SPORTS1.0: a tool for annotating and profiling non-coding RNAs optimized for rRNA- and tRNA- derived small RNAs

Junchao Shi[1,*], Eun-A Ko[1], Kenton M. Sanders[1], Qi Chen[1,*], Tong Zhou[1,*]

[1]*Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, NV 89512 USA.*

*Corresponding author: junchaoshi@nevada.unr.edu (J.S.) cqi@med.unr.edu (Q.C.), tongz@med.unr.edu (T.Z.)

Abstract:

High-throughput RNA-seq has revolutionized the process of small RNA (sRNA) discovery, leading to a rapid expansion of sRNA categories. In addition to previously well-characterized sRNAs such as miRNAs, piRNAs and snoRNAs, recent emerging studies have spotlighted on tsRNAs (tRNA-derived small RNAs) and rsRNAs (rRNA-derived small RNAs) as new categories of sRNAs that bear versatile functions. Since existing software and pipelines for sRNA annotation are mostly focusing on analyzing miRNAs or piRNAs, here we developed SPORTS1.0 (**s**mall RNA annotation **p**ipeline **o**ptimized for **r**RNA- and **t**RNA- derived **s**mall RNAs), which is optimized for analyzing tsRNAs and rsRNAs from sRNA-seq data, also with the capacity to annotate canonical sRNAs such as miRNAs and piRNAs. In addition, SPORTS1.0 can predict potential RNA modification sites basing on nucleotide mismatches within sRNAs. SPORTS1.0 is precompiled to annotate sRNAs for a wide range of 68 species across bacteria, yeast, plant and animal kingdoms additional species for analyses could be readily expanded upon end users' input. As an example, SPORTS1.0 revealed distinct tsRNA and rsRNA signatures from different mice tissues/cells; and discovered that tsRNAs bear the highest mismatch rate compared with other sRNA species, which is consistent with their highly modified nature. SPORTS1.0 is an open-source software deposited at https://github.com/junchaoshi/sports1.0**.**

Introduction

Expanding classes of small RNAs (sRNAs) have emerged as key regulators of gene expression, genome stability and epigenetic regulation [1, 2]. In addition to previously well-characterized sRNA classes such as miRNAs, snoRNAs and piRNAs, recent analysis of sRNA-seq data has led to the identification of expanding novel small RNA families, including tRNA-derived small RNAs (tsRNAs, also called tRNA-derived fragments (tRFs)) and rRNA-derived small RNAs (rsRNAs) [3]. tsRNAs and rsRNAs have been discovered in a wide range of species with evolutionary conservation, supposedly due, in part, to the highly conservative sequence of their precursors: tRNAs and rRNAs, respectively [3]. Interestingly, tsRNAs and rsRNAs have been abundantly found in unicellular organisms (e.g. protozoa), where canonical small RNA pathways such as miRNA, siRNA and piRNAs are entirely lacking [4-6]. The dynamic regulation of tsRNAs and rsRNAs in these unicellular organisms suggests that they are among the most ancient classes of small RNAs for intra- and inter-cellular communications [7]. Moreover, recent emerging evidences from mammalian species have highlighted diverse biological functions mediated by tsRNAs, including regulating ribosome biogenesis, translation initiation, retrotransposon control, cancer metastasis, stem cell differentiation, neurological diseases and epigenetic inheritance [3, 8-15]. The exact molecular mechanisms of how tsRNAs exert their function in these processes have not been fully understood, but they are involved in regulations at both post-transcriptional and translational levels [11, 15, 16]. Compared to tsRNAs, rsRNAs are more recently discovered and also show tissue specific distribution. Dynamic expression of rsRNAs was found to be associated with diseases such as metabolic disorders and inflammation [17-19]. The diverse biological functions of tsRNAs and rsRNAs and their strong disease association are now pushing the new frontier of small RNA research. While there are multiple existing generalized sRNA annotation software and pipelines [20-24], and some have been developed aiming to analyze tsRNAs [25-27], specialized tools for simultaneously analyzing both tsRNAs and rsRNAs in addition to other canonical sRNAs (miRNA, piRNA etc) is still lacking. Here, we provide SPORTS1.0 (small RNA annotation pipeline optimized for rRNA- and tRNA- derived small RNAs), which can annotate and profile canonical sRNAs such as miRNAs and piRNAs, and is also optimized to analyze tsRNAs and rsRNAs from sRNA-seq data. In addition, SPORTS1.0 can help predict potential RNA modification sites basing on nucleotide mismatches within sRNAs.

Methods

The source code of SPORTS1.0 is written in Perl and R. The whole package and installation instructions are available on Github (https://github.com/junchaoshi/sports1.0). SPORTS1.0 can apply to a wide-range of species. The annotation references of 68 species are already precompiled for download (listed in **Supplementary Table S1**).

The workflow of SPORTS1.0 consists of four main steps: pre-processing, mapping, annotation output and annotation summary (**Figure 1)**. SRA, FASTQ and FASTA are the acceptable formats for data input. By calling *Cutadapt* [28] and *Perl* scripts extracted from *miRDeep2* [29], SPORTS1.0 outputs clean reads by removing sequence adapters and discarding sequences with length beyond the defined range and those with bases other than ATUCG. The clean reads obtained in pre-processing step will be sequentially mapped against reference genome, miRbase [30], rRNA database (collected from NCBI), GtRNAdb [31], piRNA database [32, 33], Ensembl [34] and Rfam [35], upon users' setting. sRNA sequences will be first annotated by *Bowtie* [36]. Next, a *Perl* script precompiled in SPORTS1.0 will be used to identify the locations of tsRNAs regarding whether they are derived from 5' terminus, 3' terminus or 3'CCA end of tRNAs. Then an *R* script precompiled in SPORTS1.0 will be applied to obtain rsRNA expression and positional mapping information regarding their respective rRNA precursors (5.8s, 18s, 28s etc).

SPORTS1.0 can also analyze sequence mismatch information if mismatches are allowed during alignment process. This information can help predict potential modification sites that have caused nucleotides misincorporation during the reverse transcription (RT) process, as previously reported [37]. In the current version, a mismatch site will be considered epreviously designated[37]. Binomial distribution is used to address whether the observed mismatch enrichment is significantly higher than the base-call error. Here, we define p_{err} as the basecalling error rate, n_{ref} as the number of nucleotides perfectly fitting to the reference sites, n_{mut} as the number of mismatch nucleotides, and n_{tot} as the sum of n_{ref} and n_{mut}. The probability of observing not larger than k perfectly matching nucleotides out of n_{tot} can be calculated as:

$$P(k \leq n_{ref}) = \sum_{i=0}^{k} pbinom(i; n_{tot}, (1 - p_{err}))$$

SPORTS1.0 provides two ways to evaluate n_{mut} number. The first option is to simply calculate n_{mut} as the reads number of sequences containing the particular mismatch. Since some sequences may align to multiple reference loci, this method may result in an increased false-positive rate. Therefore, a second method is proposed, in which reads number of sequences from multiple matching loci are uniformly distributed and consequently generates an adjusted n_{mut}.

SPORTS1.0 summary output includes annotation details for each sequence and length distribution along with other statistics. (See sample output **Figure 2-3, Supplementary Table S2-3**). User guideline is provided online (https://github.com/junchaoshi/sports1.0).

Results

As an example, we used SPORTS1.0 to analyze previously deposited sRNA-seq datasets from mouse sperm (GSM2304822 [38]), bone marrow cells (GSM1604100 [39]) and colon (GSM1975854 [40]). Graphic output by SPORTS1.0 reveals distinct sRNA profiles in sperm **(Figure 2A)**, bone marrow cells, **(Figure 2B)** and colon **(Figure 2C)**. Particularly tsRNAs and rsRNAs are equally or more abundant than previously well-known miRNAs or piRNAs (length distribution data for each type of sRNA are exemplified in **Supplementary Table S2**).

Importantly, SPORTS1.0 found an appreciable portion of rsRNAs annotated in sperm (48.7%), bone marrow cell (11.1%) and colon (61.1%) that are previously deemed as “unmatch genome” (**Figure 2A-C upper pie-chart**). This is because these newly annotated rsRNAs are derived from rRNA genes (rDNA), which were not assembled and shown in current mouse genome (mm10) [41], and thus were discarded before analysis by previous small RNA analyzing pipelines. SPORTS1.0 can now annotate and analyze these rsRNAs, including providing the subtypes of rRNA precursors (5.8s, 18s, 28s etc) from which they are deriving from (**Figure 3A-C**), as well as the loci mapping information (**Figure 3D-F**). Very interestingly, our analyses revealed that the specific loci that generate rsRNAs are completely distinct among sperm, bone marrow cells, and colon (**Figure 3D-F**), suggesting distinct biogenesis and functions. Similarly, SPORTS1.0 also revealed tissue-specific landscape of tsRNAs regarding both their relative abundance and the tRNA loci where they are deriving from (5' terminus, 3' terminus, 3'CCA end, etc.) (**Figure 2A-C lower pie-chart, Figure 4 and Supplementary Fig S1-3**). Since tsRNAs from different loci bear distinct biological functions [3], the tissue-specific tsRNA composition may represent features that define the unique functions of respective tissue/cell types.

In addition, SPORTS1.0 also revealed distinct mismatch rates among different types of sRNAs (**Figure 5 and Supplementary Table S3**), with tsRNAs showing the highest. The detected mismatch sites represent the modified nucleosides that might have caused misincoporation of nucleotides during the RT process. The relatively higher mismatch rate detected in tsRNA sequences is consistent with their highly modified nature. The mismatch sites detected by SPORTS1.0 could provide a potential source for further analyses of RNA modifications within sRNAs.

Finally, SPORTS1.0 can analyze sRNAs of a wide range of species, depending on the availability of their genome and sRNAs references (**Figure 6 and Supplementary Table S1**). The species to be analyzed and their associated sRNA references are subject to update in future versions, or can be customized by the end users.

Conclusion

SPORTS1.0 is an easy-to-use and flexible pipeline for analyzing sRNA-seq data across a wide-range of species. Using mice as example, SPORTS1.0 provided a far more complicated sRNAs landscape than we previously saw, highlighting a tissue-specific dynamic regulation of tsRNAs and rsRNAs. SPORTS1.0 can also predict potential RNA modification sites basing on nucleotide mismatches within sRNAs, and shows a distinct pattern between different sRNA types. SPORTS1.0 may set the platform for many future new discoveries in biomedical and evolution research that relate to sRNAs.

The real voyage of discovery consists not in seeking new landscapes, but in looking with new eyes.

-Marcel Proust

Author contributions

JC, TZ and QC conceived the idea and wrote the manuscript. JC and TZ developed the SPORTS1.0 software and analyzed the RNA-seq data. JC, EK, KS, QC and TZ contributed to the interpretation of the results.

Competing interests

The authors have declared no competing interests

Acknowledgements

We want to thank Songjia Fan and Tin Nguyen for constructive suggestions for the manuscript. This work is supported by Start-up funds for Zhou and Chen lab from University of Nevada, Reno School of Medicine, and from NIH (R01DK091336 and P01DK041315 to K.M.S; R01HD092431 and P30GM110767-03 to Q.C).

Reference:

[1] Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell 2014;157:77-94.

[2] Chen Q, Yan W, Duan E. Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. Nat Rev Genet 2016;17:733-43.

[3] Kumar P, Kuscu C, Dutta A. Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). Trends Biochem Sci 2016;41:679-89.

[4] Lambertz U, Oviedo Ovando ME, Vasconcelos EJ, Unrau PJ, Myler PJ, Reiner NE. Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. BMC Genomics 2015;16:151.

[5] Garcia-Silva MR, das Neves RF, Cabrera-Cabrera F, Sanguinetti J, Medeiros LC, Robello C, et al. Extracellular vesicles shed by Trypanosoma cruzi are linked to small RNA pathways, life cycle regulation, and susceptibility to infection of mammalian cells. Parasitol Res 2014;113:285-304.

[6] Liao JY, Guo YH, Zheng LL, Li Y, Xu WL, Zhang YC, et al. Both endo-siRNAs and tRNA-derived small RNAs are involved in the differentiation of primitive eukaryote Giardia lamblia. Proc Natl Acad Sci U S A 2014;111:14159-64.

[7] Szempruch AJ, Dennison L, Kieft R, Harrington JM, Hajduk SL. Sending a message: extracellular vesicles of pathogenic protozoan parasites. Nat Rev Microbiol 2016;14:669-75.

[8] Chen Q, Yan M, Cao Z, Li X, Zhang Y, Shi J, et al. Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science 2016;351:397-400.

[9] Schorn AJ, Gutbrod MJ, LeBlanc C, Martienssen R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell 2017;170:61-71 e11.

[10] Anderson P, Ivanov P. tRNA fragments in human health and disease. FEBS Lett 2014;588:4297-304.

[11] Kim HK, Fuchs G, Wang S, Wei W, Zhang Y, Park H, et al. A transfer-RNA-derived small RNA regulates ribosome biogenesis. Nature 2017;552:57-62.

[12] Gebetsberger J, Wyss L, Mleczko AM, Reuther J, Polacek N. A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. RNA Biol 2017;14:1364-73.

[13] Schimmel P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol 2017.

[14] Martinez G, Choudury SG, Slotkin RK. tRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res 2017;45:5142-52.

[15] Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P. Angiogenin-induced tRNA fragments inhibit translation initiation. Mol Cell 2011;43:613-23.

[16] Luo S, He F, Luo J, Dou S, Wang Y, Guo A, et al. Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. Nucleic Acids Res 2018.

[17] Wei H, Zhou B, Zhang F, Tu Y, Hu Y, Zhang B, et al. Profiling and identification of small rDNA-derived RNAs and their potential biological functions. PLoS One 2013;8:e56842.

[18] Chu C, Yu L, Wu B, Ma L, Gou LT, He M, et al. A sequence of 28S rRNA-derived small RNAs is enriched in mature sperm and various somatic tissues and possibly associates with inflammation. J Mol Cell Biol 2017;9:256-9.

[19] Zhang Y, Zhang X, Shi J, Tuorto F, Li X, Liu Y, et al. Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol 2018.

[20] Wu X, Kim TK, Baxter D, Scherler K, Gordon A, Fong O, et al. sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. Nucleic Acids Res 2017;45:12140-51.

[21] Mohorianu I, Stocks MB, Applegate CS, Folkes L, Moulton V. The UEA Small RNA Workbench: A Suite of Computational Tools for Small RNA Analysis. Methods Mol Biol 2017;1580:193-224.

[22] Rueda A, Barturen G, Lebron R, Gomez-Martin C, Alganza A, Oliver JL, et al. sRNAtoolbox: an integrated collection of small RNA research tools. Nucleic Acids Res 2015;43:W467-73.

[23] Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. RNA 2013;19:740-51.

[24] Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2011;39:W112-7.

[25] Thompson A, Zielezinski A, Plewka P, Szymanski M, Nuc P, Szweykowska-Kulinska Z, et al. tRex: A Web Portal for Exploration of tRNA-Derived Fragments in Arabidopsis thaliana. Plant Cell Physiol 2018;59:e1.

[26] Zheng LL, Xu WL, Liu S, Sun WJ, Li JH, Wu J, 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.

[27] Selitsky SR, Sethupathy P. tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinformatics 2015;16:354.

[28] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011;17:3.

[29] Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008;26:407-15.

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

[31] Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 2016;44:D184-9.

[32] Zhang P, Si X, Skogerbo G, Wang J, Cui D, Li Y, et al. piRBase: a web resource assisting piRNA functional study. Database (Oxford) 2014;2014:bau110.

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

[34] Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res 2016;44:D710-6.

[35] Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 2015;43:D130-7.

[36] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.

[37] Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, Dragomir I, et al. HAMR: high-throughput annotation of modified ribonucleotides. RNA 2013;19:1684-92.

[38] Yang Q, Lin J, Liu M, Li R, Tian B, Zhang X, et al. Highly sensitive sequencing reveals dynamic modifications and activities of small RNAs in mouse oocytes and early embryos. Sci Adv 2016;2:e1501482.

[39] Tuorto F, Herbst F, Alerasool N, Bender S, Popp O, Federico G, et al. The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. EMBO J 2015;34:2350-62.

[40] Peck BC, Mah AT, Pitman WA, Ding S, Lund PK, Sethupathy P. Functional Transcriptomics in Diverse Intestinal Epithelial Cell Types Reveals Robust MicroRNA Sensitivity in Intestinal Stem Cells to Microbial Status. J Biol Chem 2017;292:2586-600.

[41] McStay B, Grummt I. The epigenetics of rRNA genes: from molecular to chromosome biology. Annu Rev Cell Dev Biol 2008;24:131-57.

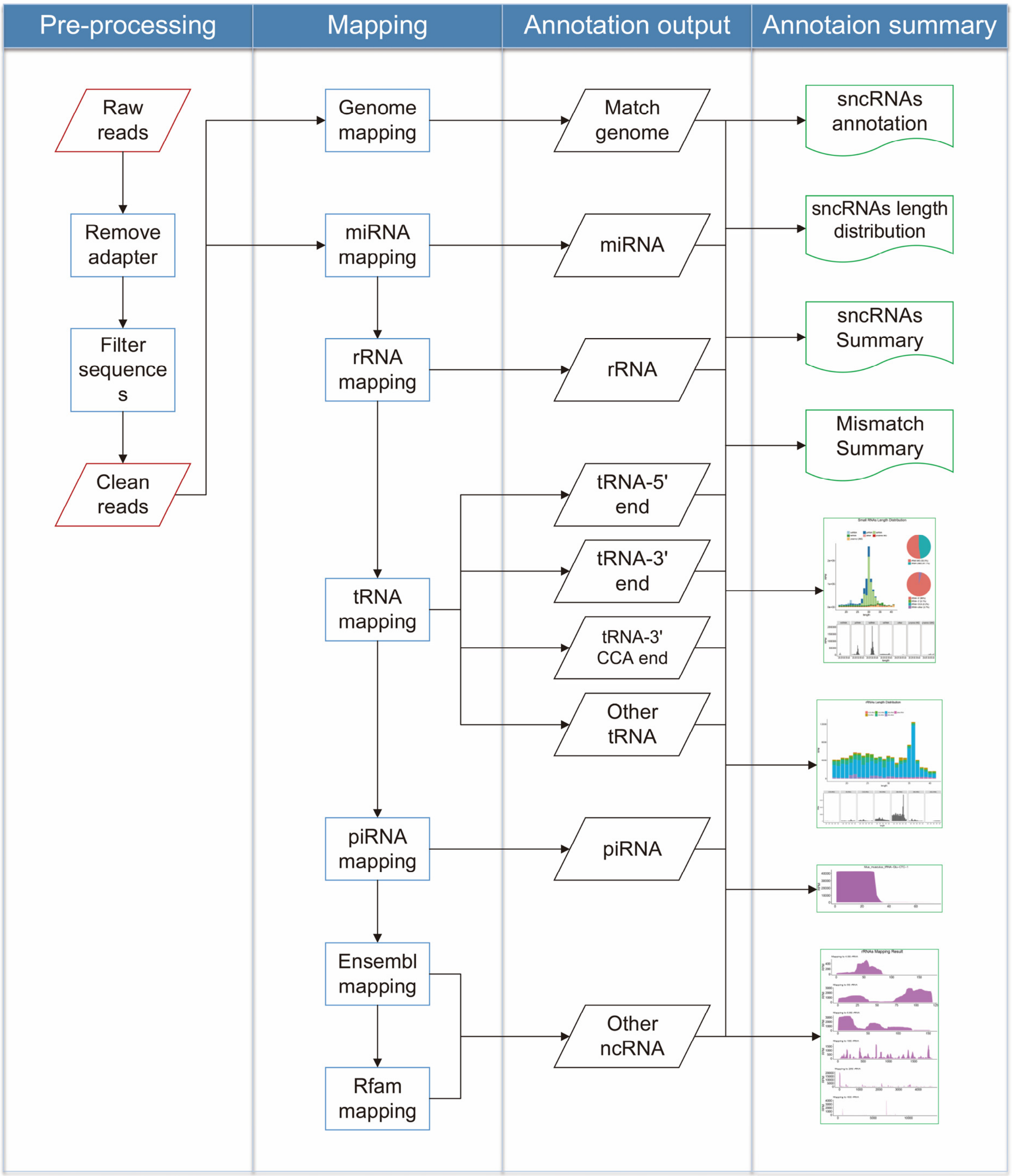

Figure 1. Workflow of SPORTS1.0 contains four main steps: pre-processing, mapping, annotation output and annotation summary, as outlined in the figure.

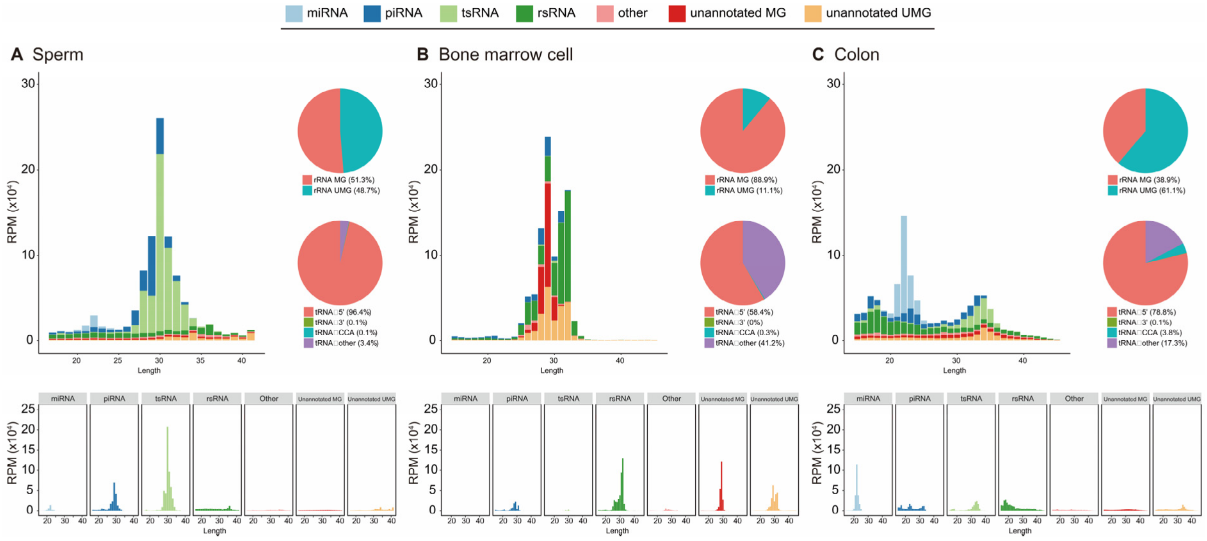

Figure 2. Exemplary annotation and profiling of sRNA-seq datasets generated by SPORTS1.0. **(A-C)** Categorization and length distribution analysis of different sRNA types in mouse sperm **(A)**, bone marrow cell **(B)** and Colon **(C).** RPM: reads per million clean reads.

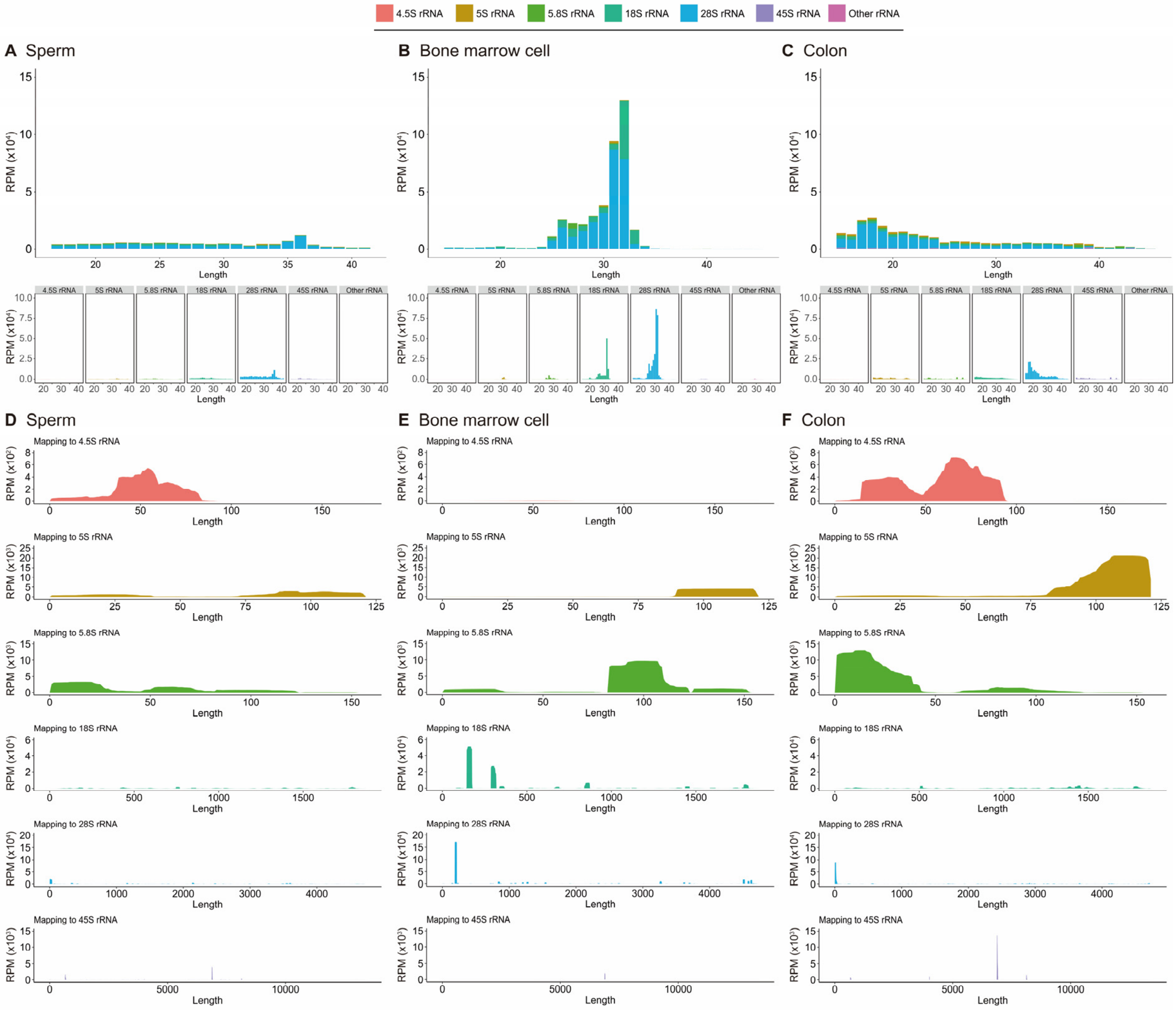

Figure 3. Tissue-specific rsRNA profiles revealed by SPORTS1.0. **(A-C)** subtypes of rRNA precursors (5.8s, 18s, 28s etc) for rsRNAs from **(A)** sperm, **(B)** bone marrow cell and **(C)** colon. **(D-F)** comparison of rsRNA-generating loci from different rRNA precursors reveals distinct pattern between **(D)** sperm, **(E)** bone marrow cell and **(F)** colon. RPM: reads per million clean reads.

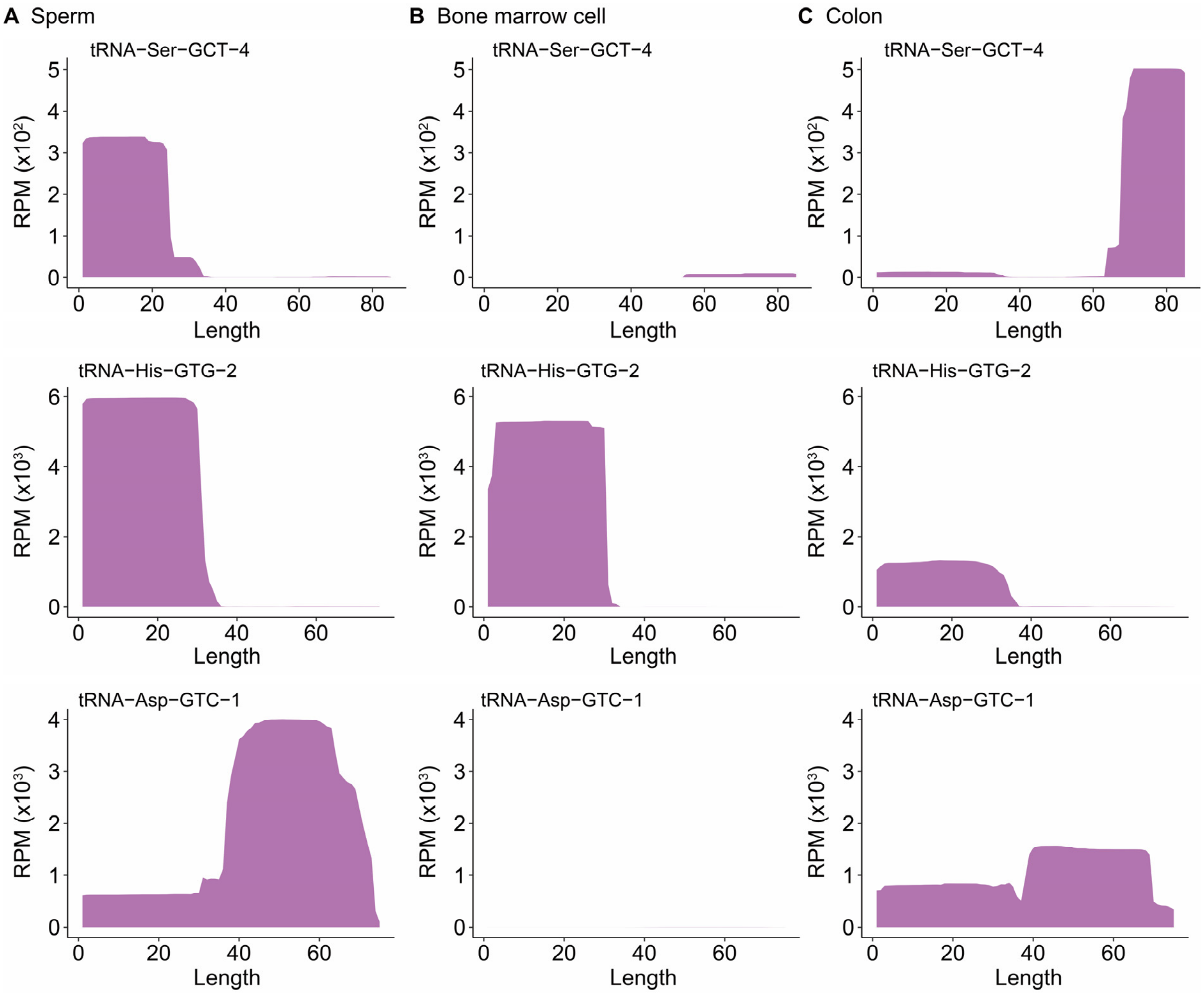

Figure 4. Examples of 3 tissue-specific tsRNA profiles revealed by SPORTS1.0 in **(A)** sperm, **(B)** bone marrow cell and **(C)** colon. Full tsRNA mapping results against tRNA loci are included in **Supplementary Fig S1-3** for (Fig S1) sperm, (Fig S2) bone marrow cell and (Fig S3) colon respectively. RPM: reads per million clean reads.

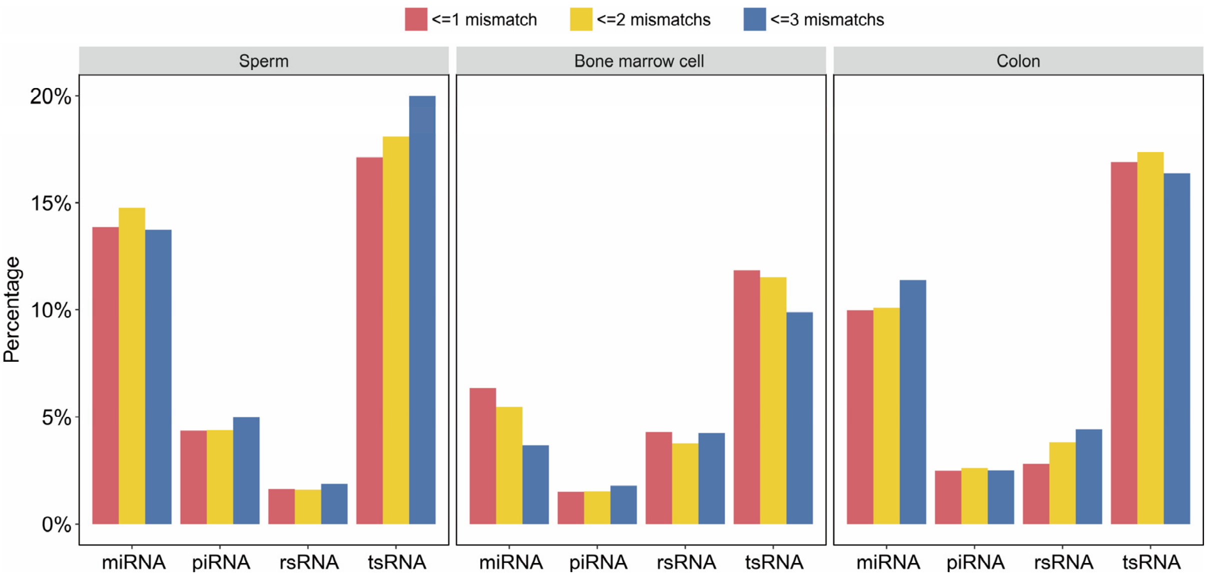

Figure 5. The percentage of unique sequences that contain significantly enriched mismatch out of total unique sequences from each subtypes of sRNAs (miRNAs, piRNAs, tsRNAs and rsRNAs) is provided in different tissues.

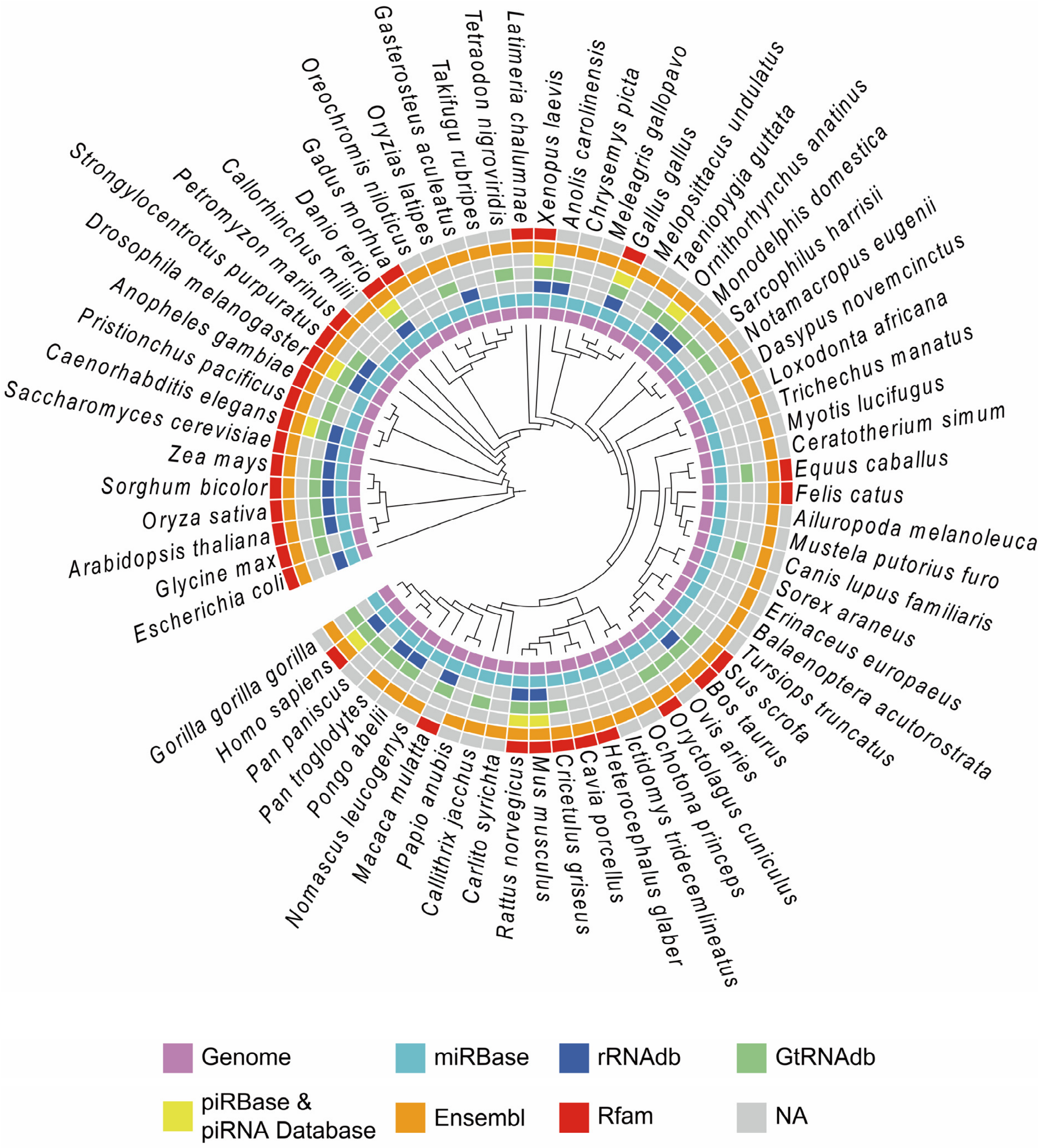

Figure 6. The 68 species and their respective reference database included in SPORTS1.0 precompiled for analysis.

Supplementary materials

Supplementary Figure S1. The mouse sperm tsRNA mapping results against tRNA loci revealed by SPORTS1.0. Mapping results for each annotated tsRNA were provided.

Supplementary Figure S2. The mouse bone marrow cell tsRNA mapping results against tRNA loci revealed by SPORTS1.0. Mapping results for each annotated tsRNA were provided.

Supplementary Figure S3. The mouse colon tsRNA mapping results against tRNA loci revealed by SPORTS1.0. Mapping results for each annotated tsRNA were provided.

Supplementary Table S1. The list of 68 species and their respective reference database that are precompiled in SPORTS1.0 ready for analyses.

Supplementary Table S2. Example output of SPORTS1.0 which includes annotation for each sequence (Table S2A), length distribution information (Table S2B) and expression level of each annotated category (Table S2C) for dataset GSM2304822.

Supplementary Table S3. Example output of SPORTS1.0 for sRNA sequence mismatch analysis for dataset GSM2304822 under the alignment criteria of mismatch ≤1 (Table S3A), ≤2 (Table S3B) and ≤3 (Table S3C).