5´XP sRNA-seq: A simple sequencing method to identify small RNA with and without 5´ phosphorylation in a single library from low input samples

U. Kugelberg[1#], D. Nätt[1#], S. Skog[1], C. Kutter[2], A. Öst[1]*

[1]Department of Clinical and Experimental Medicine, Linköping University, 581 85 Linkoping, Sweden

[2]Department of Microbiology, Tumor and Cell Biology, Science for Life Laboratory, Karolinska Institute, 171 77, Stockholm

* Correspondence to: anita.ost@liu.se

# Equal contribution

SUMMARY

Small RNA (sRNA) sequencing has been critical for our understanding of the transcriptional regulation in most biological processes. Nonetheless, the varying biochemical properties of sRNA, such as 5´ nucleotide modifications, make many sRNA subspecies incompatible with common protocols for sRNA sequencing. Here we describe 5XP-seq that outlines a novel strategy on how to overcome this limitation. By specifically tagging 5´-phosphorylated sRNA (5´P sRNA), that enriches miRNA and piRNA, we successfully discriminate these sRNA from all other sRNA species. Our results underscore the importance of more holistic approaches, in which 5´ insensitive library strategies are combined with 5´ sensitive tagging. Further advances using such strategies will overcome the increasing bias in reports of easily caught sRNA and ensure assessment of the previously overlooked transcriptome.

INTRODUCTION

High-throughput RNA-sequencing techniques have revolutionized our understanding of various small noncoding RNA (sRNA) species, typically with an arbitrarily defined length of less than 200 nucleotides. Numerous studies have provided insight into the indispensable functions of sRNAs in regulating cellular processes, such as cell development, differentiation and proliferation under physiologically normal and pathological conditions [1, 2]. In addition to well-characterized micro (miRNA), small interfering (siRNA) and PIWI-interacting RNAs (piRNA)[3, 4], new sRNA species deriving from transfer (tRNA), ribosomal (rRNA) [5], small nucleolar (snoRNA), vault (V RNA) and Y RNA [6] exists in many cell types including the transcriptionally dormant sperm [7]. These fragments are not just by-products of pervasive transcription or random RNA degradation but are precisely engineered to exert specialized regulatory functions. For example, specific fragments generated from tRNAs (tsRNA) have been implicated in translational control and cell growth [8]. The regulatory repertoire of sRNA in general, and tsRNA specifically, is expanded by the addition of small molecular modifications including methylations and pseudouridylation. Such modifications effect RNA stability, three-dimensional structure, cellular location and interactions with proteins [9-11].

A major limitation when exploring the sRNA world is the biased enrichment of different sRNA species introduced during library preparation that are caused by diverse chemical modifications at the 5´- and 3´- ends of RNA. Hydroxyl (OH) or phosphate (P) groups can be found either in the 3´- or 5´-end depending on the processing pathway, while a 2',3´-cyclic phosphate (cP) group is only present on the 3´-end [12]. In addition, sRNAs can have multiple types of 5´-caps such as mGpppC, 7mGpppG, GpppG, GpppA, and 7mGpppA in human cells [13] or NAD+ [14] and 3′-dephospho-CoA, succinyl-dephospho-CoA and acetyl-dephospho-CoA in bacteria [15].

Most high-throughput sequencing methods require the ligation of sequencing-compatible adapters first to a hydroxyl-group at the 3´-end (3´-OH) and then to a phosphate group at the 5´-end (5´-P) of the sRNA molecule. These chemical reactions therefore select for sRNA containing these very specific terminal modifications, that are subsequently PCR amplified and finally detected during sequencing.

Here, we describe 5´XP sRNA-seq, a novel sequencing method with a unique tagging system that allows the detection of both sRNA with a 5´-P (typically miRNA and piRNA) and sRNA with alternative 5´ terminal groups. We show that this method increases specificity towards miRNA and piRNA when compared to a commonly used strategy, while simultaneously allowing for additional fragments to be detected in the same library. Overall, we show that 5´XP sRNA-seq is an adequate strategy for tackling critical biases commonly prevalent in standard next generation sRNA library preparations.

RESULTS

5´P, 5´X and 5´XP sRNA-seq employ distinct library preparation strategies

5´-P and 3´-OH dependent ligations amplify some sRNA species, including piRNA and miRNA, but will exclude functional sRNA species without a 5´-P. We refer to this ligation strategy as 5´-P dependent sRNA-seq (from here on: 5P-seq). This should not be mistaken for 5´-P dependent methods recently developed for long RNA sequencing [16-18]. It has been shown that the 5´P-ligation is sensitive to secondary

structures of the RNA, which further restricts capturing a broad range of sRNA species (for review see [19]). To include highly structured RNA species without 5´-P ends, strategies have been developed in which the 5´ adapter is not ligated to the RNA but is added later to the cDNA [20, 21]. We here refer to such methods as 5´ inclusive sRNA-seq (5X-seq). Building on these two approaches we developed 5´XP sRNA-seq (5XP-seq). In this method, the first adapter is ligated to the 3´ end and a 11-nt long oligo is ligated to the 5´ end of sRNA, which allows efficient tagging of RNAs with 5´-P. Then, after reverse transcription of the RNA ligation product, a sequencing-compatible adapter is ligated to the corresponding 3´ end of all cDNA (Fig 1, Supplementary Fig. 1). Through this approach sRNA with 5´-P and 5´-X can be sequenced in the same library and then bioinformatically discerned based on the presence or absence of the 5´ oligonucleotide tag.

5P and 5X sRNA-seq generate libraries with substantially different sRNA content

To better understand how 5P and 5X sRNA-seq vary in their enrichment for different RNA species, we first compared libraries generated using a popular commercial 5P-seq kit (NEBNext Multiplex Small RNA Library prep kit for Illumina; New England Biolabs) and an adaptation of published 5X-seq protocols [20-23]. For comparative reasons, we modified the 5X-seq protocol to work with the reverse transcriptase of the 5P-seq kit. We extracted total RNA from a pool of 50 *Drosophila* embryos, 0.5-2.5 hours old, which was used as input for all sequencing library preparations. Each preparation used only 30 ng of total RNA that equals to one third of the RNA extracted from one single drosophila embryo, rendering the protocols highly efficient. We adopted sequence-based counting of the reads, for which statistical units are based on counts per sequence and not counts per feature (see methods for details). An annotated count table normalized against total library sizes as counts per million (cpm) are presented in Supplementary Table 1. Raw data is available at Sequence Read Archive under the accession number PRJNA658107.

Library stats for the 5P- and 5X-type libraries using single end sequencing for 75 cycles are found in Table 1. Data from 5P-seq libraries were in accordance with our previous observations using this kit. To better validate the 5X-seq protocol, we generated a second batch of 5X-seq using independent embryos (Batch B).

Focusing on intermediately-to-strongly expressed sRNA (>10 cpm), 5P-seq resulted in more alignments against miRNA and piRNA, and less alignments against rRNA and tRNA, when compared to 5X-seq (Fig 2). This result was similar for both cumulative counts (Fig 2A) and relative fold changes (Fig 2B), indicating a general trend across most sRNA. This difference is expected since miRNA and piRNA are commonly 5´ phosphorylated and should be preferably enriched by 5P-seq, while tRNA and rRNA show more diversity in the 5´-end and should preferably be caught by 5X-seq. When not allowing mismatches in the annotation 5X-seq annotated better to the references than 5P-seq (light grey Fig 2A-B). The advantage for applying sequence-based counting is that mismatches in the alignments between sequenced reads and small RNA reference sequences only affect the annotation without changing the raw count table, which is often the case for feature-based counting. Allowing up to 3 mismatches when annotating sequences against small RNA references removed nearly all sequences with no annotation in 5P-seq (light grey Fig 2C-D). Since libraries were generated from the same RNA pool, this indicates that 5P-seq may have suffered from misreading during cDNA-synthesis possibly by interference from RNA-modifications. Removing sequences mapping to rRNAs, however, also removed this difference (Supplementary Fig 2).

While the majority of sRNA identified by 5X-seq were also present in 5P-seq libraries, some sRNAs were completely absent in one or the other library, which resulted in a more than 20-fold change (Fig 2B and 2D; Supplementary Table 1). This indicates substantial differences in the composition of sRNA species identified by the two different library strategies. Therefore, studies that have applied only 5P-based sRNA-seq strategies may have misinterpreted the proportions of small RNA present in their samples.

5XP sRNA-seq distinguishes piRNA and miRNA

To preserve sequence diversity given the observed biases described above, in parallel to the 5P and 5X-seq libraries, we generated 5XP-seq libraries using the same two pools of RNA from *Drosophila* embryos (Batch A and B). Table 1 summarizes the general properties of 5XP-seq after being sorted into the two sub-libraries, either containing (5P-tag) and not containing (5X-notag) the 5´ oligo tag. In

theory, 5P-tag and 5X-notag of 5XP-seq would correspond to the 5P-seq and 5X-seq libraries, respectively.

Similar to the difference observed between 5P- and 5X-seq, 5XP-seq had more specificity towards miRNA and piRNA in the 5P-tag sub-libraries, than in the 5X-notag sub-libraries (Table 1). In fact, sRNA diversity was highly similar between 5X-notag sub-libraries and regular 5X libraries prepared without tagging 5´-P (Table 1). In support, expression levels of 5X-seq and 5X-notag cohered strongly, both by correlation and unsupervised clustering (Fig 3A).

In contrast, we determined higher specificity (~ 2-fold) of miRNA and piRNA species in 5P-tag sub-libraries when compared to regular 5P-seq libraries (Table 1). This indicates that ligating a short oligonucleotide tag at the 5´-terminal—as performed in 5XP-seq—captures 5´-P sRNAs more efficiently when compared to the long 5´ adapter ligation used in regular 5P-seq. Nonetheless, the 5P-tag sub-libraries only contributed with 10% to 20% of the total library sizes in 5XP-seq (Table 1) and correlated poorly with regular 5P-seq (Fig 3A). This led to two possible explanations. First, 5´-P sRNA may constitute a smaller proportion of the total sRNA population. Secondly, 5XP-seq may suffer from inefficient ligation of the 5´ oligo tag. In support of the former, a saturation analysis showed that the 5P-tag sub-libraries were equally saturated to the other libraries. Thus, irrespective of library-type, increasing the sequencing depths would result in little gain in identifying new sRNA specimens from the sRNA population targeted by each method (Supplementary Fig 3).

To further test the hypothesis that the 5P-tag sub-library of 5XP-seq targets a smaller and possibly more specific population of 5´-P sRNAs, we characterized each library type in more details. Size distribution of different sRNA species confirmed a sharp peak at 22 nt in 5P-tag sub-libraries, while the other library strategies showed more diversity (Fig 3B). The peak at 22nt is the typical size of miRNA. However, we detected that that many sequences of 22nt also mapped to rRNAs. Therefore, we extracted miRNA and piRNA that are known 5´-P subspecies for further comparison. Both the 5P-tag sub-libraries and regular 5P libraries where enriched with a sharp 22-23 nt miRNA peak, and a broader 22-29 nt piRNA peak (left panel Fig 3C). However, the 5P-tag sub-libraries of 5XP-seq had fewer non-canonical piRNA

fragments that were shorter than 21 nt and showed a higher proportion of canonical piRNA with uridine bias at their first nucleotide (middle panel Fig 3C; [24]). Interestingly, the first nucleotide uridine bias was also observed in miRNA (right panel Fig 3C), which could indicate a new subspecies of miRNA in early drosophila embryos.

A combination of a lower proportion of unannotated small RNA (light grey in Fig 3B and 2A), as well as a smaller target population of sRNA (Supplementary Fig 3) and more canonical piRNA profiles (Fig 3C), altogether indicated that the 5P-tag sub-library of 5XP-seq enriches for 5´-P RNA better than regular 5P-seq. The low correlation in the expression profiles observed between these two strategies (Fig 3A) was also rescued (Fig 3D) after noise from less studied subspecies of sRNA was reduced by analyzing miRNAs only, which represent the most well-characterized 5´-P sRNA species.

Library strategies are defined by opposite terminal enrichments of small rRNA

Loss of peak integrity of rRNA subunits is often used as an indicator of RNA degradation in studies of long RNA, often referred to as the RNA integrity number (RIN). While a similar measurement is lacking for sRNA, all our libraries showed very sharp peaks indicating specifically processed rRNA with high sRNA integrity (Fig 4A-B).

The 5P-tag sub-libraries in 5XP-seq were dominated by a peak centered at 22 nt (Fig 3B), which primarily contained an rRNA fragment (rRF) originating from the 5´-end of 18S rRNA subunit (Fig 4A). While this rRF is not classified as a miRNA in mirBase, recent findings in Zebrafish show that it may interact with Argonaute proteins in a miRNA-like matter [5, 25]. Low detection of this rRF in regular 5P-seq (Fig 4B) would have been overlooked in other studies. Nonetheless, unlike 5X-type libraries that showed very similar rRF profiles, 5P-type libraries were at first dissimilar (Fig 4A-B). A more detailed view on the rRNA subunits revealed, however, that most peaks were present in both the 5P-tag and regular 5P libraries but varied in detection level (Fig 4C-E). Together this indicates that the 5´ RNA ligation strategy—by either using long or short oligos/adapters—severely affects the enrichment of specific rRF, which could make cross-study validation challenging.

We also noticed a 5´ bias in 5P-type and a 3´ bias in 5X-type libraries (Fig 4A-B), and therefore classified each fragment by their 5´ start or 3´ end of a mature rRNA subunit. This confirmed that rRFs generated from the 5´ terminals were enriched in 5P-seq, while rRFs deriving from the 3´ terminals were enriched in 5X-seq (Fig 4F). Importantly, 5XP-seq contained rRFs originating from both the 5´ and 3´ terminal but kept them separated in the 5P-tag and 5X-notag sub-libraries, respectively. These differences indicate that 5P-type libraries lose sRNA primarily at the 3´ end, possibly as a consequence of interference with RNA-modifications during reverse transcription. Interestingly, 5´ rRF bias has recently been reported in humans using an independent 5P-seq kit {Nätt, 2020 #2760;Hua, 2019 #2565}.

Coverage of small tRNA fragments strongly depends on library preparation strategy.

Like rRFs, tRNA derived sRNA (tRFs) can also originate from the 5´ or 3´ terminal. In addition, internal (i´) fragments, which neither start nor end in the terminals of full-length tRNA have been described [26, 27]. We used these classifications as an independent confirmation of the differences in 5´ and 3´ affinities between the 5P- and 5X-type libraries that we observed in sRNA derived from rRNA. As expected, 5P-seq had a clear 5´ preference, while 5X-seq progressively increased towards the 3´ terminal (Fig 5A). Similar to the rRFs, 5XP-seq contained tRFs originating from both the 5´ and 3´ terminal and successfully separated them into the two sub-libraries.

In addition to 5´, i´ and 3´ stratification, we classified each tRF according to the isodecoder of the mature tRNA in order to better understand the possible biases in the tRF coverage introduced by different library strategies. This enables an overview of tRF subtypes, such as tRNA derived stress-induced small RNA (tiRNA) that are generated by angiogenin dependent anticodon cleavage of specific isodecoders resulting in tRNA halves [28], as well as the recently discovered nuclear internal T-loop tRFs (nitRNA) [7]. As expected, 5X-notag sub-libraries generated primarily i´ and 3´ fragments that was strongly replicated by regular 5X-seq, but poorly replicated by 5P-seq (Fig 5B). Again, while the most covered tRFs in the 5P-tag sub-libraries of 5XP-seq showed similarities to regular 5P-seq, we found more variability on the isodecoder level (Fig 5C). In highly abundant isodecoders, there were also an

enrichment of sRNA species derived from the 5´ end of tRNAs but this was less pronounced in less abundant isodecoders.

In line with the rRNA analysis, more in-depth analysis of a selection of tRNA isodecoders showed that most tRFs were detected by each method but detection levels varied greatly across methods. For example, the 5P-tag sub-library was superior in detecting both the 5´and 3´ halves generated by angiogenin-mediated cleavage of AspGUC, which generates a known tiRNA (Fig 5D). In another confirmed tiRNA originating from GlyGCC, the 5´-half was detected in both 5P-tag and regular 5P libraries (Fig 5E). At the 3´-end a short fragment generated by T-loop cleavage were detected in the 5P-tag, 5X-notag and regular 5X libraries but to a much lesser degree in regular 5P libraries (Fig 5E). LysCUU has previously been shown to generate many different tRFs [7, 29]. This was confirmed by 5X-notag, regular 5X and regular 5P, but to a lesser degree in 5P-tag libraries (Fig 5F). Interestingly, primarily 5X-notag and regular 5X libraries detected a weakly expressed long T-loop internal tRFs comparable to the nitRNA that was recently detected in human sperm [7]. Similarly, 5X-notag and regular 5X, but not 5P-tag and regular 5P, detected a short nitRNA in ArgCCU generated by cleavage in the T- and anticodon- loops that was almost identical to the one previously reported in human sperm [7].

Together these findings show that the choice between 5P-sensitive and 5P-insensitive methods are critical when studying tRFs using sequencing. Thus, using a method that conserves the diversity of both phosphorylated and non-phosphorylated sRNA—like 5XP sRNA-seq—is highly desirable.

DISCUSSION

Here we presented a novel method—5XP sRNA-seq—that sequences sRNA with and without 5´ phosphorylation (5´-P) using the same library. Our protocol was optimized for low starting material which makes it attractive for use with precious samples. By ligating a short 5´-P oligonucleotide tag, we demonstrated that 5´-P sRNAs can be separated from other sRNAs in the downstream analysis. We reported multiple examples in which a one-sided approach would have had severe consequences for the interpretation and cross-validation of sRNA experiments. We

also provided evidence that ligating a short oligonucleotide tag enhances the recovery of 5´-P sRNAs, such as piRNA and miRNA.

It has become increasingly clear that the 5´ terminal of RNA is subject to diverse modifications. As this is the start site for transcription, it means that if the first molecule is a nucleotide, it will initially have a triphosphate. If transcribed by RNA polymerase II, to become for example mRNA, a cap consisting of modified guanine (G) nucleotide is first added to the initial nucleotide during transcription. Primary piRNAs are also transcribed by RNA polymerase II and receive a similar cap. These 5´ caps regulate the stability and determine the downstream fate of the transcript. The process of generating sRNAs from longer precursors by nuclease digestion will result in either -P or -OH terminals, depending on the nuclease. Intriguingly, it was recently shown that sRNAs are capped as well, indicating that not all caps are added during transcription but rather after the cleavage to shorter transcripts [13, 30]. In addition to the guanine-based caps, recent discoveries show that the sRNA may have other types of 5´-caps.

First reported in bacteria, and suggested to be an alternative to the eukaryotic m^7G cap, 5´ NAD caps have now been demonstrated in several species including humans (see review: [31]). Interestingly, NAD^+, as well as NADH and dpCoA, can be incorporated into RNA during transcription initiation, thus serving as non-canonical initiating nucleotides [32]. In human cells, the amount of RNA with 5´-NAD caps change in response to shifting cellular NAD concentrations [33], while data from Arabidopsis shows enrichment of NAD-caps for transcripts involved in redox responses [31]. These two findings, together with the fact that NAD+/NADH is one of the most important intracellular redox pairs, suggests that NAD-capping of RNA may regulate gene expression as a function of the cell's redox state. Moreover, members of the NUDIX hydrolase superfamily have been shown to remove not only 5´-NAD, but also 5´-FAD and 5´-CoA, indicating that metabolite-containing 5´ caps might be a more widespread phenomenon than previously thought [34]. Whether NAD or other metabolic 5´-caps are used to regulate sRNA, however, is yet to be discovered.

These recent findings demonstrate the urge to develop new sequencing methods that not only include 5´ modifications other than phosphate, but also correctly identify

each modification in downstream analyses. By using RNA ligases with specificity toward different 5´ terminals together with unique oligonucleotide tags, 5´XP sRNA-seq is a first step towards a more holistic sequencing approach. By tagging 5´-P in this way, our data suggest that 5´-P sRNAs constitute only 10-20% of the total pool of sRNA in drosophila embryos. Therefore, up to 90% of sRNA species may have previously escaped our attention since most commercial kits are based on 5´-P dependent adapter ligation. Using 5´XP sRNA-seq on RNA from *Drosophila* embryos, we discovered several new sRNA species without 5´-P, including cross-species validation of a zebrafish 5´-rRF from the 18S rRNA subunit [25], and nitRNA derived from internal tRNA T-loop cleavage, previously described only in human sperm [7].

Our results represent an important step toward exploring the complete world of sRNAs in a single experiment. Future technologies will be aimed at further expanding the number of specific terminal RNA modifications that can be identified by genome-wide sequencing approaches. We anticipate discovering a rich repertoire of functional RNA modifications that will greatly expand our understanding of how genomes are regulated.

Methods

Experimental Methods

Reagents and Oligos

Many of the reagents used in 5XP-seq are included in the NEBNext Multiplex Small RNA Library prep kit for Illumina (New England Biolabs), and are referred to as 'NEB-kit' below. Since we used lower input (30 ng) than the recommended 100 ng total RNA, we diluted the 3´SR Adapter, 5´SR Adapter, and SR RT Primer 1:2 in nuclease free H2O. The SR Primer and all Index Primers used for amplification, were used in original concentrations.

Custom oligos not included in the NEB-kit were all HPLC-purified:
5P-tag RNA oligo (long: 5´-UGGCAACGAUC-3´; short: 5´-UGGGAUC-3´; both 3.75 µM).

R1R DNA Adapter (5´ Phos-GAT CGT CGG ACT GTA GAA CTC TGA ACG TGT AG-SpcC3 3´; 100 µM). Adenylation of R1R was done using 5´ DNA Adenylation kit (New England Biolabs) according to manufacturer's instructions and cleaned with Oligo Clean & Concentrator (Zymo Research) as described here [23]. Oligos was eluted and diluted with nuclease-free water to 10 µM.

Illumina multiplex PCR primer (5´ -AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TTC AGA GTT CTA CAG TCC GAC GAT C- 3´; 10 µM)

Since *Drosophila melanogaster* expresses large amounts of 2S rRNA in the same size as many sRNA, we blocked this transcript by adding anti-sense oligos at the 5´ RNA ligation (First), and at the cDNA ligation steps (Second). For species with less 2S rRNA, these blocking oligos can be exchanged for nuclease-free H2O or custom oligos targeting other transcripts that may occupy a large amount of the sequencing capacity.

First 2SrRNA block oligo (5´-TAC AAC CCT CAA CCA TAT GTA GTC CAA GCA-SpcC3 3´; 10 µM) [35].

Second LNA 2SrRNA block oligo (5´- TGC-**T**TG-GAC-**T**AC-ATA-**T**GG-TT**G**-AGG-**G**TT-G**T**A-SpcC3 3´; 10 µM, where bold letter indicate LNA incorporation).

RNA Extraction from Drosophila Embryos

RNA was isolated from *Drosophila melanogaster* (W1118) embryos 0.5-2.5h of age. Embryos were decorionated in pools of 50 embryos using sodium hypochlorite 3.5% (RECTAPUR) and washed extensively with RNase free H2O. RNA was isolated using the miRNeasy Micro kit (Qiagen) according to manufacturer's instructions. In brief, 500 µl Qiazol (Qiagen) was added to the embryos and homogenized for 2 min at 40 Hz using a Tissue Lyser LT (Qiagen) and 5 mm stainless steel beads (Qiagen). Phase separation was done by mixing 100 µl of Chloroform followed by centrifugation at 12000xg for 15 min in 4ºC. RNA was then collected by columns, washed and eluted in 14 µl nuclease free H2O. RNA integrity was confirmed by Bioanalyzer (Agilent) and concentrations were determined by Nanodrop (ThermoFisher).

5´XP sRNA-seq Library Preparation

A side-by-side comparison of the workflows of the different methods used in this study is available in Supplementary Fig 1.

The 5´XP sRNA-seq method includes the following steps:

i.	Ligation of 3´ RNA adapter	18.5 h	Proceed immediately (over-night)
ii.	Ligation of 5´ RNA tag	2.5 h	Proceed immediately
iii.	cDNA synthesis	2.5 h	Pause point
iv.	Ligation of 5´ cDNA adapter	3.0 h	Pause point
v.	Barcoding & Amplification	1.5 h	Pause point
vi.	Size selection	>12 h	Pause point (over-night)
vii.	Pooling barcoded samples	3 h	Pause point

In detail, the 3´ adapter was first ligated using 0.5 µl of 3´SR Adapter for Illumina (NEB-kit) added to 30 ng of input RNA diluted in 3.0 µl nuclease free H_2O. Samples were heated to 70°C for 2 min and immediately cooled on ice. A mix of 5 µl 3´Ligation

Reaction Buffer 2x (NEB-kit) and 1.5 µl of 3´Ligation Enzyme Mix (NEB-kit) were added and incubated for 18 h at 16°C over-night. A 2S rRNA blocking oligo was added simultaneous to hybridization of the reverse transcription primer by adding a mix of 0.5 µl of First 2SrRNA block oligo, 0.5 µl of SR RT primer (NEB-kit) and 1.75 µl of nuclease-free water to each sample followed by incubation at 90°C for 30 s, 65°C for 5 min, 37°C for 15 min, and finally 25°C for 15 min. 5´-tagging was then performed by adding 0.5 µl 5P-tag RNA oligo (denatured at 70°C for 2 min), 0.5 µl 5´Ligation Reaction Buffer (NEB-kit) and 1.25 µl of 5´Ligation Enzyme Mix (NEB-kit) followed by a 1 h incubation at 25°C. To synthesis cDNA, a mix of 4 µl First strand Synthesis Reaction Buffer (NEB-kit), 0.5 µl Murine RNase Inhibitor (NEB-kit) and 0.5 µl ProtoScript II RT (NEB-kit) was added to each sample, and incubated first at 50°C for 1 h, and then at 70°C for 15 min. Enzymatic inactivation and hydrolysis of RNA was done by incubating samples at 95°C for 3 min with 17.5 µl nuclease-free water, 10 µl 0.5M EDTA and 2.5 µl 5M NaOH. After room temperature equilibration pH was lowered by adding 1 µl 5M HCl. Samples were then cleaned using Oligo Clean & Concentrator kit (Zymo Research) according to manufacturer's recommendation but with an extra wash in 250 µl DNA wash buffer (Zymo Research) and cDNA elution in 11 µl of nuclease-free water. Ligation of pre-adenylated R1R adapter with 2S RNA blocking was done by adding 10 µl of cDNA, 2 µl 10x NEBuffer 1 (New England Biolabs), 2 µl 50 mM MnCl2, 2 µl Thermostable 5´App ligase (New England Biolabs), 4 µl 10 µM pre-adenylated R1R Adapter, and 0.5 µl 10 µM Second LNA 2SrRNA block oligo, followed by 2 h incubation at 65°C. Ligated cDNA samples were cleaned and eluted in 24 µl Elution Buffer (Qiagen) using Oligo Clean & Concentrator (Zymo Research) according to manufacturer's instructions. Clean samples were amplified by adding 25 µl 2x Phusion High-Fidelity PCR master mix (ThermoFisher), 1 µl 10 µM Illumina Multiplex Primer, 1 µl 10 µM Illumina barcode (NEB-kit) to 23 µl of cDNA, and incubated in a thermocycle at 98°C for 5 s, followed by 15 cycles of 98°C for 5 s, 60°C for 10 sec, 72°C for 30s, ending with 72°C for 1 min. Amplified libraries were cleaned using Agencourt AMPure XP (Beckman Coulter), and size selected for 130-200 nt fragments on a pre-casted 6% polyacrylamide Novex TBE gel (Invitrogen). Gel extraction was done using Gel breaker tubes (IST Engineering Inc) in DNA Gel Elution Buffer (NEB-kit). Disintegrated gels were incubated at 37°C for 1h on a shaker, quickly frozen for 15 min at -80°C, followed by another incubation for 1 h. Remaining gel debris were removed by Spin-X 0.45 µm centrifuge tubes. The

libraries were precipitated overnight at -80°C by adding 1 µl of GlycoBlue (Invitrogen) as co-precipitant, 0.1 times the volume of Acetate 3M (pH 5.5), and three times the volume of 100% ethanol. Library concentrations were estimated using QuantiFluor ONE ds DNAsystem on a Quantus fluorometer (Promega). Pooled libraries were sequenced on NextSeq 500 with NextSeq 500/550 High Output Kit v2.5, 75 cycles (Illumina).

5´P sRNA-seq Library Preparation

Library preparation was done with NEBNext Multiplex Small RNA Library prep kit for Illumina (New England Biolabs) according to the manufactures instructions except for downscaling all samples to half volume and using 30 ng of input RNA (100 ng recommended) with an appropriate 1:2 dilution of all adapters. This corresponds to the 5XP-seq preparation as above, but with the following changes:

In Step [ii.], 5´ RNA ligation was performed using 0.5 µl 5´ SR Adapater (NEB-kit) instead of the 5P RNA oligo.

In Steps [iii-iv.], clean up of cDNA, ligation of the 5´ DNA adapter and the Second LNA 2SrRNA oligo block was not performed. The cDNA reaction mix was immediately amplified according to kit protocol.

In Step [v.], 20 µl cDNA reaction mix was amplified by adding 25 µl of Long Amp Taq 2x Master Mix (NEB-kit), 1.25 µl SR primer for Illumina (NEB-kit), 2.5 µl nuclease-free H_2O, 1.25 µl 10 µM Illumina barcode (NEB-kit) and incubated, starting with 94°C for 30 s, followed by 15 cycles of 94°C for 15 sec, 62°C for 30 sec, 70°C for 15 sec, ending with 70°C for 2 min.

5´X sRNA-seq Library Preparation

5X-seq library preparation was identical to 5XP-seq library preparation, except for the following adjustments:

In Steps [ii-iii.], 5´ RNA ligation of the 5P-tag RNA oligo and the First 2SrRNA block was not performed. The cDNA reaction mix was therefore compensated by adding 0.5 µl 5´Ligation Reaction Buffer 10x (NEB-kit) and 1.75 µl nuclease-free H_2O.

Computational Methods

Quality Control and Pre-processing

Raw fastq data files have been deposited in Sequence Read Archive under the accession number PRJNA658107. All libraries passed Illumina's default quality control. Demultiplexed fastq files were downloaded from BaseSpace using BaseMount (Illumina) and lanes were merged for each sample. For all library-types the 3´ adapters were trimmed using Cutadapt 2.3 [36] with following input: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAT --discard-untrimmed --nextseq-trim=20 -O 5 -m 5. This specifically trims nextseq type sequences between 5-70 nt with at least 5 nt of the 3´ adapter present in the 3´ end and discarding all other sequences. For 5XP-seq libraries we reloaded the trimmed sequences into Cutadapt using the following input: -g TGGCAACGATC -m 5 --untrimmed-output. This saves sequences with or without the 5´-tag in separate fastq output files. All trimmed fastq files were quality filtered using fastq_quality_filter -q 20 -p 80 -v available in the FASTX Toolkit 0.0.14 (https://github.com/agordon/fastx_toolkit), which only retain sequences with PHRED score > 20 in more than 80% of nucleotides. The integrity of trimmed and quality filtered fastq files were further verified using FastQC 0.11.9 (https://github.com/s-andrews/FastQC).

Sequence Counting, Filtering and Annotation

Data was summarized over counts of unique sequences across all samples. Tools for such sequence-based approaches are available—such as Sports and MintMap [26, 37]—and may be contrasted against feature-based counting approaches, where sequences are counted over genomic features—such as the genomic coordinates of miRNA and piRNA. One benefit of unique sequence-based approaches is that the sequence of the original read is maintained during counting, which is often lost du initial mismatch allowance in feature-based approaches. Here we counted unique sequences in trimmed fastq files using customized scripts in R 3.4.4 [38]. For parallel and efficient large data processing, foreach 1.5.0 [39] and the data.table [40] packages were used. Reading and processing fastq-files were done using the ShortRead package [41]. Prior to normalization, noise was reduced by only including sequences with at least 5 counts in 100% of samples within a given method (5P-seq, 5X-seq, 5XP-seq etc). This dataset contained 330 971 unique small RNA sequences

with mean total counts per sample of 8 993 458. Normalized counts in reads per million (rpm) were generated by dividing the individual sequence counts (reads) with the total sequence counts for each sample. For some analysis that targeted highly abundant sRNAs we further reduced the data by only including sequences with at least 10 rpm in 100% of the samples within a given method included in the analysis. Some critical values of the datasets are available in Table 1.

Sequence annotation was performed by mapping unique sequences against small RNA sequence reference databases using bowtie 1.2.2 [42]. This was done in cycles, where each cycle allowed one additional mismatch, from 0 to 3 mismatches. A sequence was only destined for another bowtie annotation cycle if it failed to align to any of the databases in the previous cycle. Fasta reference sequence files were attained from the following databases: miRNA = miRbase and Ensembl, tRNA = GtRNAdb (nuclear) and Ensembl (mitochondrial), rRNA = Ensembl (both nuclear and mitochondrial), piRNA = pirBase, Other sncRNA = Ensembl. We used dm6 versions across all databases. To resolve multimapping issues when sequences matched multiple databases within the same mismatch category, we applied the following hierarchy miRNA > Mt_tRNA > tRNA > MT_rRNA > rRNA > piRNA > other sncRNA. A sequence mapped to miRNA with 1 mismatch and piRNA with 0 mismatch, would therefore annotate as piRNA, while if both had 0 mismatch it would annotate as miRNA.

Statistical Methods and Visualization

All statistical analysis was done in R 3.4.4 [38]. Data visualization was primarily done using the ggplot2 package in R [43] and finalized using Inkscape 0.92. For fitting the saturation plots we used non-linear least square regression with an asymptotic self-starter (r functions: nls and SSasymp). For hierarchical clustering and correlation plots we used the corrplot package [44]. For tRNA/rRNA coverage we used custom scripts wrapped around the vmatchPattern function in the Biostrings package [45]. A version of this script has been published [7] and are available here: https://github.com/Danis102/Natt_et_al_2019_Human_Sperm_Rapid_Response_to_Diet/blob/master/S1_Text.R. Reference sequences were obtained from GtRNAdb (tRNA) and Ensembl (rRNA). GtRNAdb ss-files were used to map tRNA loops.

Future scripts for 5´XP sRNA-seq will be posted here: https://github.com/Danis102/.

References

1. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet. 2009;10(12):833-44. doi: 10.1038/nrg2683. PubMed PMID: 19920851.
2. Wery M, Kwapisz M, Morillon A. Noncoding RNAs in gene regulation. Wiley Interdiscip Rev Syst Biol Med. 2011;3(6):728-38. Epub 2011/03/04. doi: 10.1002/wsbm.148. PubMed PMID: 21381218.
3. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10(2):94-108. doi: 10.1038/nrg2504. PubMed PMID: 19148191; PubMed Central PMCID: PMCPMC2724769.
4. Ernst C, Odom DT, Kutter C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nature Communications. 2017;8(1):1411. doi: 10.1038/s41467-017-01049-7.
5. Lambert M, Benmoussa A, Provost P. Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA. Non-Coding RNA. 2019;5(1):16. PubMed PMID: doi:10.3390/ncrna5010016.
6. Tuck AC, Tollervey D. RNA in pieces. Trends Genet. 2011;27(10):422-32. Epub 2011/07/06. doi: 10.1016/j.tig.2011.06.001. PubMed PMID: 21741109.
7. Nätt D, Kugelberg U, Casas E, Nedstrand E, Zalavary S, Henriksson P, et al. Human sperm displays rapid responses to diet. PLOS Biology. 2019;17(12):e3000559. doi: 10.1371/journal.pbio.3000559.
8. Shi J, Zhang Y, Zhou T, Chen Q. tsRNAs: The Swiss Army Knife for Translational Regulation. Trends Biochem Sci. 2019;44(3):185-9. Epub 2018/10/10. doi: 10.1016/j.tibs.2018.09.007. PubMed PMID: 30297206; PubMed Central PMCID: PMCPMC6379142.
9. Wu R, Jiang D, Wang Y, Wang X. N (6)-Methyladenosine (m(6)A) Methylation in mRNA with A Dynamic and Reversible Epigenetic Modification. Mol Biotechnol. 2016;58(7):450-9. doi: 10.1007/s12033-016-9947-9. PubMed PMID: 27179969.
10. Jia G, Fu Y, He C. Reversible RNA adenosine methylation in biological regulation. Trends Genet. 2013;29(2):108-15. Epub 2012/12/04. doi: 10.1016/j.tig.2012.11.003. PubMed PMID: 23218460; PubMed Central PMCID: PMCPMC3558665.
11. El Yacoubi B, Bailly M, de Crécy-Lagard V. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. Annu Rev Genet. 2012;46:69-95. Epub 2012/08/16. doi: 10.1146/annurev-genet-110711-155641. PubMed PMID: 22905870.
12. Shigematsu M, Kawamura T, Kirino Y. Generation of 2',3'-Cyclic Phosphate-Containing RNAs as a Hidden Layer of the Transcriptome. Front Genet. 2018;9:562. Epub 2018/11/27. doi: 10.3389/fgene.2018.00562. PubMed PMID: 30538719; PubMed Central PMCID: PMCPMC6277466.
13. Abdelhamid RF, Plessy C, Yamauchi Y, Taoka M, de Hoon M, Gingeras TR, et al. Multiplicity of 5' cap structures present on short RNAs. PLoS One. 2014;9(7):e102895. Epub 2014/07/31. doi: 10.1371/journal.pone.0102895. PubMed PMID: 25079783; PubMed Central PMCID: PMCPMC4117478.
14. Chen YG, Kowtoniuk WE, Agarwal I, Shen Y, Liu DR. LC/MS analysis of cellular RNA reveals NAD-linked RNA. Nat Chem Biol. 2009;5(12):879-81. Epub 2009/10/11. doi:

589 10.1038/nchembio.235. PubMed PMID: 19820715; PubMed Central PMCID:
590 PMCPMC2842606.
591 15. Kowtoniuk WE, Shen Y, Heemstra JM, Agarwal I, Liu DR. A chemical screen for
592 biological small molecule-RNA conjugates reveals CoA-linked RNA. Proc Natl Acad Sci U S A.
593 2009;106(19):7768-73. Epub 2009/04/28. doi: 10.1073/pnas.0900528106. PubMed PMID:
594 19416889; PubMed Central PMCID: PMCPMC2674394.
595 16. Zhang Y, Pelechano V. High-throughput 5'P sequencing reveals environmental
596 regulated ribosome stalls at termination level. bioRxiv. 2020:2020.06.22.165134. doi:
597 10.1101/2020.06.22.165134.
598 17. Nersisyan L, Ropat M, Pelechano V. Improved computational analysis of ribosome
599 dynamics from 5'P degradome data using fivepeseq. bioRxiv. 2020:2020.01.22.915421. doi:
600 10.1101/2020.01.22.915421.
601 18. Attaf-Bouabdallah N, Cervera-Marzal I, Dong C, Gil L, Renand A, Spinelli L, et al.
602 FB5P-seq: FACS-based 5-prime end single-cell RNAseq for integrative analysis of
603 transcriptome and antigen receptor repertoire in B and T cells. bioRxiv. 2019:795575. doi:
604 10.1101/795575.
605 19. Raabe CA, Tang TH, Brosius J, Rozhdestvensky TS. Biases in small RNA deep
606 sequencing data. Nucleic Acids Res. 2014;42(3):1414-26. Epub 2013/11/05. doi:
607 10.1093/nar/gkt1021. PubMed PMID: 24198247; PubMed Central PMCID: PMCPMC3919602.
608 20. Pak J, Fire A. Distinct populations of primary and secondary effectors during RNAi
609 in C. elegans. Science. 2007;315(5809):241-4. Epub 2006/11/23. doi:
610 10.1126/science.1132839. PubMed PMID: 17124291.
611 21. Munafó DB, Robb GB. Optimization of enzymatic reaction conditions for generating
612 representative pools of cDNA from small RNA. Rna. 2010;16(12):2537-52. Epub 2010/10/06.
613 doi: 10.1261/rna.2242610. PubMed PMID: 20921270; PubMed Central PMCID:
614 PMCPMC2995414.
615 22. Xu H, Yao J, Wu DC, Lambowitz AM. Improved TGIRT-seq methods for
616 comprehensive transcriptome profiling with decreased adapter dimer formation and bias
617 correction. Sci Rep. 2019;9(1):7953. Epub 2019/05/28. doi: 10.1038/s41598-019-44457-z.
618 PubMed PMID: 31138886; PubMed Central PMCID: PMCPMC6538698.
619 23. Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, et al. High-
620 throughput sequencing of human plasma RNA by using thermostable group II intron reverse
621 transcriptases. RNA. 2016;22(1):111-28. Epub 2015/11/09. doi: 10.1261/rna.054809.115.
622 PubMed PMID: 26554030; PubMed Central PMCID: PMCPMC4691826.
623 24. Homolka D, Pandey Radha R, Goriaux C, Brasset E, Vaury C, Sachidanandam R, et
624 al. PIWI Slicing and RNA Elements in Precursors Instruct Directional Primary piRNA Biogenesis.
625 Cell Reports. 2015;12(3):418-28. doi: 10.1016/j.celrep.2015.06.030.
626 25. Locati MD, Pagano JFB, Abdullah F, Ensink WA, van Olst M, van Leeuwen S, et al.
627 Identifying small RNAs derived from maternal- and somatic-type rRNAs in zebrafish
628 development. Genome. 2018;61(5):371-8. Epub 2018/02/09. doi: 10.1139/gen-2017-0202.
629 PubMed PMID: 29425468.
630 26. Loher P, Telonis AG, Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear
631 and mitochondrial tRNA fragments from short RNA-seq data. Scientific reports. 2017;7:41184-
632 . doi: 10.1038/srep41184. PubMed PMID: 28220888.
633 27. Nätt D, Öst A. Male reproductive health and intergenerational metabolic responses
634 from a small RNA perspective. J Intern Med. 2020. Epub 2020/05/16. doi:
635 10.1111/joim.13096. PubMed PMID: 32415866.

28. Lyons SM, Fay MM, Akiyama Y, Anderson PJ, Ivanov P. RNA biology of angiogenin: Current state and perspectives. RNA Biology. 2017;14(2):171-8. doi: 10.1080/15476286.2016.1272746.
29. Zhou J, Liu S, Chen Y, Fu Y, Silver AJ, Hill MS, et al. Identification of two novel functional tRNA-derived fragments induced in response to respiratory syncytial virus infection. J Gen Virol. 2017;98(7):1600-10. Epub 2017/07/15. doi: 10.1099/jgv.0.000852. PubMed PMID: 28708049; PubMed Central PMCID: PMCPMC5721923.
30. Project AET, Project CSHLET. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature. 2009;457(7232):1028-32. Epub 2009/01/25. doi: 10.1038/nature07759. PubMed PMID: 19169241; PubMed Central PMCID: PMCPMC2719882.
31. Kiledjian M. Eukaryotic RNA 5'-End NAD. Trends Cell Biol. 2018;28(6):454-64. Epub 2018/03/12. doi: 10.1016/j.tcb.2018.02.005. PubMed PMID: 29544676; PubMed Central PMCID: PMCPMC5962413.
32. Bird JG, Zhang Y, Tian Y, Panova N, Barvík I, Greene L, et al. The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. Nature. 2016;535(7612):444-7. Epub 2016/07/06. doi: 10.1038/nature18622. PubMed PMID: 27383794; PubMed Central PMCID: PMCPMC4961592.
33. Grudzien-Nogalska E, Bird JG, Nickels BE, Kiledjian M. "NAD-capQ" detection and quantitation of NAD caps. RNA. 2018;24(10):1418-25. Epub 2018/07/25. doi: 10.1261/rna.067686.118. PubMed PMID: 30045887; PubMed Central PMCID: PMCPMC6140466.
34. Sharma S, Grudzien-Nogalska E, Hamilton K, Jiao X, Yang J, Tong L, et al. Mammalian Nudix proteins cleave nucleotide metabolite caps on RNAs. Nucleic Acids Res. 2020;48(12):6788-98. doi: 10.1093/nar/gkaa402. PubMed PMID: 32432673; PubMed Central PMCID: PMCPMC7337524.
35. Wickersheim ML, Blumenstiel JP. Terminator oligo blocking efficiently eliminates rRNA from Drosophila small RNA sequencing libraries. Biotechniques. 2013;55(5):269-72. Epub 2013/11/13. doi: 10.2144/000114102. PubMed PMID: 24215643; PubMed Central PMCID: PMCPMC4077022.
36. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17(1):3. Epub 2011-08-02. doi: 10.14806/ej.17.1.200.
37. Shi J, Ko E-A, Sanders KM, Chen Q, Zhou T. SPORTS1. 0: a tool for annotating and profiling non-coding RNAs optimized for rRNA-and tRNA-derived small RNAs. Genomics, proteomics & bioinformatics. 2018;16(2):144-51.
38. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
39. foreach: Provides Foreach Looping Construct [Internet]. 2020. Available from: https://CRAN.R-project.org/package=foreach.
40. data.table: Extension of `data.frame` [Internet]. 2020. Available from: https://CRAN.R-project.org/package=data.table.
41. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. 2009;25(19):2607-8. Epub 2009/08/06. doi: 10.1093/bioinformatics/btp450. PubMed PMID: 19654119; PubMed Central PMCID: PMCPMC2752612.
42. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25.

43. Wickham H. ggplot2: elegant graphics for data analysis: Springer; 2016.
44. Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix. 2017.
45. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms v2.48.0. R package. 2018. doi: citeulike-article-id:11644278.

Table 1. Library stats and sequence diversity

				General library stats:					*Sequences with >10 cpm in all samples*								
Batch	Sub-library	n libraries	Selected size	Trimmed library size		Mean counts/cpm* (max/min)	Library size past 10 cpm		Total[A]	miRNA[B]		piRNA[C]		tsRNA[D]		rsRNA[E]	
			(nt)	Mean M counts (max/min)			Mean M reads (max / min)	%	n	n	%	n	%	n	%	n	%
5P-seq																	
A	-	3	10-70	18.21	(20.57/14.95)	14.33 (16/12)	8.35 (9.45/6.69)	46	4193	164	3.91	2835	67.61	117	2.79	932	22.23
5X-seq																	
A	-	3	10-70	14.74	(16.81/13.17)	13.67 (16/12)	5.05 (5.79/4.44)	34	2395	49	2.05	1628	67.97	144	6.01	1303	54.41
B	-	3	10-70	20.12	(20.85/18.91)	19.33 (20/18)	9.97 (10.31/9.36)	50	2260	79	3.50	1502	66.46	294	13.01	1118	49.47
5XP-seq																	
A	5X-notag	3	10-70	10.89	(12.73/7.94)	9.33 (11/7)	3.74 (4.56/2.67)	34	2135	37	1.73	1460	68.38	132	6.18	1306	61.17
	5P-tag	(3)	10-70	1.87	(2.14/1.41)	2.00 (2/2)	1.19 (1.38/0.90)	64	2298	179	7.79	1905	82.90	134	5.83	837	36.42
B	5X-notag	3	10-70	16.41	(18.23/14.45)	14.33 (16/13)	6.84 (7.92/5.95)	42	1895	64	3.38	1306	68.92	233	12.30	1076	56.78
	5P-tag	(3)	10-70	1.27	(1.40/1.15)	1.33 (2/1)	0.66 (0.74/0.60)	52	4194	219	5.22	3742	89.22	259	6.18	888	21.17

* Counts per million reads (cpm) was calculated on the affective library size where only sequences found in at least two independent samples were included.
[A] n unique sequences surpassing a filter threshold of at least 10 cpm in all samples of the batch.
[B] n unique sequences surpassing the filter in [A] and were listed in mirBase.
[C] n unique sequences surpassing the filter in [A] and were listed in pirBase.
[D] n unique sequences surpassing the filter in [A] and were listed in GtRNAdb.
[E] n unique sequences surpassing the filter in [A] and mapped to rRNA in Ensembl ncRNA.

690

Figures:

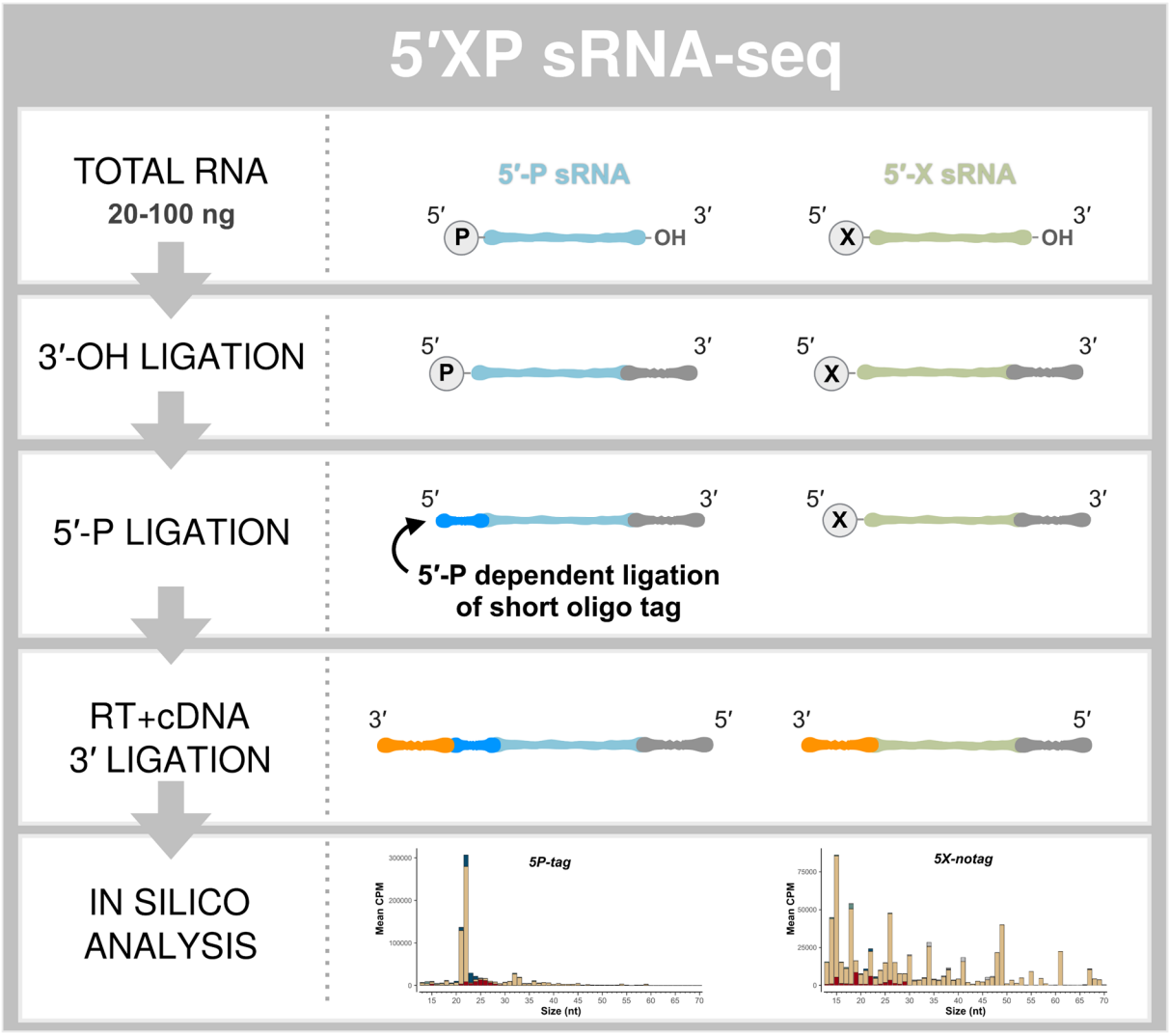

Fig 1. The principle steps of 5′XP sRNA-seq. Small RNA sequencing using 5′XP sRNA-seq generates two sub-libraries, either sensitive (5P-tag) or insensitive (5X-notag) to a phosphate (P) in the 5′-terminal of the original RNA. This is done by tagging RNA fragments with 5′-P using a sequence specific oligo that identifies the 5′-P in the downstream bioinformatic analysis.

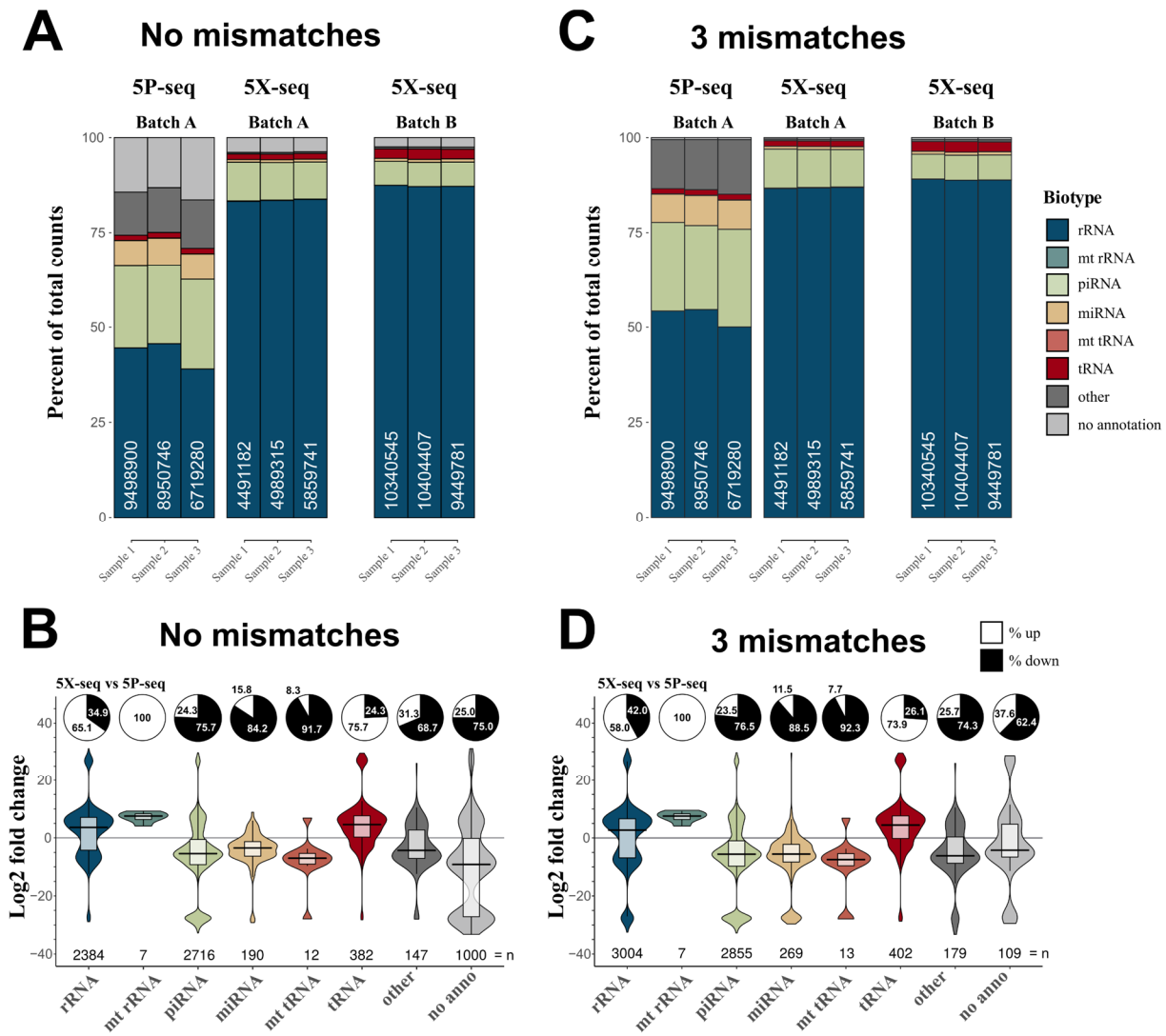

Fig 2. 5X-seq catches different small RNAs compared to 5P-seq. In order to generate a sequencing method more inclusive of RNAs with 5′ modifications we compared two adapter ligation strategies: 5P-seq adds the 5′ adapter on RNA and is dependent on 5′ phosphorylation, and 5X-seq adds the 5′ adapter, to the corresponding 3′ side of cDNA and is therefore independent of 5′ RNA phosphorylation. Stacked bars showing the sRNA composition of 5P and 5X libraries, allowing for **(A)** no mismatches and **(C)** 3 mismatches in the annotation against sRNA reference sequences. Batch A and B indicates samples isolated from the two independent pools of RNA. Violin plots with log2 fold differences between 5X and 5P per unique sequence when **(B)** 0 mismatches and **(D)** 3 mismatches are allowed. Pie charts show percent up and down regulated sequences. Only unique sequences that passed 10 CPM in all samples of a method (either 5P or 5X) was included. mt=mitochondrial. CPM = counts per million reads.

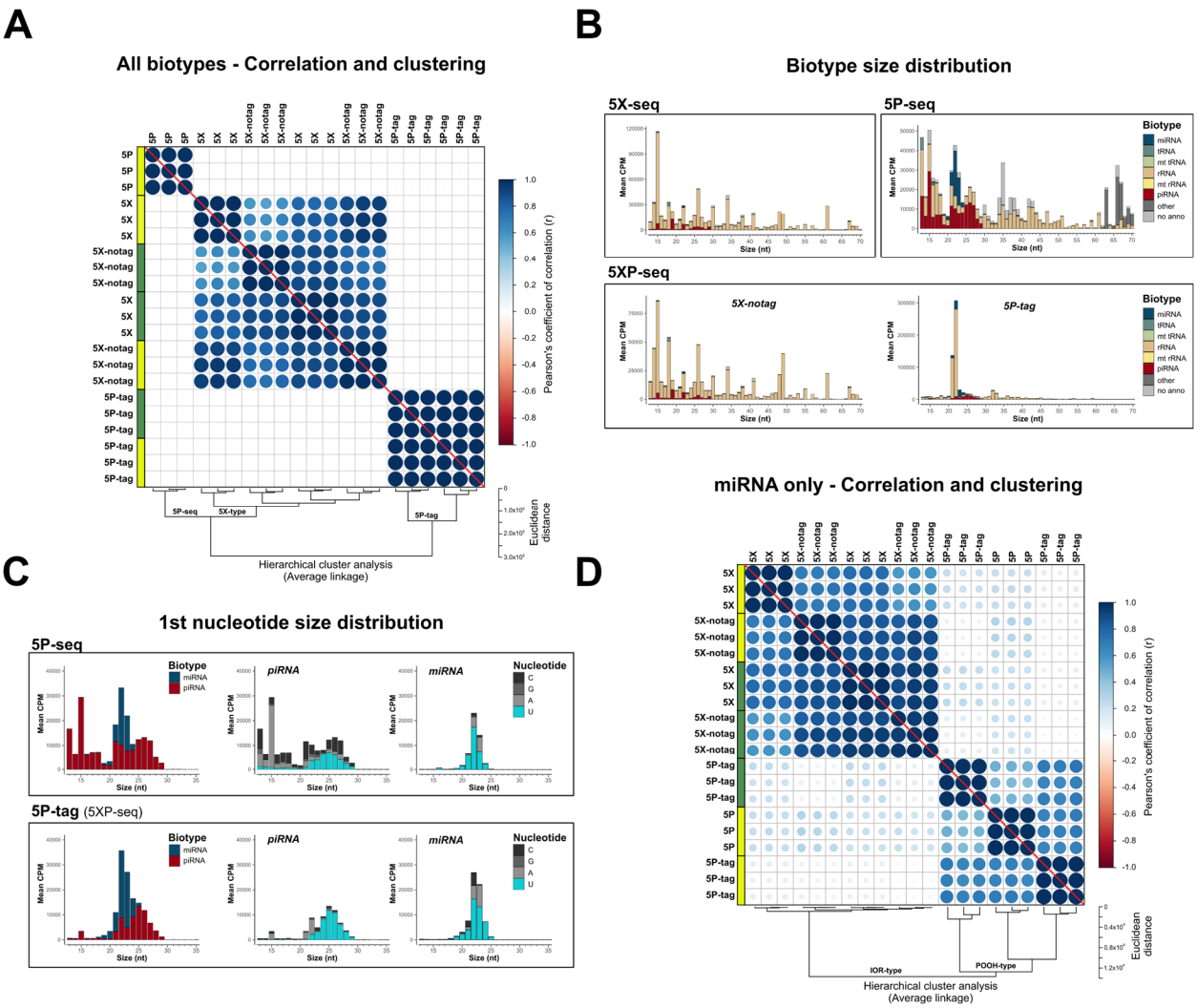

Fig 3. 5XP-seq replicates 5X-seq and identifies 5′ phosphorylated sRNA species. 5XP-seq distinguishes sRNA with (5P-tag) and without (5X-notag) 5′-terminal phosphorylation which characterize canonical miRNA and piRNA. **(A)** Correlation plot ordered by unsupervised hierarchical clustering showing three separate clusters, where all 5X-type libraries (5X-seq; 5X-notag sub-libraries of 5XP-seq) generates similar sequencing libraries, while 5P-seq and the 5P-tag sub-libraries of 5XP-seq forms two separate clades. **(B)** Histograms showing the mean expression of biotypes across different fragment sizes. **(C)** Size distributions of known miRNA and piRNA (left panels) show that the 5P-tag sub-libraries of 5XP-seq generates a cleaner canonical piRNA peak in the expected size range, with a typical Uridine (U) bias on the 1st nucleotide (middle panels), than 5P-seq. Size distribution and 1st nucleotide bias of miRNA are more similar between libraries (right panel). **(D)** Correlation plot ordered by hierarchical clustering of only annotated miRNA confirms that 5P-seq and 5P-tag sub-libraries enrich the same miRNA. Analysis was performed on sRNA that reached 10 CPM in all replicates of a method. Yellow and green boxes in the correlation plots shows libraries generated from two separate pools of RNA from *drosophila* embryos. CPM = counts per million reads.

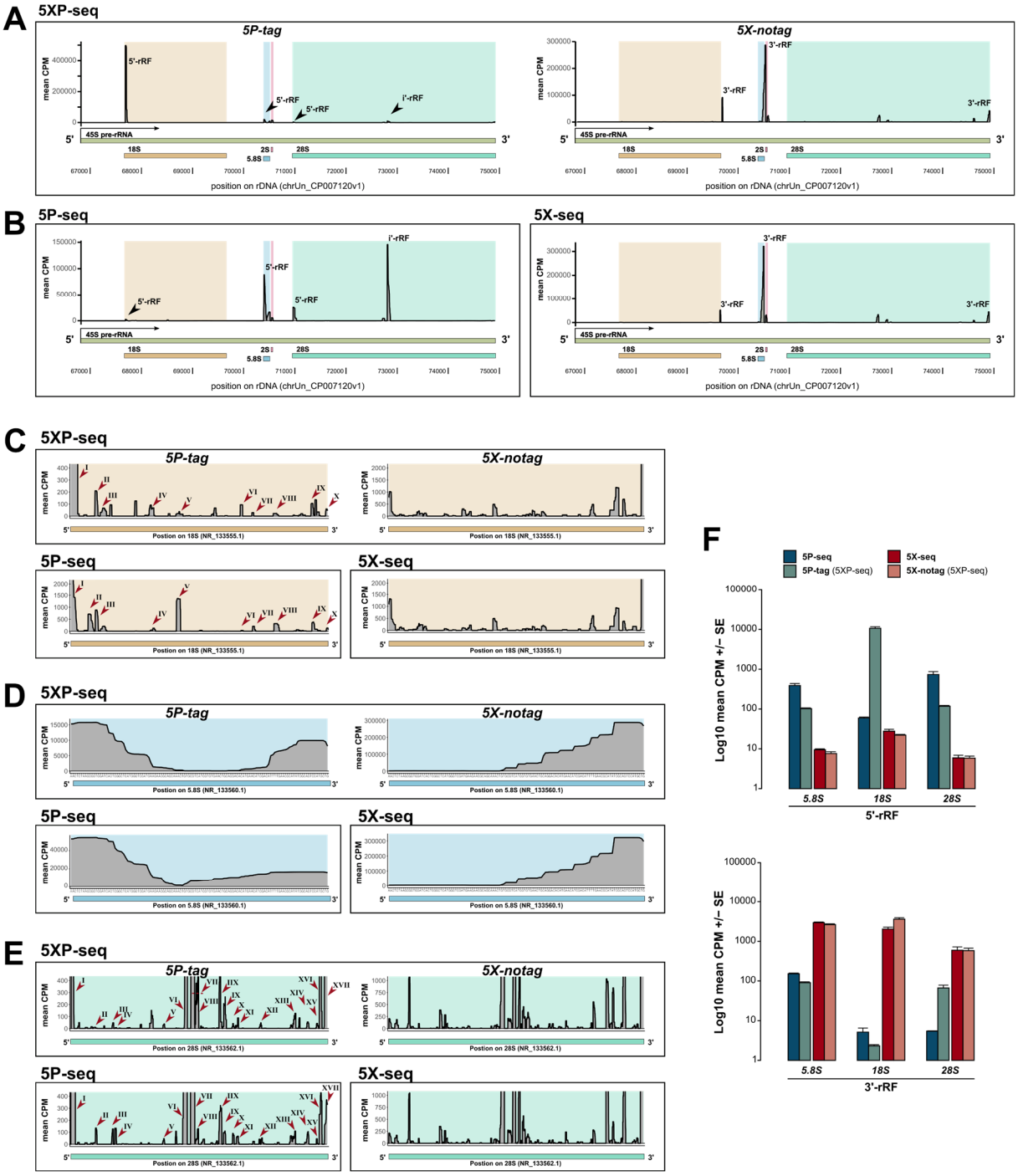

Fig 4. 5P-type libraries enrich 5´ rRNA fragments while 5X-type libraries enrich 3´ fragments. Graphs shows 5´ and 3´ stratification in sRNA derived from rRNA (rRF). **(A-B)** Coverage plots of rRFs across an example pre-rRNA sequence (45S; green) located on the repetitive rDNA loci (chrUn_CP007120v1), annotated with mature rRNA subunits 18S (light brown), 5.8 (light blue), 2S (light red) and 28S (light turquoise). **(A)** Shows mean CPM coverage of 5XP-seq libraries separated by 5P-tag (left panel) and 5X-notag (right panel) sub-libraries. **(B)** Instead shows 5P-seq and 5X-seq control libraries. **(C-E)** Zoomed-in coverage plots of the 18S, 5.8S and 28S subunits. For 5P-type libraries, roman numbers indicate major peaks present in both regular 5P-seq libraries and the 5P-tag sub-libraries of 5XP-seq. **(F)** Bars show 5´-rRF enrichment in 5P-type libraries (upper panel) and 3´-rRF enrichment in 5X-type libraries (lower panel). CPM = counts per million reads.

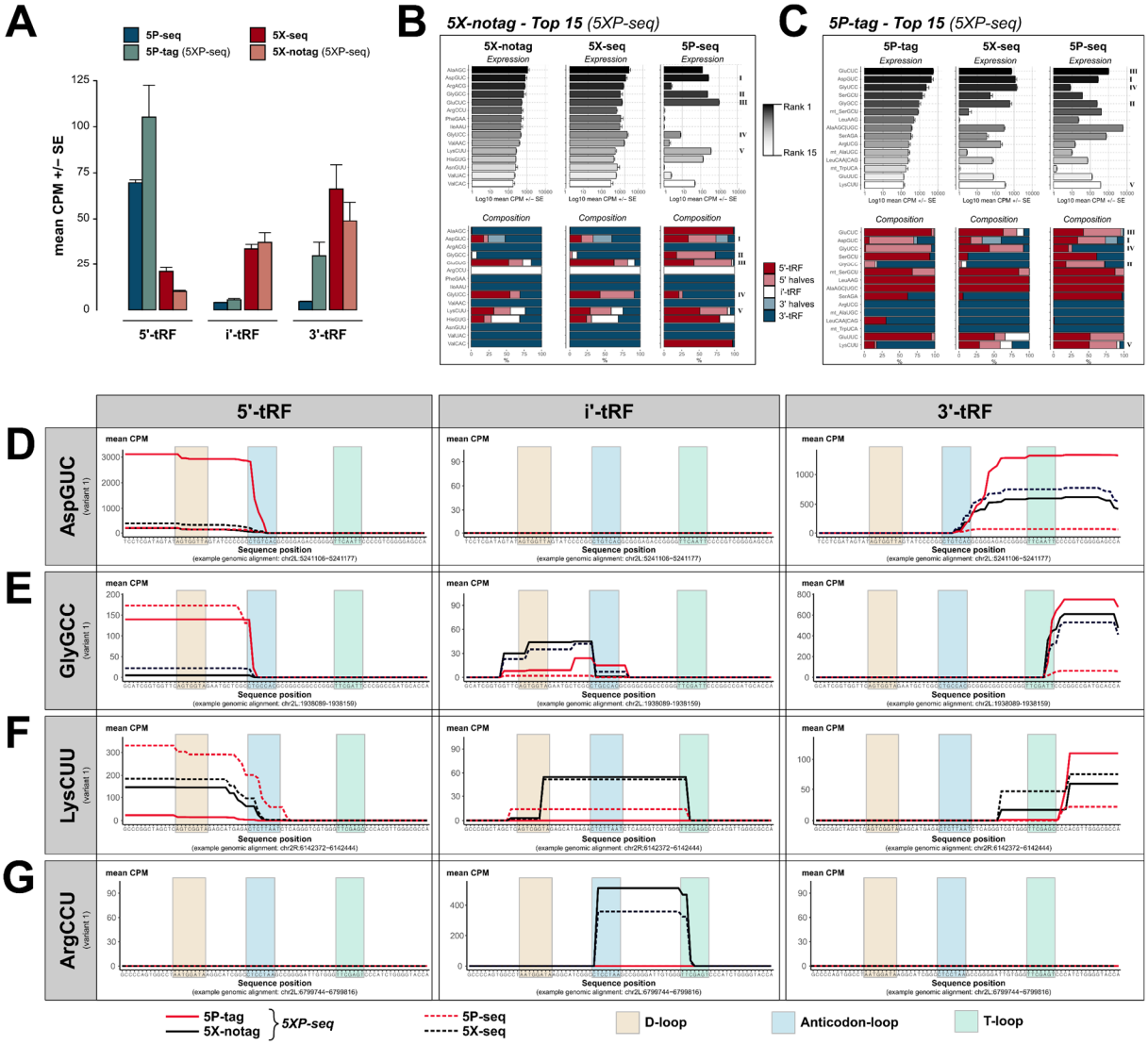

Fig 5. tRNA isodecoder analysis reveals the sRNA complexity generated by different library preparation strategies. Graphs show tRNA derived sRNA (tRFs) grouped by their classification into 5´, i' and 3´ sub-species, and tRNA isodecoder family. **(A)** Bars show 5´-tRF enrichment in 5P-type libraries and 3´-tRF enrichment in 5X-type libraries. **(B)** Top 15 expressed isodecoders in 5X-notag and their corresponding expression levels (black-grey-white bars) and 5´/i´/3´ ratios (blue/light blue stacked bars) in regular 5X-seq and 5P-seq libraries. **(C)** Same as (B), but the top expressed isodecoders in 5P-tag sub-libraries. **(D-G)** Show tRF coverage plots of a selection of isodecoders, separated by 5´, i' and 3´ tRF classification (left, middle and right panels, respectively). **(D)** AspGCU, illustrating anticodon-loop cleavage with both 5´ and 3´ halves. **(E)** GlyGCC, illustrating a 5´ half tRF cleaved in the anticodon-loop and a short 3´ tRF cleaved in the T-loop. **(F)** LysCUU, illustrating many 5´- and 3´-tRFs, but also an i'-tRF resulting from D- and T-loop cleavage. **(G)** ArgCCU showing a i'-tRF cleaved between the anticodon- and T-loops. CPM = counts per million reads.