **9**, e110808.