A comprehensive profile of circulating RNAs in human serum

Sinan Uğur Umu, Hilde Langseth, Cecilie Bucher-Johannessen, Bastian Fromm, Andreas Keller, Eckart Meese, Marianne Lauritzen, Magnus Leithaug, Robert Lyle & Trine B. Rounge

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

🔓 OPEN ACCESS ⏻ Check for updates

A comprehensive profile of circulating RNAs in human serum

Sinan Uğur Umu [ID]a, Hilde Langseth [ID]a, Cecilie Bucher-Johannessen [ID]a, Bastian Fromm [ID]b, Andreas Kellerc, Eckart Meesed, Marianne Lauritzen [ID]a, Magnus Leithauge, Robert Lylee,f, and Trine B. Rounge [ID]a

aDepartment of Research, Cancer Registry of Norway, Oslo, Norway; bDepartment of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Nydalen, Oslo, Norway; cDepartment of Clinical Bioinformatics, Saarland University, Saarbruecken, Germany; dDepartment of Human Genetics, Saarland University, Homburg/Saar, Germany; eDepartment of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway; fPharmaTox Strategic Research Initiative, School of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway

ABSTRACT

Non-coding RNA (ncRNA) molecules have fundamental roles in cells and many are also stable in body fluids as extracellular RNAs. In this study, we used RNA sequencing (RNA-seq) to investigate the profile of small non-coding RNA (sncRNA) in human serum. We analyzed 10 billion Illumina reads from 477 serum samples, included in the Norwegian population-based Janus Serum Bank (JSB). We found that the core serum RNA repertoire includes 258 micro RNAs (miRNA), 441 piwi-interacting RNAs (piRNA), 411 transfer RNAs (tRNA), 24 small nucleolar RNAs (snoRNA), 125 small nuclear RNAs (snRNA) and 123 miscellaneous RNAs (misc-RNA). We also investigated biological and technical variation in expression, and the results suggest that many RNA molecules identified in serum contain signs of biological variation. They are therefore unlikely to be random degradation by-products. In addition, the presence of specific fragments of tRNA, snoRNA, Vault RNA and Y_RNA indicates protection from degradation. Our results suggest that many circulating RNAs in serum can be potential biomarkers.

Introduction

Human serum and plasma contain various classes of RNA molecules [1-3] such as protein-coding messenger RNAs (mRNA) [4], miRNAs [3,5-10], piRNAs [1,11,12], tRNAs and miscellaneous other ncRNA molecules [1,11]. These circulating RNAs are usually packed in extracellular vesicles and have considerable potential as minimally-invasive biomarkers [4,5,8,11,13,14], since they are stable and some have been associated with disease phenotypes [5,6,11,15,16].

miRNAs are the best characterized class of sncRNA molecules. They are approximately 22 nucleotides (nts) in length and regulate cellular gene expression via RNA-RNA antisense binding [17-19]. They can also be found as circulating RNAs [3,5-8,20]. Many studies have investigated the biomarker potential of miRNAs [2,5-9,16,21,22] and their isoforms, isomiRs [23-25]. Small nucleolar RNAs (snoRNAs) are another well-known member of sncRNA molecules. They play a crucial role in ribosomal RNA (rRNA) maturation [26] and can be found as extracellular RNAs [4,12]. piRNAs, initially discovered in germline cells [27,28], are a less studied class of small RNA molecules, however, recent studies have identified them as circulating RNAs [1,11,12]. Besides regulatory sncRNAs, protein-coding mRNAs and tRNAs are also found as circulating RNAs [11] despite their roles in protein synthesis.

Furthermore, tRNA-derived small RNAs or tRNA-derived fragments (tRFs) are known to have specific cellular expression patterns [29,30] and are associated with some cancer types [31]. This makes extracellular tRNAs and their fragments potential biomarkers.

Large portions of the human genome are biochemically and transcriptionally active [32-34]. Efforts have been made to deduce the roles of cellular RNAs and their fragments [35-40]. Different body fluids, including serum, have been investigated for extracellular RNAs [4,41,42]. The functionality of these RNA molecules is an open question [4,11], since they can be mere degradation by-products, experimental noise or have alternative roles in circulation. The studies so far have mostly focused on analyzing circulating miRNAs to understand their function and determine biomarker potential. Yet, it has been shown that the variation of circulating miRNA expression can be influenced by different biological (e.g. disease, age, sex, body mass index etc.) [2,20,42,43] and technical factors (e.g. lab processing, platform, noise etc.) [11,44,45], which can greatly affect profiles of highly expressed miRNAs [1,12,20,42]. Therefore, it is important to understand 'normal' RNA content of human serum before utilizing RNAs as biomarkers.
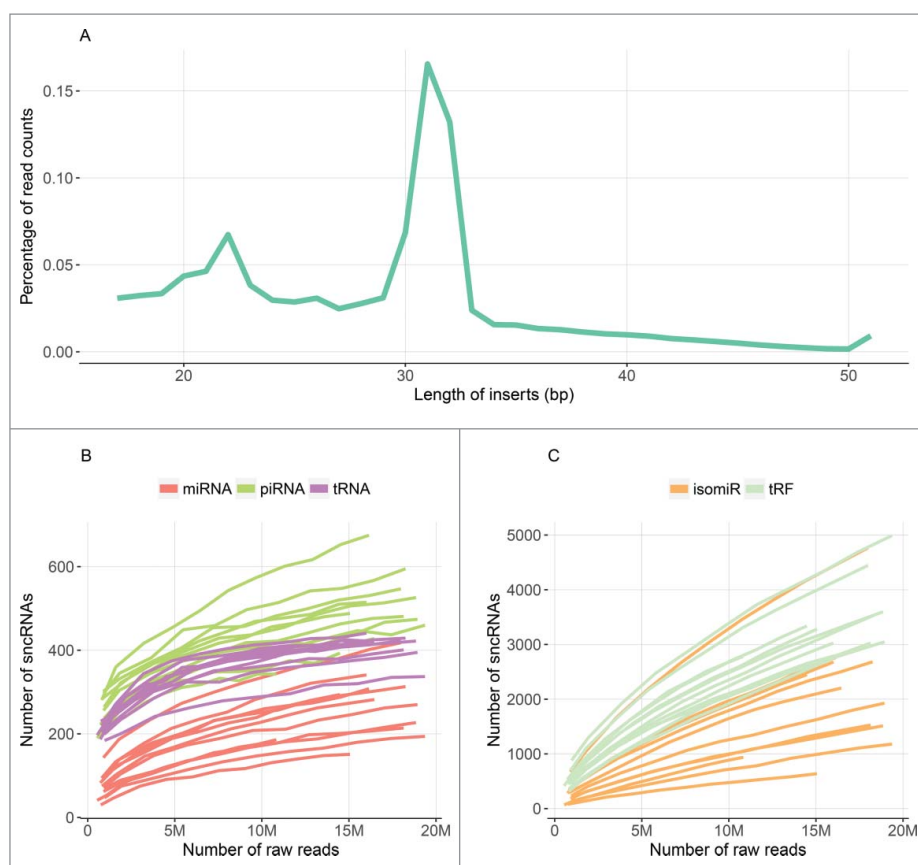
Figure 1. (A) The line shows the distribution of trimmed RNA molecule sizes for the serum samples. Our theoretical input library size is between 17 and 47 nts. There are two peaks for the reads at 22 and 31 nts length. This enabled us to detect numerous RNA types including fragments of lncRNAs and mRNAs. (B) The saturation lines of canonical genes (i.e. miRNAs, piRNAs, and tRNAs) for a randomly selected subset of serum samples (n = 12) are shown. The number of identified genes are still increasing for piRNAs (the dark green lines) but the others are about to reach plateau. (C) The non-canonical isoforms (i.e. isomiRs and tRFs) identified are also increasing with the sequencing depth and far from reaching plateau.

The aim of this study was to profile RNA molecules in human serum. We analyzed small RNA-seq data from a large (N = 477) set of long-term archived serum samples. To assess potential functionality, we analyzed biological variation of sncRNAs and expression/degradation patterns of RNA fragments. To date, this is the most comprehensive analysis of the sncRNA repertoire in human serum.

Results

Overall RNA profiles

We analyzed the RNAs in the size range of 17 to 47 nts (Fig 1A). This entails mostly sncRNAs, but it also includes fragments of long non-coding RNAs (lncRNA), mRNAs and other longer transcripts. miRNAs are represented with a peak at 22 nts. The completeness of the profiles relies on sequencing depth, and the saturation analyses showed that canonical miR-NAs and tRNAs are approaching plateau with a sequencing depth of about 10–15 Million reads (Fig. 1B). However, the number of piRNAs, isomiRs and tRFs are still increasing at 15 Million reads (Fig. 1B, C).

We found a total of 258 miRNA, 441 piRNA, 411 tRNA, 24 snoRNA, 125 snRNA and 123 misc-RNA genes that passed the expression threshold that we set (median expression > = 10 reads), representing the core RNA expression profile of serum.

In addition, 87 lncRNAs and 1334 mRNAs were detected because of the RNA fragments mapped to these annotations. The transcript origin of RNA reads mapping to multiple genomic locations cannot be determined when mapping qualities are equal for several locations. For comparability to previous studies, we show profiles using both uniquely and multi-mapped reads (Fig. 2). Multi-mapped sequence counts enriches the abundance of high-copy number genes (e.g. piRNA and tRNA). We also used this approach for RNA identification in this study.

The overall RNA expression profile shows that some RNA classes are highly expressed compared to others and the top expressed RNAs are listed in Table 1. The misc-RNA class includes Y_RNAs, Signal Recognition Particle (SRP) RNA and Vault RNAs etc. (Table 1). The snoRNAs include U3, U8 and some other related C/D or H/ACA box snoRNAs (Table S4). The snRNAs include U2, U1, U6 and related snRNA genes (Table S5). Complete lists of all identified RNAs are in supplementary tables (Tables S1-S8).

Isoform profiles of miRNAs and tRNAs

We identified 1642 isomiRs in the serum samples, which passed the detection threshold (i.e. median expression > = 10 among samples). The average GC contents of serum isomiRs, canonical forms and miRNA precursors are 0.51, 0.50 and 0.52
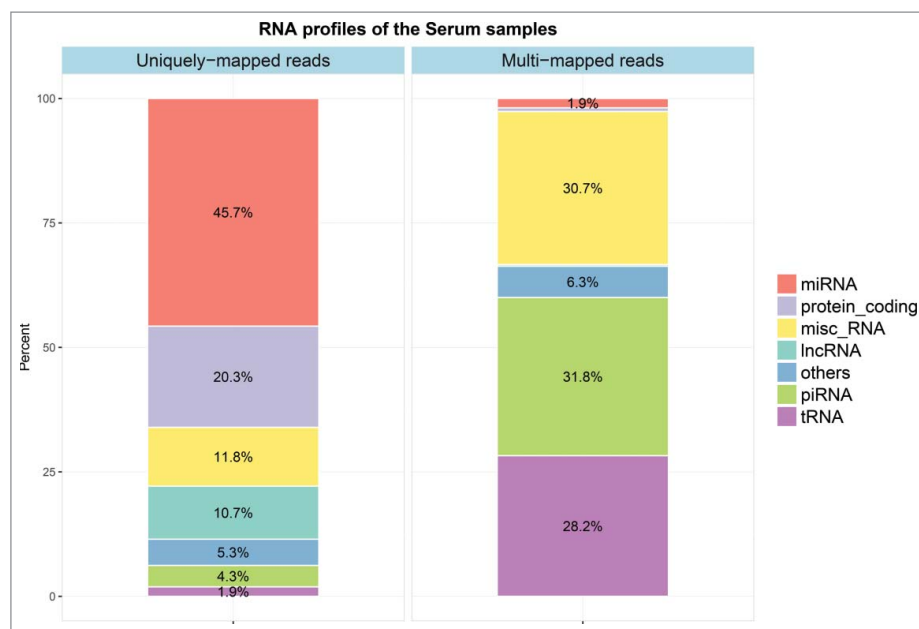
Figure 2. An overall classification of the mapped reads of the serum samples (n = 477). This pie-chart on the left, generated using uniquely-mapped reads, shows an abundance of miRNA hits followed by protein-coding mRNAs and misc-RNAs. Allowing multi-mapped reads is affecting overall RNA profiles (on the right). For multi-mapped reads, piRNAs (green) are the most abundant RNA type followed by misc-RNAs (yellow) and tRNAs (purple). The annotations of GENCODE v26 and piRBase were used to create these plots. Similar pie-charts for the technical replicates are at the supplementary (Fig. S2).

respectively. The isomiRs are mostly 3′ isomiRs (78%), followed by 5′ (27%), substitution (22%) and canonical forms (8%). The identified isomiRs are generally an isoform of highly expressed miRNAs (Table 1). For example, hsa-miR-320a, hsa-miR-423-5p, hsa-miR-122-5p and hsa-miR-1246 have 159, 138, 73 and 55 isoforms respectively. A detailed list of the serum isomiRs and their precursors is provided in supplementary (Table S1A).

We identified 1900 tRFs in the serum samples. The average length of these tRNA fragments is ∼29 nts and the average GC content is 0.53. A detailed examination of tRFs showed that they originated from either the 5′ or 3′ end of mature tRNAs (Fig. 3A). This suggests there are no mature tRNAs in serum. The 3′ end of tRNAs was the most abundant region with a uniform distribution throughout a 30 nts region (Fig. 3A).

Profiles of RNA fragments

We also analyzed the profiles of RNA molecules mapped to other annotated regions, including snoRNAs, Vault RNAs, Y_RNAs, mRNAs and lncRNAs. First, U3 snoRNAs are the most abundant within the snoRNA class (Table S4) and the average size of all U3 snoRNA mapped reads is around 29 nts with an average GC content of 0.51. These reads usually come from two regions, the first 20 nts or the last 22 nts region (Fig. 3B), but there are also two smaller peaks between nts 48–74 and 169–195. Second, Vault RNAs have a consistent signal of expression with reads derived from a region covering 75th – 95th nts, while the total size of the Vault MSA is 101 nts (Fig. 3C). These reads also have higher average GC contents, 0.62, than their host Vault RNAs, 0.52. Third, Y_RNAs constitute most of the misc-RNA group's expression (Table 1). The MSA of Y_RNAs consist of 51 Y_RNAs and 179 nts (Fig. 3D). The expression profiles of Y_RNAs showed that the reads were mapped to the first 1–50 nts region. The average GC content of these reads is 0.51 with an average length of 37 nts. Lastly, as mentioned in the Materials and Methods, we counted the reads only mapped to exonic regions of mRNAs and lncRNAs. The fragments mapped to exonic regions of longer annotations (i.e. mRNA and lncRNA) have average sizes of 29 nts for mRNAs and 30 nts for lncRNAs with GC contents of 0.52 and 0.51 respectively.

Table 1. A summary table of highly expressed RNAs identified in the serum samples.

Expression Rank	miRNA	piRNA	misc-RNA	lncRNA	mRNA
1	hsa-miR-423-5p	piR-hsa-25779	Y_RNA	RP11-1151B14.3	NSRP1
2	hsa-miR-320a	piR-hsa-25780	RNY4	RP11-20B24.2	WDR74
3	hsa-miR-1246	piR-hsa-12790	RNY1	LINC00910	VMP1
4	hsa-miR-122-5p	piR-hsa-2106	RN7(x)	LINC00324	HOXB4
5	hsa-miR-1290	piR-hsa-25783	RNY3	LINC01783	ATP5G3
6	hsa-miR-21-5p	piR-hsa-25782	SRP	RP11-108M9.3	MTRNR2L8
7	hsa-miR-486-5p	piR-hsa-18709	VTRNA1(x)	RP11-473M20.16	C9orf3
8	hsa-miR-148a-3p	piR-hsa-2107	KCNQ1OT1_5	CARMN	MTRNR2L12
9	hsa-miR-451a	piR-hsa-25781	7SK	RNU11	MTRNR2L1
10	hsa-miR-101-3p	piR-hsa-1207	Vault RNA	RP11-160E2.6	FAM212A

Note: *these miRNAs are challenged, see the Discussion. **similar annotations are collapsed for misc-RNAs. The extended lists are available in Supplementary Tables S1-S8.
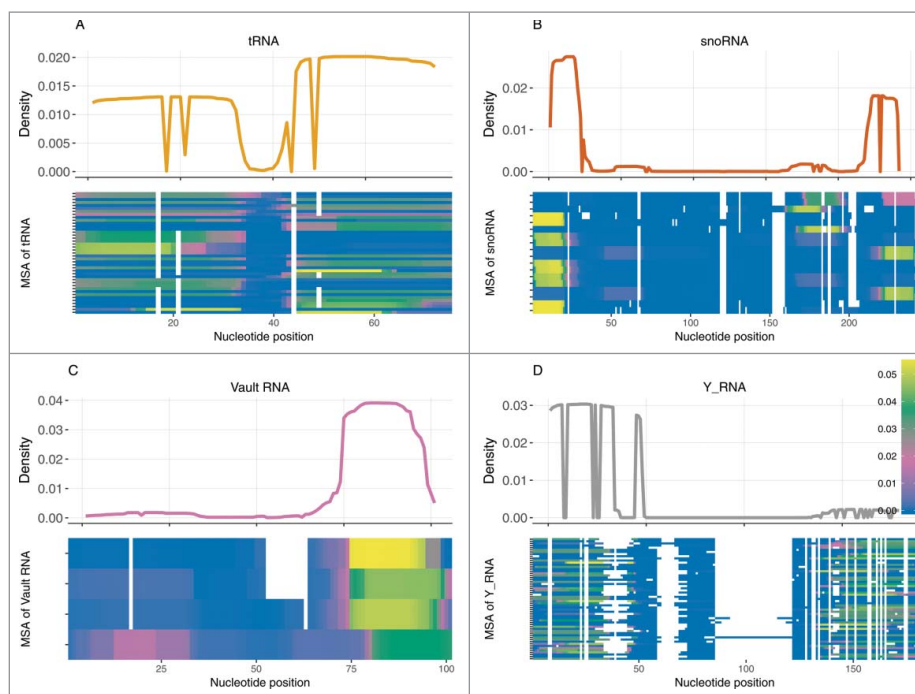
Figure 3. The profiles of mapped reads from highly expressed (A) tRNAs (n = 41), (B) U3 snoRNAs (n = 18), (C) Vault RNAs (n = 4) and (D) Y_RNAs (n = 57). Each panel has a multiple sequence alignment (MSA) at the bottom and a corresponding density plot at the top. The x-axes of all plots display a nt position on their MSAs. For example, the MSA of tRNAs is 75 nts long which can be seen at the bottom of the plots. The density plots shows the overall mapping profiles and their x-axes also display nt positions. The heat-maps provide colored representation of the density plot per RNA in the alignment. Yellow and green correspond to the top expressed regions (i.e. high depth), while blue contain almost no mapped reads. White are the gaps in the alignment. (A) The reads mapped to mature tRNAs are mostly coming from the 3′ ends (density plot). (B) There is a peak at the 5′ end of the snoRNA density plot that corresponds to a 20 nts long region. (C) The Vault RNAs identified have a clear signal of expression at their 3′ ends (density plot and yellow bricks at the heatmap). (D) The Y_RNA reads are mostly originating from 5′ ends and there is a small peak at the 3′ end (density plot).
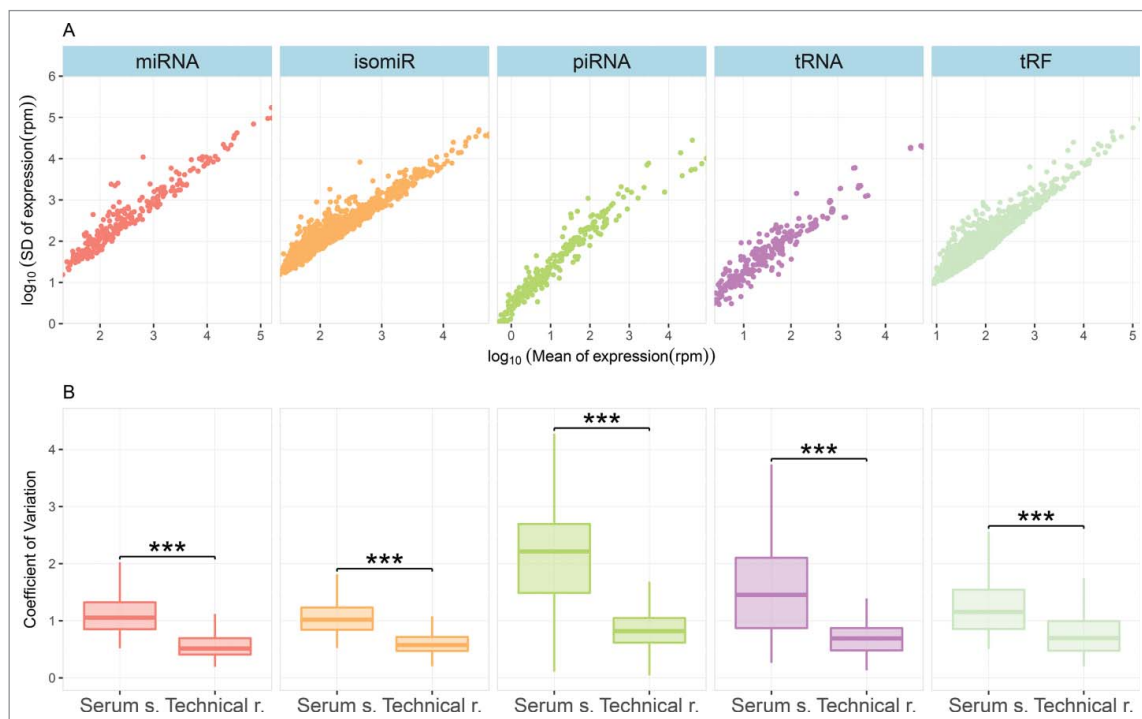
Figure 4. (A) The y-axis shows the $\log_{10}$ of standard deviations of normalized expression and the x-axis shows the $\log_{10}$ mean expression of identified sncRNAs. (B) The boxplots show the distribution of CV values in the serum samples and the technical replicates. A pairwise MWU test (*** $p << 0.0001$) confirmed higher CV values in the serum samples than the technical replicates suggesting higher biological variation for the serum samples than the technical replicates. Randomly generated subsamples of the serum samples (n = 17) also produces similar results (Fig. S3) excluding variation due to different samples sizes.

Coefficient of variation (CV) analyses of sncRNA expression

We analyzed variation in expression of identified sncRNAs to investigate biological signals. In the serum samples, there is a linear relationship between log-normalized mean expression and the standard deviation of identified sncRNAs (Fig. 4A), which shows that the variation is higher for the highly expressed sncRNAs.

A CV value measures dispersion of a distribution and is a standardised measure of the standard deviation. Distributions of CV values per sncRNA class for both the serum samples and the technical replicates were calculated. We hypothesized that RNA expression in the serum samples will vary more than the technical replicates due to biological variance, because the variation in RNA expression of the serum samples is a combination of technical and biological factors. We tested the null hypothesis: there is no difference in CV values of these two sample sets in three sncRNA types (i.e. miRNA, piRNA and tRNA) and in two different isoforms. We found that the RNA expression varies more in the serum samples than the technical replicates (one sided Mann-Whitney U test (MWU), $p << 0.0001$ for all) (Fig. 4B). This means that the CV values of RNA expression in the technical replicates are consistently lower than in the serum samples for all sncRNA types, including isoforms (i.e. isomiRs and tRFs).

Low technical variation is preferable for a biomarker 44], so removing the sncRNAs with high technical variation should create a better set of biomarkers. As an example we tested this with cluster analyses using isomiRs identified both in the serum and technical replicates. The detected isomiRs were divided into four groups based on their CV: all isomiRs (n = 1642, identified in both sample groups), low CV (lower than median CV, n = 797), very low CV (lower than first quantile, n = 403) and high CV isomiRs (higher than median CV, n = 845). The four dendrograms created from these groups showed that the low CV and very low CV isomiRs can successfully cluster a set of randomly selected serum samples (n = 17) and technical replicates (Fig. S4). However, all isomiRs and the high CV isomiRs cannot successfully cluster these two sample types (Fig. S4). We detected a GC difference between the high CV (0.52) and low CV (0.49) isomiRs (two sided MWU, p = 0.003) which may be a reason for the additional technical variation in some isomiRs. Their average internal folding energies, −1.19 kcal/mol for the high and −1.17 kcal/mol for the low CV group, are also slightly different (two sided MWU, p = 0.014), which is most likely an effect of the GC difference.

Discussion

A biomarker is a measurable indicator of a biological state or a phenotype [46,47]. There is increasing interest in early-detection of diseases using RNA biomarkers, and numerous studies have investigated circulating miRNAs as candidate non-invasive biomarkers [2,5-9,16,21,22]. We expanded previous research by generating the most comprehensive RNA profile of serum which reports existence of some RNA classes in human serum for the first time. Our in-depth analyses include not only miRNAs, but also piRNAs, tRNAs, snoRNAs, snRNAs, misc-RNAs, lncRNAs, mRNAs and RNA fragments such as isomiRs, tRFs, RNA derived particles.

To be able to analyse all the sncRNAs, a size filtering of 15–40 nts is sufficient [46]. With our insert size selection (17-47 nts) we were able to do a complete profiling of serum sncRNAs (Fig. 1A). The fragments of lncRNAs, mRNAs and other longer transcripts were also detected in serum. Sequencing depth influences sensitivity of RNA-seq (Fig. 1B) and this is especially notable for isoforms (Fig. 1C). The average sequencing depth is high and selection of a lower threshold (i.e 5) would allow identification of 23% more miRNAs (i.e. 318), 10% more piRNAs (i.e. 482) and 11% (i.e. 457) more tRNAs, compared to the reported core set (Tables S1-S8). The total number of identified miRNAs in serum was reported between 90 and 700 in the previous profiling studies [7,10,20,45], which was between 123 and 500 for plasma samples [1,12,20,42]. The total piRNA counts in plasma samples were reported to be around 120 [11,12], while our data identified at least three times as many piRNAs. The serum samples in this study can be up to 40 years old, however, the results suggest that many RNA classes are still recoverable with a high expression signal. There is a slight difference between the overall RNA contents of the serum (Fig. 1) and the (fresh) technical replicates (Fig. S2). This difference is most likely an artifact of pooling several samples together rather than of degradation. Although our data revealed some loss of miRNAs and isomiRs over time, the effects ($R^2 = 0.11$ and $R^2 = 0.14$, respectively) are low (Fig. S5).

The core set of RNAs were reported by selecting a high expression threshold, which filtered out the RNA products with less stable expression. Our analyses produced comparable results with previous circulating sncRNA profiling of different body fluids in terms of RNA diversity. However, the RNA profiles can vary between studies, which is also true for highly expressed RNAs [1]. We found examples of highly expressed serum sncRNAs that were previously reported as circulating RNAs. For example, the highly expressed miRNAs in our serum samples, hsa-miR-423-5p, hsa-miR-320a, hsa-miR-122-5p, hsa-miR-486-5p, hsa-miR-486-3p were detected in blood samples [1,6,46]. Hsa-miR-451a, among our top 10 expressed miRNAs, was reported to be the most abundant miRNA in plasma [12]. Hsa-miR-1290 and hsa-miR-1246 were detected in serum and associated with metastasis of lung cancer tumors [48]. Some of the highly expressed piRNAs in our serum samples (e.g. piR-hsa-2106 (pir-001311), piR-hsa-27493 (pir-019825), piR-hsa-23209 (pir-020496), piR-hsa-28223 (pir-020388), piR-hsa-28527 (pir-020582), piR-hsa-28374 (pir-020485)) are known to exist in plasma and a few of them were also associated with cancer phenotypes [11].

A single miRNA locus can produce various isomiRs with distinct length or sequence [49] and they have been associated with phenotypes and diseases [23-25]. Both in animal and plants, 3′ isomiRs are the most common ones [49], consistent with our results. We found that only 8% of the isomiRs are the canonical forms from miRBase, and highly expressed potential isomiRs can be identified in serum. tRFs are another less-known class of sncRNAs which are isoforms of tRNA genes [29]. They are derived either from mature tRNAs or 3′ of tRNA precursors [29,30] and expressed under various stress conditions [50,51]. Many tRFs were associated with different cancer phenotypes [30,31] and some were found to be functional like a regulatory miRNA

[52]. Random degradation of tRNAs should give a uniform distribution of tRFs covering the entire tRNA annotation [30]. However, we found that tRFs have non-uniform expression patterns (Fig. 3A), suggesting a regulated cleavage. This is consistent with known tRF biogenesis [29]. We also found potentially functional tRNA derived fragments. For example, tRF-5001 was detected in prostate cells in high amounts [29]. Moreover, 107 tRFs identified were associated with Argonaute family proteins and predicted to have possible mRNA targets [53]. One of these 5′ end tRFs have the maximum median expression in our serum samples (Table S3A). It was deposited to MINTbase tRF database (id tRF-30-PNR8YP9LON4V) [54] and also found to bind 12 different mRNAs (e.g. EI24, SUGP2 etc.) according to CLASH data [53].

There are RNA fragments originating from well-known annotations, such as snoRNA, misc-RNA, lncRNA and mRNA, that can be functional independent of their host gene [38,40,55]. In our dataset these RNA fragments are abundant (at least 40% of the all RNA molecules). SnoRNA derived fragments can act like miRNAs to suppress target gene expression [39] and Figure 3B shows that snoRNA in serum also have a non-uniform expression pattern, similar to tRFs. Y_RNAs are short misc-RNAs with functional roles in DNA replication and RNA stability [56,57]. These fragments, previously found as circulating RNAs in mammals [14,56], have been associated with apoptosis in human cells [58]. Vault RNAs and their fragments were also associated with drug resistance [55,59]. Vault RNAs are a part of ribonucleoprotein complexes [60,61]. They were identified as circulating RNAs in mammals [14]. Both Y_RNAs and Vault RNAs are highly abundant in our serum (Tables 1 and S6) and have a non-uniform expression patterns (Fig. 3C, D). Furthermore, lncRNA and mRNA fragments are known to have different roles such as competing for protein/oligonucleotide binding [62,63], and target gene regulation [64,65]. The RNA fragments mapped to them have similar size and GC distribution with other sncRNA fragments in our dataset. The expression is often high and stable for these fragments and they cover only small fractions of their host gene (i.e. non-uniformity).

An important question is whether the discovered sncRNAs and their fragments are genuine functional products. The above mentioned high expression pattern and regulated cleavage suggest function. Random degradation and experimental noise from RNA-seq studies [66–69] might introduce false positive prediction of biological function and associations due to lack of RNA-seq sensitivity [66,70]. We proposed that CV analysis (Fig. 3 and Fig. S4) is suited for suggesting biological variation, because in an ideal setting, technical replicates should contain no biological variation, only technical variation. However, variation in serum samples is a sum of both biological and technical variability. We identified a statistically significant difference in average CV between technical and serum samples for all sncRNA classes (including isoforms) that shows higher variation for serum samples. This supports a biological signal in serum RNA expression and suggests potential function for circulating RNA molecules.

Technical variation in RNA-seq may vary depending on RNA molecule characteristics such as expression level, size,

sequence and secondary structure. We measured a range of CV values in our technical replicates even though we expected them to be closer to zero (Fig. 4B). High technical variation can decrease biomarker value by influencing reproducibility. This can be observed in our cluster analysis: the low CV and very low CV isomiRs best discriminate the serum and technical replicate group. We detected a statistically significant difference between the GC contents of high and low CV isomiRs which may partly explain technical variation. Some of those highly discriminatory isomiRs (e.g. isomiRs of hsa-miR-192-5p, hsa-miR-375 etc.) were successfully clustering various cancer tissues in a binary classification approach [23]. Another 5′ end isoform of hsa-miR-101-3p, with a low technical variation in our study, was also found to have a role in gene silencing in brain tissues [25]. In short, this analysis showed that a set of isomiRs with low CV is less prone to technical variation and they successfully cluster the two groups.

The large sample size, high coverage and the diversity of RNA products analyzed are the strengths of our study. We extensively profiled abundant RNA fragments in serum, and showed specific cleavage patterns of some RNA fragments for the first time. We also utilized a set of technical replicates to measure biological signal of serum RNA expression. This analysis suggested functionality for RNA fragments. However, there are potential limitations that we should address.

First, long-term storage may degrade some unstable RNAs, though our results suggested that the degradation effect is not strong for sncRNAs (Figure S5). It has been proven for miRNAs that they remain stable in severe conditions [10] and in circulation [9]. They can be extracted from long-term serum [7,71]. Moreover, any RNA found in serum stored up to 40 years is evidently quite stable, which is one of the critical criteria for good biomarkers. Second, although all samples are processed in the same way, slight differences in laboratory processing may still introduce some technical variance into expression which cannot be removed totally. We addressed this variation (Fig. 4B) using the technical replicate samples and CV values, which showed that higher technical variation was introduced into some sncRNAs than the others. Third, the lab and bioinformatic analysis methods chosen may compromise generalizability of results. For example, differences in gel cut size will change proportions of sncRNAs and narrower cut will limit detection of certain sncRNA classes. Detection threshold and allowing multi-mapped reads will also change the overall RNA profiles substantially (Fig. 2). Selection of read mapper and algorithm parameters are other bioinformatics related factors that can influence overall results [72]. Furthermore, high quality annotations are also essential to correctly identify transcripts [73], which is still a major barrier even for well-studied human miRNAs [74]. For example, highly expressed miRNAs, hsa-miR-1246 and hsa-miR-320a, are questioned for not being a miRNA gene [74]. Since they are part of miRBase, we reported them (and their isoforms) as miRNAs to be consistent with the literature. However, improving annotation quality is an on-going process and still far from perfect. It is also reasonable to consider possible alternative functions of the RNA fragments derived from longer host genes rather than counting them as a single piece of a large annotation. For instance, counting tRFs or misc-RNA derived fragments as their host genes would have

overshadowed the specific expression patterns that we reported in Figure 3.

Conclusion

Here we present a comprehensive characterization of human serum sncRNA content. Our results unveiled that most of the RNAs identified in serum are not random by-products but most likely have roles as circulating RNAs. This conclusion is supported by (1) stable high expression, (2) biological signal and (3) distinct expression patterns of many identified RNA molecules. Our results suggest new opportunities for novel biomarker discovery in serum, but they are also transferable to other body fluids and tissues.

Materials and methods

Study design

The JSB cohort is a population-based cancer research biobank containing pre-diagnostic serum samples from 318 628 Norwegians [75]. By linking data from the Cancer Registry of Norway [76] with the JSB cohort, we identified serum donors (n = 477) that were cancer-free at least 10 years after sample collection (male/female ratio: 2.13, average age at sampling: 49 years (range 19–77 years)). We do not have any information about non-malignant conditions. A previous study showed that miRNA (and other sncRNA) discovery is possible in long-term archived serum samples [7]. In addition to investigate technical variation, fresh serum from 6 individuals were pooled into one sample and divided into 17 aliquots. They were analysed as technical replicate samples. The downstream analyses were identical for all samples (Fig. S1). The donors have given broad consent for the use of the samples in cancer research. The study was approved by the Norwegian regional committee for medical and health research ethics (REC no: 2016/1290).

Laboratory processing

RNA was extracted from $2 \times 200 \ \mu$l serum using phenol-chloroform phase separation and the miRNeasy Serum/Plasma kit (Cat. no 1071073, Qiagen) on a QIAcube (Qiagen). Glycogen (Cat. no AM9510, Invitrogen) was used as carrier during the RNA extraction step. Small RNA-seq was performed using NEBNext® Small RNA Library Prep Set for Illumina (Cat. No E7300, New England Biolabs Inc.). Size selection was performed using a 3% Agarose Gel Cassette (Cat. No CSD3010) on a Pippin Prep (Sage Science) with a cut size optimized to cover RNA molecules from 17 to 47 nt in length. Sequencing libraries were indexed and 12 samples were sequenced per lane of a HiSeq 2500 (Illumina).

Bioinformatics analyses

The total number of reads generated was approximately 10 billion. The average sampling depths of the serum and technical replicate samples were 17.9 and 19.5 million raw reads, respectively. The reads were initially trimmed for adapters using AdapterRemoval v2.1.7 [77]. We then mapped the collapsed reads (generated by FASTX v0.14) to the human genome (hg38) using Bowtie2 v2.2.9 (10 alignments per read were allowed). We compiled a comprehensive annotation set from miRBase/MirGeneDB [74,78] for miRNAs, pirBAse/pirnabank for piRNAs [79,80], GENCODE [73] for other RNAs and tRNAs. We used SeqBuster [81] to get isomiR and miRNA profiles of our samples. To count the reads mapped on other RNAs, HTSeq [82] was utilized in a Python script. We used a threshold of 10 median read count per sncRNA to get a robust signal of expression. For longer transcripts (e.g. mRNA or lncRNA), we counted reads only mapped to exonic regions. However, this does not mean that the non-exonic mapped reads are not important. We are interested in bona fide fragments of longer genes but many non-exonic reads usually overlap with other short annotations, so it can be hard to determine their correct origin. Read counts were normalized to get reads per million (RPM) values. The coefficient of variation (CV) was calculated based on RPM values for the genes identified both in the serum and technical replicates in order to test biological and technical variation.

In order to get isoform and coverage profiles of tRNAs, we counted the reads mapped to tRNAs. There are 649 mature tRNA annotations available in GENCODE. We selected 41 tRNAs accounting for 99% of all reads mapped to tRNA annotations. The tRNAs were aligned to Rfam model (RF00005) using the *cmalign* tool 83] to get a multiple sequence alignment (MSA) of expressed tRNAs. Similar analyses were conducted for U3 snoRNAs and other misc-RNA (the models are RF00012, RF00006 and RF00019). Misc-RNAs denote RNA transcripts that are not classified into any other groups [73], which were taken from Rfam [84].

Data availability statement

The processed data is available upon request.

ORCID

Sinan Uğur Umu http://orcid.org/0000-0001-8081-7819
Hilde Langseth http://orcid.org/0000-0002-9446-4855
Cecilie Bucher-Johannessen http://orcid.org/0000-0003-3277-269X
Bastian Fromm http://orcid.org/0000-0003-0352-3037
Marianne Lauritzen http://orcid.org/0000-0002-0761-2558
Magnus Leithaug http://orcid.org/0000-0002-0271-0677
Trine B. Rounge http://orcid.org/0000-0003-2677-2722

References

[1] Danielson KM, Rubio R, Abderazzaq F, et al. High Throughput Sequencing of Extracellular RNA from Human Plasma. PLoS One. 2017;12:e0164644.

[2] Keller A, Leidinger P, Gislefoss R, et al. Stable serum miRNA profiles as potential tool for non-invasive lung cancer diagnosis. RNA Biol. 2011;8:506–16.

[3] Hornick NI, Huan J, Doron B, et al. Serum Exosome MicroRNA as a Minimally-Invasive Early Biomarker of AML. Sci Rep. 2015;5:11295.

[4] Kim KM, Abdelmohsen K, Mustapic M, et al. RNA in extracellular vesicles. Wiley Interdiscip Rev RNA [Internet]. 2017;8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28130830

[5] Inns J, James V. Circulating microRNAs for the prediction of metastasis in breast cancer patients diagnosed with early stage disease. Breast. 2015;24:364–9.

[6] Leidinger P, Backes C, Deutscher S, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 2013;14:R78.

[7] Rounge TB, Lauritzen M, Langseth H, et al. microRNA Biomarker Discovery and High-Throughput DNA Sequencing Are Possible Using Long-term Archived Serum Samples. Cancer Epidemiol Biomarkers Prev. 2015;24:1381–7.

[8] Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. 2008;105:10513–8.

[9] Arroyo JD, Chevillet JR, Kroh EM, et al. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A. 2011;108:5003–8.

[10] Chen X, Ba Y, Ma L, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res. 2008;18:997–1006.

[11] Yuan T, Huang X, Woodcock M, et al. Plasma extracellular RNA profiles in healthy and cancer patients. Sci Rep. 2016;6:19413.

[12] Freedman JE, Gerstein M, Mick E, et al. Diverse human extracellular RNAs are widely detected in human plasma. Nat Commun. 2016;7:11106.

[13] An T, Qin S, Xu Y, et al. Exosomes serve as tumour markers for personalized diagnostics owing to their important role in cancer metastasis. J Extracell Vesicles. 2015;4:27522.

[14] Nolte-'t Hoen ENM, Buermans HPJ, Waasdorp M, et al. Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions. Nucleic Acids Res. 2012;40:9272–85.

[15] Nomura S. Extracellular vesicles and blood diseases. International Journal of Hematology. 2017;105:392–405.

[16] Maierthaler M, Benner A, Hoffmeister M, et al. Plasma miR-122 and miR-200 family are prognostic markers in colorectal cancer. International Journal of Cancer. 2017;140:176–87.

[17] Ambros V. The functions of animal microRNAs. Nature. 2004;431:350–5.

[18] Chen X. MicroRNA metabolism in plants. Curr Top Microbiol Immunol. 2008;320:117–36.

[19] Umu SU, Gardner PP. A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life. Bioinformatics. 2017;33:988–96.

[20] Wang K, Yuan Y, Cho J-H, et al. Comparing the MicroRNA spectrum between serum and plasma. PLoS One. 2012;7:e41561.

[21] Flatmark K, Høye E, Fromm B. microRNAs as cancer biomarkers. Scand J Clin Lab Invest Suppl. 2016;245:S80–3.

[22] Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. Cell. 2012;148:1172–87.

[23] Telonis AG, Magee R, Loher P, et al. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. Nucleic Acids Res. 2017;45:2973–85.

[24] Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 2008;18:610–21.

[25] Llorens F, Bañez-Coronel M, Pantano L, et al. A highly expressed miR-101 isomiR is a functional silencing small RNA. BMC Genomics. 2013;14:104.

[26] Kiss T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell. 2002;109:145–8.

[27] Klattenhoff C, Theurkauf W. Biogenesis and germline functions of piRNAs. Development. 2008;135:3–9.

[28] Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006;442:199–202.

[29] Lee YS, Shibata Y, Malhotra A, et al. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 2009;23:2639–49.

[30] Zheng L-L, Xu W-L, Liu S, et al. tRF2Cancer: A web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. Nucleic Acids Res. 2016;44:W185–93.

[31] Goodarzi H, Liu X, Nguyen HCB, et al. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. Cell. 2015;161:790–802.

[32] Pennisi E. Genomics. ENCODE project writes eulogy for junk DNA. Science. 2012;337:1159–1161.

[33] Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.

[34] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

[35] Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Front Genet. 2015;6:2.

[36] Clark MB, Amaral PP, Schlesinger FJ, et al. The reality of pervasive transcription. PLoS Biol. 2011;9:e1000625;discussion e1001102.

[37] Pauli A, Valen E, Schier AF. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. Bioessays. 2015;37:103–12.

[38] Tuck AC, Tollervey D. RNA in pieces. Trends Genet. 2011;27:422–32.

[39] Scott MS, Ono M. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. Biochimie. 2011;93:1987–92.

[40] Röther S, Meister G. Small RNAs derived from longer non-coding RNAs. Biochimie. 2011;93:1905–15.

[41] Tzimagiorgis G, Michailidou EZ, Kritis A, et al. Recovering circulating extracellular or cell-free RNA from bodily fluids. Cancer Epidemiol. 2011;35:580–9.

[42] Tonge DP, Gant TW. What is normal? Next generation sequencing-driven analysis of the human circulating miRNAOme. BMC Mol Biol. 2016;17:4.

[43] Keller A, Leidinger P, Bauer A, et al. Toward the blood-borne miR-Nome of human diseases. Nat Methods. 2011;8:841–3.

[44] Kahraman M, Laufer T, Backes C, et al. Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase. Clin Chem [Internet]. 2017; Available from: http://dx.doi.org/10.1373/clinchem.2017.271619

[45] Keller A, Rounge T, Backes C, et al. Sources to variability in circulating human miRNA signatures. RNA Biol. 2017;1–8.

[46] Lopez JP, Diallo A, Cruceanu C, et al. Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. BMC Med Genomics. 2015;8:35.

[47] Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;5:463.

[48] Zhang WC, Chin TM, Yang H, et al. Tumour-initiating cell-specific miR-1246 and miR-1290 expression converge to promote non-small cell lung cancer progression. Nat Commun. 2016;7:11702.

[49] Neilsen CT, Goodall GJ, Bracken CP. IsomiRs-the overlooked repertoire in the dynamic microRNAome. Trends Genet. 2012;28:544–9.

[50] Thompson DM, Parker R. Stressing out over tRNA cleavage. Cell. 2009;138:215–9.

[51] Saikia M, Jobava R, Parisien M, Putnam A, Krokowski D, Gao X-H, Guan B-J, Yuan Y, Jankowsky E, Feng Z, et al. Angiogenin-cleaved tRNA halves interact with cytochrome c, protecting cells from apoptosis during osmotic stress. Mol Cell Biol. 2014;34:2450–63.

[52] Maute RL, Schneider C, Sumazin P, et al. tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. Proc Natl Acad Sci U S A. 2013;110:1404–9.

[53] Kumar P, Anaya J, Mudunuri SB, et al. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily

conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. 2014;12:78.

[54] Pliatsika V, Loher P, Telonis AG, et al. MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. Bioinformatics. 2016;32:2481–9.

[55] Persson H, Kvist A, Vallon-Christersson J, et al. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. Nat Cell Biol. 2009;11:1268–71.

[56] Kowalski MP, Krude T. Functional roles of non-coding Y RNAs. Int J Biochem Cell Biol. 2015;66:20–9.

[57] Mosig A, Guofeng M, Stadler BMR, et al. Evolution of the vertebrate Y RNA cluster. Theory Biosci. 2007;126:9–14.

[58] Rutjes SA, van der Heijden A, Utz PJ, et al. Rapid nucleolytic degradation of the small cytoplasmic Y RNAs during apoptosis. J Biol Chem. 1999;274:24799–807.

[59] Izquierdo MA, Scheffer GL, Schroeijers AB, et al. Vault-related resistance to anticancer drugs determined by the expression of the major vault protein LRP. Cytotechnology. 1998;27:137–48.

[60] Kedersha NL, Miquel MC, Bittner D, et al. Vaults. II. Ribonucleoprotein structures are highly conserved among higher and lower eukaryotes. J Cell Biol. 1990;110:895–901.

[61] van Zon A, Mossink MH, Schoester M, et al.. Multiple human vault RNAs. Expression and association with the vault complex. J Biol Chem. 2001;276:37715–21.

[62] Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014;505:344–52.

[63] Kulcheski FR, Christoff AP, Margis R. Circular RNAs are miRNA sponges and can be used as a new class of biomarker. J Biotechnol. 2016;238:42–51.

[64] Pircher A, Bakowska-Zywicka K, Schneider L, et al. An mRNA-derived noncoding RNA targets and regulates the ribosome. Mol Cell. 2014;54:147–55.

[65] Rogler LE, Kosmyna B, Moskowitz D, et al. Small RNAs derived from lncRNA RNase MRP have gene-silencing activity relevant to human cartilage-hair hypoplasia. Hum Mol Genet. 2014;23:368–82.

[66] McIntyre LM, Lopiano KK, Morse AM, et al. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12:293.

[67] Backes C, Sedaghat-Hamedani F, Frese K, et al. Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. Anal Chem. 2016;88:2088–95.

[68] Tarazona S, García-Alcalde F, Dopazo J, et al. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21:2213–23.

[69] Marioni JC, Mason CE, Mane SM, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509–17.

[70] Todd EV, Black MA, Gemmell NJ. The power and promise of RNA-seq in ecology and evolution. Mol Ecol. 2016;25:1224–41.

[71] Zhu W, Qin W, Atasoy U, et al. Circulating microRNAs in breast cancer and healthy subjects. BMC Res Notes. 2009;2:89.

[72] Ziemann M, Kaspi A, El-Osta A. Evaluation of microRNA alignment techniques. RNA. 2016;22:1120–38.

[73] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.

[74] Fromm B, Billipp T, Peck LE, et al. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. Annu Rev Genet. 2015;49:213–42.

[75] Langseth H, Gislefoss RE, Martinsen JI, et al. Cohort Profile: The Janus Serum Bank Cohort in Norway. Int J Epidemiol. 2017;46:403–4.

[76] Larsen IK, Småstuen M, Johannesen TB, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009;45:1218–31.

[77] Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

[78] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73.

[79] Zhang P, Si X, Skogerbø G, et al. piRBase: a web resource assisting piRNA functional study. Database. 2014;2014:bau110.

[80] Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res. 2008;36:D173–7.

[81] Pantano L, Estivill X, Martí E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Res. 2010;38:e34.

[82] Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.

[83] Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.

[84] Nawrocki EP, Burge SW, Bateman A, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43:D130–7.