SPORTS1.0: A Tool for Annotating and Profiling Non-coding RNAs Optimized for rRNA- and tRNA-derived Small RNAs

Junchao Shi, Eun-A Ko, Kenton M. Sanders, Qi Chen, Tong Zhou

Department of Physiology and Cell Biology, Reno School of Medicine, University of Nevada, Reno, NV 89512, USA

*Corresponding authors.

E-mail: junchaoshi@nevada.unr.edu (Shi J), cqi@med.unr.edu (Chen Q), tongz@med.unr.edu (Zhou T).

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 the previously well-characterized sRNAs such as microRNAs (miRNAs), Piwi-interacting RNA (piRNAs), and small nucleolar RNA (snoRNAs), recent emerging studies have spotlighted on tRNA-derived sRNAs (tsRNAs) and rRNA-derived sRNAs (rsRNAs) as new categories of sRNAs that bear versatile functions. Since existing software and pipelines for sRNA annotation are mostly focused on analyzing miRNAs or piRNAs, here we developed the sRNA annotation pipeline optimized for rRNA- and tRNA- derived sRNAs (SPORTS1.0). SPORTS1.0 is optimized for analyzing tsRNAs and rsRNAs from sRNA-seq data, in addition to its capacity to annotate canonical sRNAs such as miRNAs and piRNAs. Moreover, SPORTS1.0 can predict potential RNA modification sites based 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, while additional species for analyses could be readily expanded upon end users' input. For demonstration, by analyzing sRNA datasets using SPORTS1.0, we reveal that distinct signatures are present in tsRNAs and rsRNAs from different mouse cell types. We also find that compared to other sRNA species, tsRNAs bear the highest mismatch rate which is consistent with their highly modified nature. SPORTS1.0 is an open-source software and can be publically accessed at https://github.com/junchaoshi/sports1.0.

Keywords: Small RNA; RNA-seq data analysis; tsRNA; rsRNA; Annotation pipeline

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 the previously well-characterized sRNA classes such as microRNAs (miRNAs), Piwi-interacting RNA (piRNAs), and small nucleolar RNA (snoRNAs), recent analysis of sRNA-seq data has led to the identification of expanding novel sRNA families. These include tRNA-derived sRNAs (tsRNAs; also known as tRNA-derived fragments, tRFs) and rRNA-derived sRNAs (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 respective precursors, *i.e.*, tRNAs and rRNAs [3]. Interestingly, tsRNAs and rsRNAs have been abundantly found in unicellular organisms (*e.g.*, protozoa), where canonical sRNA 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 sRNAs for intra- and inter-cellular communications [7].

Moreover, recent emerging evidence from mammalian species have highlighted the 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]. Although tsRNAs are known to be involved in regulating these processes at both post-transcriptional and translational levels [11,14,16], the exact molecular mechanisms of how tsRNAs exert their functions have not been fully understood . Compared to tsRNAs, rsRNAs are more recently discovered and also exhibit tissue-specific distribution. Dynamic expression of rsRNAs is associated with diseases such as metabolic disorders and inflammation [17–19]. The diverse biological functions of tsRNAs and rsRNAs and their strong disease associations are now pushing the new frontier of sRNA research.

Currently, there are multiple existing general sRNA annotation software and pipelines [20–24], and some have been developed aiming to analyze tsRNAs [25–27]. However, there still lack the specialized tools that can simultaneously analyze both tsRNAs and rsRNAs in addition to other canonical sRNAs. Here, we provide SPORTS1.0, 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 based on nucleotide mismatches within sRNAs.

Method

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 and the annotation references of 68 species are precompiled for downloading (Table S1).

The workflow of SPORTS1.0 consists of four main steps, *i.e.*, 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 are 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 are first annotated by Bowtie [36]. Next, a *Perl* script precompiled in SPORTS1.0 is 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 is applied to obtain rsRNA expression level and positional mapping information regarding their respective rRNA precursors (5.8S, 18S, 28S, *etc*.).

SPORTS1.0 can also be used to analyze sequence mismatch information if mismatches are allowed during alignment process. Such information can help predict potential modification sites that have caused nucleotide misincorporation during the reverse transcription (RT) process as previously reported [37]. In the current version, a mismatch site is designated using criteria as previously described [37]. Binomial distribution is used to address whether the observed mismatch enrichment is significantly higher than the base-calling error. Here, we define p_{err} as the base-calling error rate, n_{ref} as the number of nucleotides perfectly fitted to the reference sites, n_{mut} as the number of mismatched nucleotides, and n_{tot} as the sum of n_{ref} and n_{mut}. The probability of observing not larger than k perfectly matched 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 methods to evaluate n_{mut} number. The first option is to simply calculate n_{mut} as the read number of sequences containing particular mismatches. Since some sequences may align to multiple reference loci, using this method may result in an increased false-positive rate. A second method is thus included, in which read number of sequences from multiple matching loci are uniformly distributed (based on the assumption that each of these multiple sites will equally express RNAs) 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** and **Figure 3,** Table S2 and Table S3). User guideline is provided online (https://github.com/junchaoshi/sports1.0).

Results

As an example, we used SPORTS1.0 to analyze sRNA-seq datasets from mouse sperm (GSM2304822 [38]), bone marrow cells (GSM1604100 [39]), and intestinal epithelial cells (GSM1975854 [40]) samples. Graphic output by SPORTS1.0 reveals distinct sRNA profiles in sperm (**Figure 2**A), bone marrow cells (Figure 2B), and intestinal epithelial cells (Figure 2C) samples. tsRNAs and rsRNAs are found equally or more abundantly than previously well-known miRNAs or piRNAs (length distribution data for each type of sRNA are exemplified in Table S2). In particularly, tsRNAs are dominant in sperm, rsRNAs are highest in bone marrow cells, and intestinal epithelial cells contains an appreciable amount of both tsRNAs and rsRNAs in addition to a miRNA peak.

Importantly, SPORTS1.0 found an appreciable portion of rsRNAs annotated in sperm (48.7%), bone marrow cell (11.1%) and intestinal epithelial cell (61.1%) samples that are previously deemed as "unmatch genome" (UMG) (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 sRNA 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 derived from (**Figure 3**A-C), as well as the loci mapping information (Figure 3D–F). Interestingly, our analyses revealed that the specific loci that generate rsRNAs are completely distinct among sperm, bone marrow cell, and intestinal epithelial cell samples (Figure 3D-F), suggesting distinct biogenesis and functions of these rsRNAs. Similarly, SPORTS1.0 also revealed tissue-specific landscape of tsRNAs in terms of their relative abundance (Figure 2A-C lower pie chart) and the tRNA loci where they are derived from (5'

terminus, 3' terminus, 3'CCA end, *etc*.) (**Figure 4**, and Figure 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 Table S3), with tsRNAs showing the highest. The detected mismatch sites represent the modified nucleotides that might have caused misincorporation 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 reference genome and sRNA sequences (**Figure 6** and Table S**1**). 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 provides a far more complicated sRNA landscape than having been previously seen, highlighting a tissue-specific dynamic regulation of tsRNAs and rsRNAs. SPORTS1.0 can also predict potential RNA modification sites based on nucleotide mismatches within sRNAs, and show a distinct pattern between different sRNA types. SPORTS1.0 may set the platform for many future new discoveries in biomedical and evolution research that is related to sRNAs.

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

-Marcel Proust

Authors' contributions

JS, TZ, and QC conceived the idea and wrote the manuscript. JS and TZ developed the SPORTS1.0 software and analyzed the RNA-seq data. JS, EK, KMS, QC, and TZ contributed to the interpretation of the results. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests

Acknowledgments

We want to thank Songjia Fan and Tin Nguyen for their constructive suggestions for the manuscript. This work is supported by Start-up funds for Zhou and Chen labs from University of Nevada, Reno School of Medicine, and from National Institutes of Health, USA (Grant Nos. R01DK091336 and P01DK041315 to KMS; Grant Nos. R01HD092431 and P30GM110767-03 to QC).

References

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

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

[15] Schimmel P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol 2018;19:45–58.

[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, doi: 10.1093/nar/gky189.

[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, doi: 10.1038/s41556-018-0087-2.

[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 J 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.

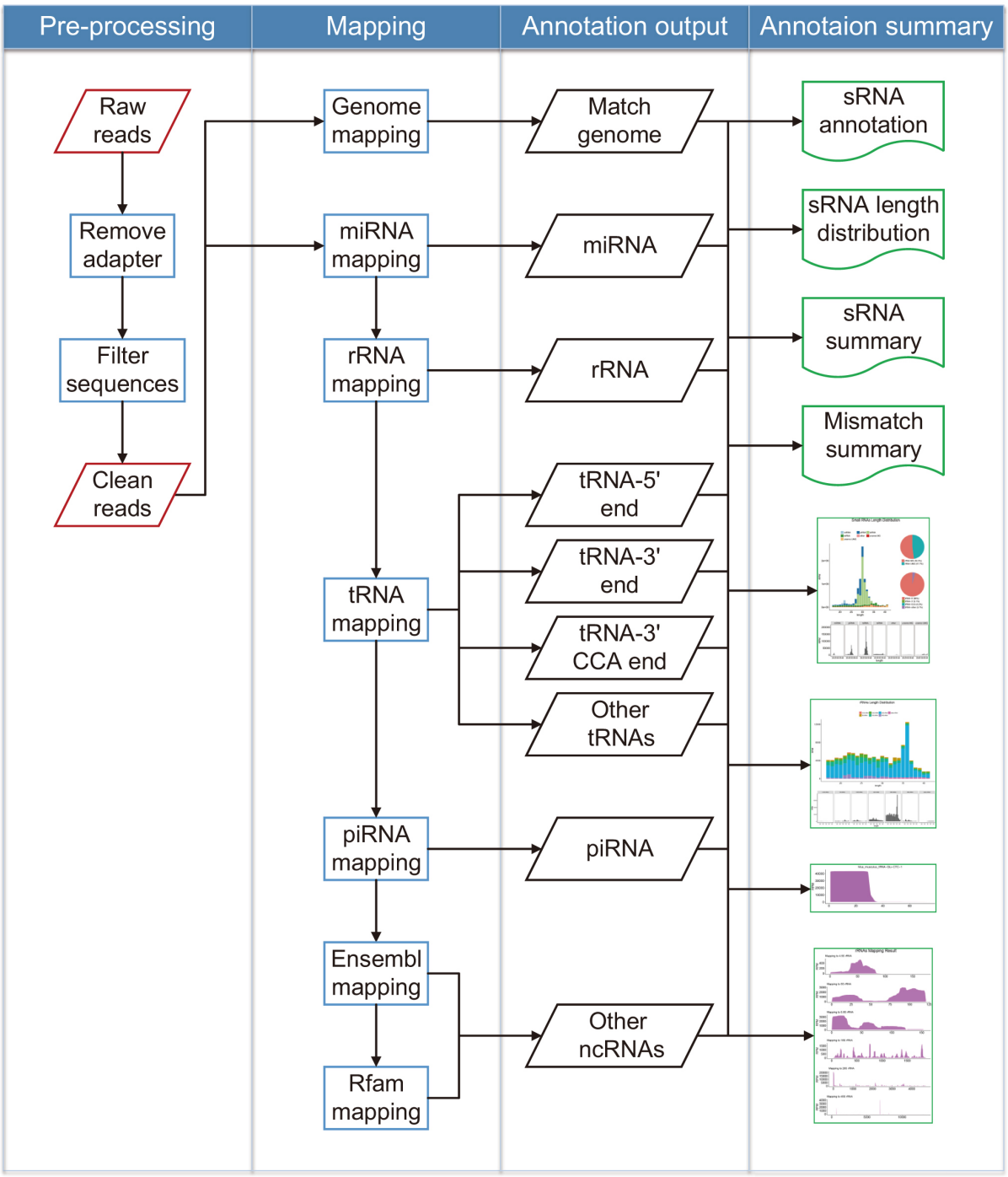

Figure 1 Workflow of SPORTS1.0

SPORTS1.0 contains four main steps, *i.e.*, pre-processing, mapping, annotation output, and annotation summary, as outlined in the figure.

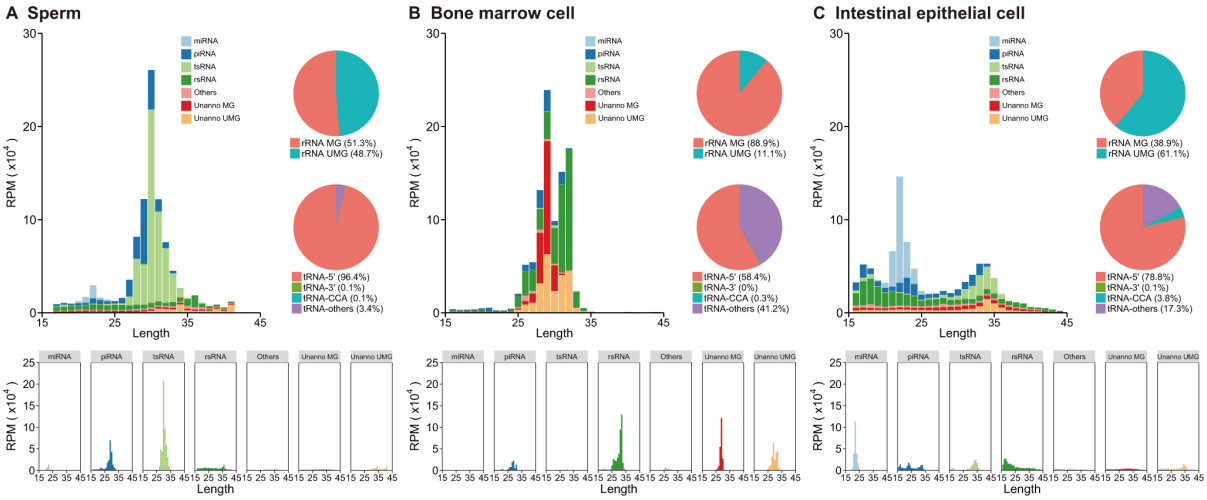

Figure 2 Exemplary annotation and profiling of sRNA-seq datasets generated by SPORTS1.0

Categorization and length distribution analysis of different sRNA types in mouse sperm (**A**), bone marrow cell (**B**), and intestinal epithelial cell (**C**) samples. RPM, reads per million clean reads; Unanno: unannotated; MG: match genome; UMG: unmatch genome.

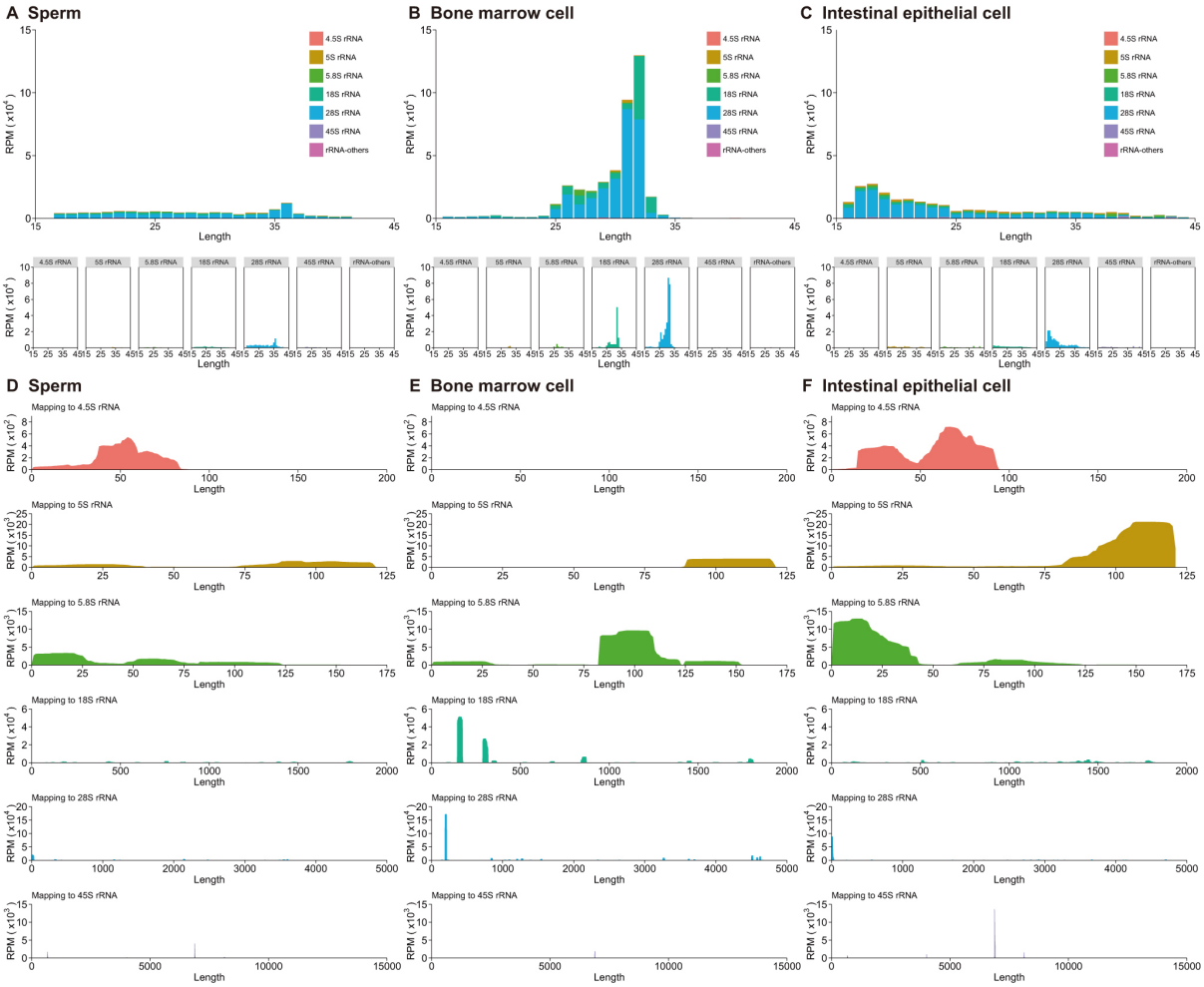

Figure 3 Cell-specific rsRNA profiles revealed by SPORTS1.0

Subtypes of rRNA precursors (5.8S, 18S, 28S, *etc*.) for rsRNAs from mouse sperm (**A**), bone marrow cell (**B**), and intestinal epithelial cell (**C**) samples. Comparison of rsRNA-generating loci from different rRNA precursors reveals distinct pattern between sperm (**D**), bone marrow cell (**E**), and intestinal epithelial cell (**F**). RPM, reads per million clean reads.

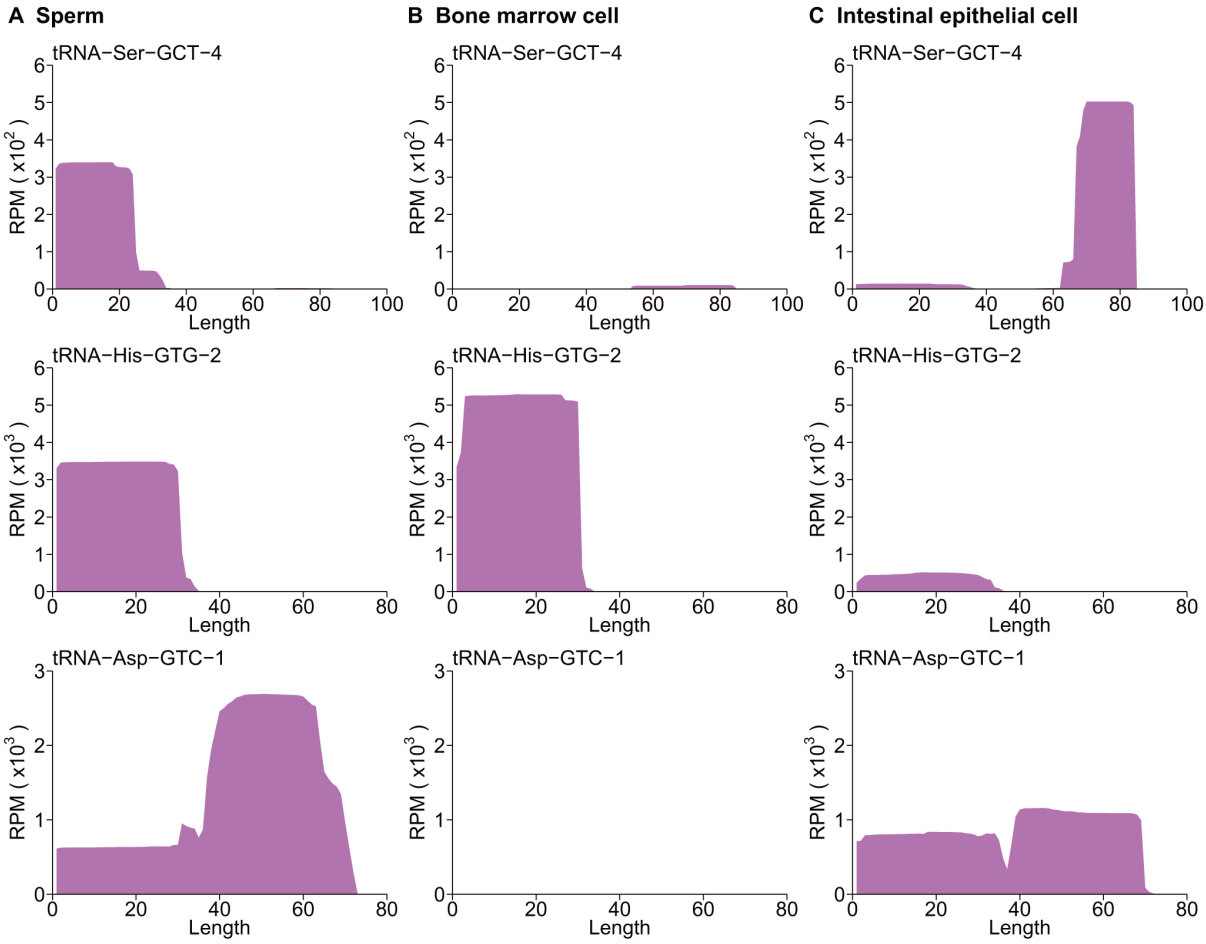

Figure 4 Cell-specific tsRNA profiles revealed by SPORTS1.0

Examples of 3 cell-specific tsRNA profiles revealed in mouse sperm (**A**), bone marrow cell (**B**), and intestinal epithelial cell (**C**) samples. Full tsRNA mapping results against tRNA loci are included in Figure S1–S3 for sperm (Figure S1), bone marrow cell (Figure S2), and intestinal epithelial cell (Figure S3) respectively. RPM, reads per million clean reads.

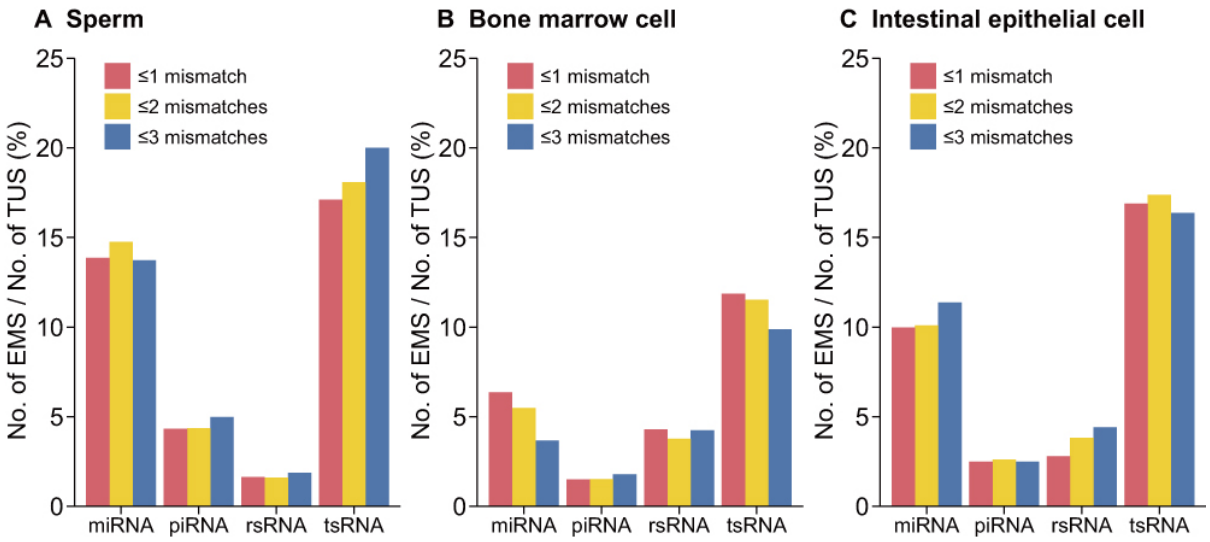

Figure 5 sRNA mismatch statistics by SPORTS1.0

The percentage of unique sequences that contain significantly-enriched mismatches out of total number of unique sequences from each subtype of sRNAs (miRNAs, piRNAs, tsRNAs, and rsRNAs) is provided for sperm (**A**), bone marrow cell (**B**), and intestinal epithelial cell (**C**) samples. EMS: enrichment mismatch sequences; TUS: total unique sequences.

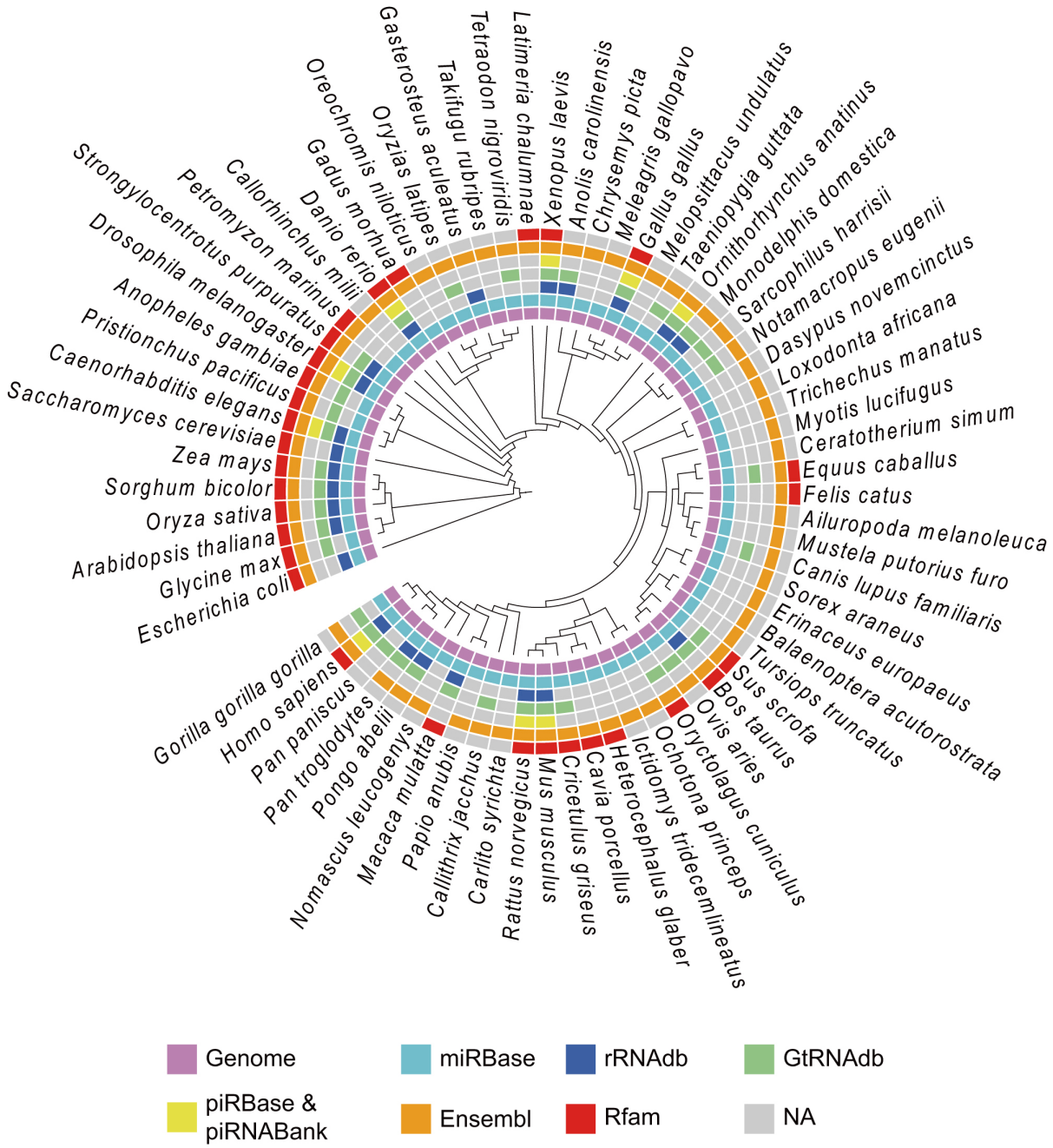

Figure 6 Species recompiled for analysis by SPORTS1.0

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

Supplementary materials

Figure S1 The mouse sperm tsRNA mapping results against tRNA loci revealed by SPORTS1.0

Mapping result for each annotated tsRNA was provided.

Figure S2 The mouse bone marrow cell tsRNA mapping results against tRNA loci revealed by SPORTS1.0

Mapping result for each annotated tsRNA was provided.

Figure S3 The mouse intestinal epithelial cell **tsRNA mapping results against tRNA loci revealed by SPORTS1.0**

Mapping result for each annotated tsRNA was provided.

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

Table S2 Example output of SPORTS1.0 which includes annotation for each sequence (A), length distribution information (B) and expression level of each annotated category (C) for dataset GSM2304822

Table S3 Example output of SPORTS1.0 for sRNA sequence mismatch analysis for dataset GSM2304822 under the alignment criteria of mismatch ≤ 1 (A), ≤ 2 (B), and ≤ 3 (C)