BMC Bioinformatics

SOFTWARE

Open Access

SmithHunter: a workflow for the identification of candidate smithRNAs and their targets

Giovanni Marturano[1†], Diego Carli[2†], Claudio Cucini[1], Antonio Carapelli[1,3], Federico Plazzi[2], Francesco Frati[1,3], Marco Passamonti[2*] and Francesco Nardi[1,3]

[†]Giovanni Marturano and Diego Carli: Equal contribution.

*Correspondence: marco.passamonti@unibo.it

[1] Department of Life Sciences, University of Siena, 53100 Siena, Italy
[2] Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126 Bologna, Italy
[3] National Biodiversity Future Center (NBFC), 90133 Palermo, Italy

Abstract

Background: SmithRNAs (Small MITochondrial Highly-transcribed RNAs) are a novel class of small RNA molecules that are encoded in the mitochondrial genome and regulate the expression of nuclear transcripts. Initial evidence for their existence came from the Manila clam *Ruditapes philippinarum*, where they have been described and whose activity has been biologically validated through RNA injection experiments. Current evidence on the existence of these RNAs in other species is based only on small RNA sequencing. As a preliminary step to characterize smithRNAs across different metazoan lineages, a dedicated, unified, analytical workflow is needed.

Results: We propose a novel workflow specifically designed for smithRNAs. Sequence data (from small RNA sequencing) uniquely mapping to the mitochondrial genome are clustered into putative smithRNAs and prefiltered based on their abundance, presence in replicate libraries and 5′ and 3′ transcription boundary conservation. The surviving sequences are subsequently compared to the untranslated regions of nuclear transcripts based on seed pairing, overall match and thermodynamic stability to identify possible targets. Ample collateral information and graphics are produced to help characterize these molecules in the species of choice and guide the operator through the analysis. The workflow was tested on the original Manila clam data. Under basic settings, the results of the original study are largely replicated. The effect of additional parameter customization (clustering threshold, stringency, minimum number of replicates, seed matching) was further evaluated.

Conclusions: The study of smithRNAs is still in its infancy and no dedicated analytical workflow is currently available. At its core, the SmithHunter workflow builds over the bioinformatic procedure originally applied to identify candidate smithRNAs in the Manila clam. In fact, this is currently the only evidence for smithRNAs that has been biologically validated and, therefore, the elective starting point for characterizing smithRNAs in other species. The original analysis was readapted using current software implementations and some minor issues were solved. Moreover, the workflow was improved by allowing the customization of different analytical parameters, mostly focusing on stringency and the possibility of accounting for a minimal level of genetic differentiation among samples.

Keywords: SmithRNAs, Small non coding RNAs, Nuclear–mitochondrial interactions, Small transcriptome, Epigenetic regulation

Background

Proteins must interact with each other to assist in the dynamic processes of living cells in a functional organism. However, the intricate dance of protein interactions often deviates from the proportions dictated by their genome occurrences. Consequently, the modulation of gene transcription becomes a critical factor in maintaining the balance necessary for proper cellular function. One possible mechanism for achieving this equilibrium is through post-transcriptional modifications, a process frequently involving small noncoding RNAs—short fragments capable of modulating gene expression by silencing genes [1]. Among these entities, microRNAs (miRNAs) have emerged as some of the most pervasive gene-regulatory molecules in the animal kingdom [2–6].

Elements such as miRNAs guide the RNA-induced silencing complex (RISC) to regulate the translation of specific mRNAs through sequence complementarity [1, 7, 8] and their post-transcriptional silencing activity extends to various developmental processes and diseases [9–13].

Notably, miRNAs have been predominantly studied in the context of nuclear-nuclear interactions (i.e., miRNAs encoded in the nuclear genome that modulate the expression of nuclear transcripts), although some mitochondrial targets have also been described for miRNAs encoded in the nuclear genome [14]. At the other extreme, mitochondrial-mitochondrial interactions have been described, where mitochondrially-encoded microRNAs can actually regulate gene expression in the mitochondrion [15–18]. Only limited consideration, in turn, has been given to the mitochondrial genome as a potential source of RNA interference acting on the nuclear genome [19].

In animal cells, mitochondrial DNA (mtDNA) is a small (~ 16 Kb) molecule, that is usually characterized by the absence of introns, a circular structure, and a conserved repertoire of 13 protein-coding genes, two ribosomal genes and 22 tRNA-coding genes [20], but see [21]. Molluscan mtDNA is unique in many respects [22] including, among others, its peculiar process of vertical transmission to offspring observed in bivalves (DUI, Doubly Uniparental Inheritance) [23, 24], but generally conforms, at least in its structure, to the model previously described for animals [22–24]. Given its unique transcription mechanism, which involves the production of long transcripts that are further cleaved to produce single gene transcripts and liberate functional RNA molecules (rRNAs and tRNAs [25]), it is reasonable to hypothesize that the mitochondrial genome may serve as an efficient source of miRNA-like molecules [26]. This possibility has been explored in the Manila clam *Ruditapes philippinarum*, that was selected as a model species for the study of mitochondrially encoded microRNAs for several reasons. First of all, it harbors two genetically distinct mitochondrial genomes (male and female, inherited according to the DUI model [23, 27], thus allowing a solid establishment of the mitochondrial source of miRNAs by comparing the small transcriptome against a male or female background [19]. Moreover there is evidence that the mitochondrial genome (namely, the presence and activity of either the male or the female genome in the developing embryos) is involved in sex determination [28]. Finally, it is worth noting that, at variance with the typical metazoan mitochondrial DNA, bivalve mitochondrial genomes

are characterized by large intergenic spacers and unassigned regions [29, 30], that may be deployed to develop novel roles in the cell, including, following maturation, regulatory RNAs [26].

The small transcriptome of the Manila clam has been characterized in detail, and multiple highly transcribed small RNAs of mitochondrial origin have been identified [19] in silico. Two of these were further validated in vivo through RNA injection experiments that demonstrated their biological activity [31]. Results of these studies have highlighted an intriguing interplay between mitochondrial and nuclear transcripts, possibly leading to gonad formation in *R. philippinarum* [19, 31].

This, in turn, led to the proposal of smithRNAs (Small MITochondrial Highly-transcribed RNAs) as a novel class of small RNAs defined as (a) of mitochondrial origin, (b) highly transcribed, and (c) regulating a nuclear transcript [19]. While (b) is somewhat arbitrary and (c) requires that their function is confirmed experimentally by RNA injection, this provides a clear definition of the class. Worth of note, some overlaps are envisionable with other classes of small RNAs, such as tRNA fragments (tRFs [32, 33]), rRNA fragments (rRFs [34]) and degradation fragments [35, 36]. Nevertheless, the unique combination of structure/origin (i.e. mitochondrial, highly transcribed) and function (i.e. regulating a nuclear target) provides a clear and workable definition of this novel small RNA class.

Concurrently, other studies have proposed the involvement of smithRNAs in sex determination in the bivalve *Potamilus streckersoni* [37]. One smithRNA, encoded in the male mitochondrial genome, was identified and predicted to target a nuclear transcript that is a) differentially regulated in males vs. females, and b) presumably involved in female development. Henceforth, while not biologically validated in the strict sense, i.e. by RNA injection experiments as in [31], this smithRNA receives substantial support in [37]. Incidentally, while the term is never used in [37], its features nicely conform to the definition of smithRNAs introduced above.

While still awaiting in vivo validation, the presence of smithRNAs have been further suggested, based on bioinformatic analyses, in *Danio rerio*, *Drosophila melanogaster* and *Mus musculus* [31], where they appear to be characterized by a high degree of sequence conservation in line with other functional mitochondrial loci. Adding to their significance, it has been proposed that new smithRNAs can readily evolve from mitochondrial RNAs through an exaptation process [26]. This evolutionary trajectory, combined with a remarkably high probability of finding nuclear targets [26], underlines the central role of sncRNAs in mediating the elaborate interaction between mitochondrial and nuclear genomes during metazoan evolution.

Nevertheless, and despite secondary evidence for smithRNA in other species [31, 37], the only confirmed evidence for biologically functional smithRNAs, at present, comes from the Manila clam [31]. As such the question remains open whether smithRNAs are a species-specific mechanism related to sex determination in the Manila clam (i.e., an 'odd feature of an odd system' [19]), or rather a mechanism of more general interest, possibly shared by the entire Metazoa.

To evaluate the potential of this phenomenon as a novel signaling pathway in the broader context of mito-nuclear cross-talking, a comprehensive investigation across a panel of representative metazoan species is therefore imperative. This systematic

approach will contribute to our understanding of the functional implications of retrograde mitochondrial RNAi across animal lineages, with implications going as far as to the origin of the eukaryotic cell [26].

To this end, we developed SmithHunter, a new workflow designed for the identification and characterization of candidate smithRNAs. The pipeline reproduces, in a unified workflow, the original procedure used to identify smithRNAs in the Manila clam [19] and in other metazoans [31]. Improvements in this implementation rely on the possibility of using replicate samples, remapping on the nuclear genome to select reads of unequivocal mitochondrial origin, improved clustering and cluster filtering methods, and computation of free energy of pre-smithRNAs secondary structures. Moreover, multiple analytical parameters of SmithHunter can be customized, allowing users to adapt the analysis to the organism/data studied.

Implementation

Overview

The pipeline is composed of two main components (Fig. 1). The first module is essentially data-driven and focuses on the identification and filtering of presumptive smithRNA sequences, defined as centroids of clusters with significant transcription levels and narrow transcription boundaries. One or more small RNA libraries (replicates), the sequence of the mitochondrial genome and, optionally, the sequence of the nuclear genome of the species of interest, are used as inputs. The main output is a list of presumptive smithRNA sequences, as well as graphics depicting: (a) raw coverage over the mitochondrial genome; (b) cluster position/abundance on the mitochondrial genome; and (c) 5′ and 3′ transcription end conservation.

The second module is more predictive in nature and deals with the identification of possible nuclear targets and pre-miRNA-like precursor structures for presumptive smithRNAs. The list of smithRNAs identified by the first module and the transcriptome of the species of interest, with annotated 5′ and 3′ UTR regions, are used as inputs. The main output is a list of nuclear transcripts putatively targeted by individual smithRNAs, information regarding the Gibbs free energy (dG) of smithRNA/target pairs, and putative precursor structures. After passing filters in the first module, sequences that find a putative nuclear target in the second module are regarded as candidate smithRNAs. Presumptive and candidate smithRNAs are given a unique identifier (e.g. their name) in the form of a string reporting, in order, the cluster number, its depth, start and end position with respect to the genome and strand (e.g.: clusterid0_size11210_pos12846_12871_strand+; see below).

The main pipeline is written in bash and R [38], with calls to external software. The pipeline setup entails cloning of the GitHub repository and running an installation script that will install SmithHunter as well as, if required, conda [39] and PITA [40]. Additionally, the installation script will create a conda environment within which SmithHunter is executed. The two modules are invoked independently via command line and all inputs (file paths, options) are specified as command line arguments. The pipeline was created and tested on the Ubuntu 22.04.3 LTS platform, but is portable to any Linux OS provided that the conda environment can be successfully created and PITA can be installed.

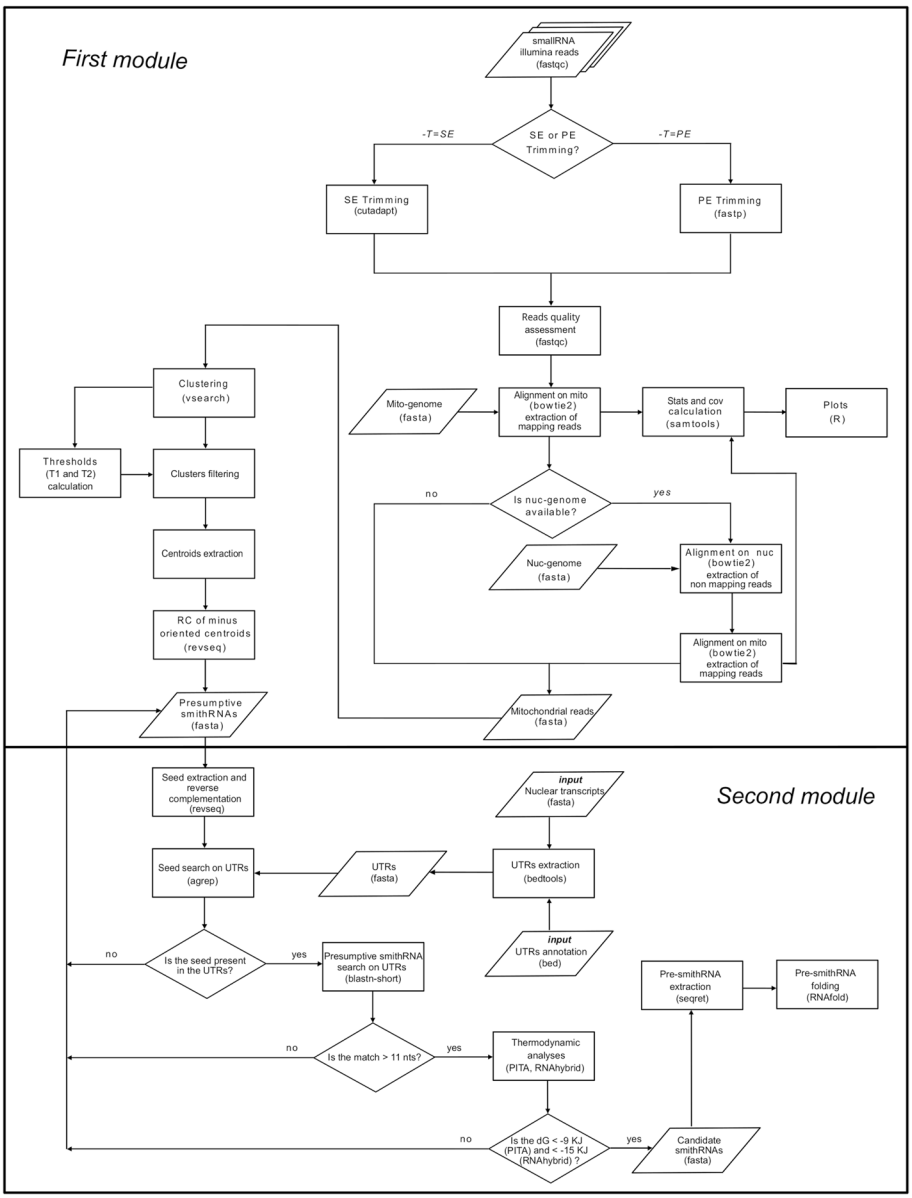

Fig. 1 SmithHunter workflow: first module (data to presumptive smithRNAs) and second module (presumptive to candidate smithRNAs)

Detailed installation and usage information, with examples, can be found in the project's GitHub repository https://github.com/ESZlab/SmithHunter.

First module

Sequence reads (fastq.gz files; multiple replicates are accepted) are preprocessed using CUTADAPT (ver. 4.3; [41]) if single-end (SE) or fastp (ver. 0.23.2; [42]) if paired-end (PE). This behavior is selected with option `-T`, which takes as values either `SE`, for single end, or `PE`, for paired end. Value `NO` can be used for pre-trimmed data. In PE mode, reverse reads are exploited to correct overlapping regions in forward reads and

subsequently discarded. Forward and reverse adapters are specified via options `-a` and `-A`, respectively.

The trimmed reads are initially aligned to the mitochondrial genome with Bowtie2 (ver. 2.5.1; [43]) using the end-to-end option and allowing a single nucleotide mismatch. The reason for this is that we are interested in microRNAs transcribed from the mitochondrial, and not the nuclear, genome [19]. Mitochondrion-mapping reads are extracted with SAMtools (ver. 1.13; [44]) and aligned, as above, to the nuclear genome, if available. This step is justified by the need to exclude reads that map to both the mitochondrial and the nuclear genome, as these may originate from nuclear mitochondrial pseudogenes (Numts; [45]) and therefore their mitochondrial origin cannot be assured. In the absence of a nuclear genome sequence for the species under study, this step is not performed. Nuclear nonmapping reads, representing *bona fide* mitochondrial-unique reads, are eventually remapped on the mitochondrial genome for counting and coverage calculation. Alignment statistics are calculated at each step using the *flagstats* and *depth* modules of SAMtools and mitochondrial-unique alignments are produced from each remapping file using the *bamtobed* module of BEDtools (ver. 2.30.0; [46]).

Mitochondrial-unique reads are sorted by abundance and clustered using VSEARCH (ver. 2.22.1; [47]). The percent identity used for clustering can be specified via option `-I`. Clusters are then subjected to several filters. To discard clusters characterized by low copy number and/or variable expression, two abundance thresholds are calculated based on the empirical distribution of cluster depths and the stringency parameter defined by option `-S`. One (T1, global) is calculated on combined reads and the second (T2, possibly different across replicates) is calculated on reads from individual replicates. Both thresholds are defined as the percentile of order S of the relevant distribution of unique cluster sizes. This definition of abundance thresholds is inherently different from the manually defined, hard coded, threshold of > 200 reads used in [19]. This option, nevertheless, has the advantage that different thresholds are applied to different datasets, thus accounting for differences in sequencing depth across species/replicates while retaining comparability across different datasets by the use of a unique stringency value.

Clusters not reaching T1 overall are discarded. Clusters are then filtered based on a minimum number of replicates where each cluster reaches T2. The number of replicates can be specified by the user via option `-M`. This step is different from that of [19], where data from different replicates were combined (equivalent to `-M 1`). We nevertheless envision the utility of setting `-M` depending on the nature of the samples and the expectations of the user (see the "Discussion" section). All the clusters not passing both filters are discarded, while the others are retained as multifasta files. Representative sequences for each cluster, selected as the most abundant within the cluster, are extracted and those mapping in minus orientation are reverse/complemented using the *revseq* module of EMBOSS (ver. 6.6.0.0; [48]) to obtain smithRNAs in 5′ to 3′ orientation. Custom R scripts are used to generate coverage plots of the mitochondrial alignments and visualize cluster size and distribution, as well as the distribution of the 5′ and 3′ ends of the reads within each cluster (i.e., transcription end conservation; equivalent to Figure 3 in [19]).

The endpoint of this first module is a multifasta of presumptive smithRNAs in 5′ to 3′ orientation. This list can be manually inspected and edited by the user. Among other possible ad hoc analyses, we envision the possibility that the user may visualize the

plot of transcription end conservation and manually select only those smithRNAs that appear to have the tightest 5′ and 3′ transcription boundaries. A script that automatically identifies smithRNAs with narrow 5′ and 3′ ends is provided to optionally help the user in this step (experimental, see GitHub repository for its documentation).

Second module

5′ and 3′ UTRs of nuclear transcripts are extracted from the transcriptome sequence as a multifasta file based on UTR annotations provided as a BED file. Seed regions (nucleotides 4–10, as in [19], see below) of each smithRNA are identified, reverse/complemented and searched against 5′ and 3′ UTRs of nuclear transcripts using the approximate grep algorithm (`agrep`; [49]). For each matching target, UTR regions are converted in a BLAST database and the full length sequence of the smithRNA is used as query in a BLAST search executed with the blastn-short option [50]. A minimum of 11 nucleotides aligned on the minus strand (alignment length) are necessary for the presumptive smithRNA and target pair to pass to the next step. Following the observation that the default e-value threshold can bias the results in favor of smithRNAs with a limited number of targets (i.e. due to differences in database size) the e-value threshold was removed and filtering is therefore based on seed matching, alignment length and RNA/RNA stability.

While the definition of the seed as nucleotides 4–10 was retained, based on [19], we acknowledge that different options exist. Furthermore, while we tentatively assimilate smithRNAs with miRNAs, the exact fine scale molecular processes involved in smithRNA-target interaction are largely unknown. Through sequence complementarity, miRNAs and their targets interact into a seed region represented by nucleotides 2–8 of the 5′ region of the miRNA and the 3′ untranslated region (UTR) of the target mRNA [51]. Although a perfect seed match between the miRNA and its target is generally needed, noncanonical matches, as well as cases where the seed is shifted, have been reported and appear to be common [4, 52, 53]. Moreover, central seed pairing has been reported to be more predictive in mammals [53]. To overcome a fixed definition of the seed region, and take into account noncanonical seed matches and seed shifts, the possibility was implemented for the user to define the seed region (options `-X` and `-Y` for the first and last nucleotides, included) and to allow 0 (by default) to 2 mismatches in the seed region (option `-m`).

Presumptive smithRNAs with evidence of similarity with UTRs in the seed region are subjected to additional thermodynamic analyses using PITA (ver. 6; [40]) and RNAHybrid (ver. 2.1.2; [54]). Those smithRNA-target pairs with dG levels < -15 kJ and 3–10 helix constraints from RNAhybrid, as well as DG < -9 kJ from PITA, are eventually retained. The former threshold (< -15) differs from the threshold (< -20) applied by [19, 31]. This finds a justification in the observation that the dG values calculated here are marginally, but consistently lower than those reported in [19]. This observation, coupled with the fact that the current workflow is aimed at finding candidate smithRNAs that need to be further validated biologically, and therefore a false positive is better tolerated than a false negative, suggested that a less stringent position was to be taken in this respect.

Presumptive smithRNAs that find at least one target in the nuclear transcriptome are further referred to as candidate smithRNAs. Putative pre-smithRNA sequences of candidate smithRNAs are extracted using the *seqret* module of EMBOSS [48]. Under the assumption that, in line with miRNAs, mature RNAs can be located on either strand of the pre-miRNA hairpin, we tentatively identified pre-smithRNA sequences as either region −15 to +50, or region −50 to +15, relative to the smithRNA. The secondary structure of both possible pre-smithRNAs is computed using RNAfold (ver. 2.3.3; [55]) at the default folding temperature of 25 °C and both structures are reported. Users can adjust the folding temperature through the `-R` parameter and pre-smith coordinates with the `-1` and `-2` parameters. The identification of putative pre-smithRNA regions differs from that of [19] and [31], who, in turn, conducted this calculation on a region that was manually selected based on gene annotation of the mitochondrial genome. Briefly, the entire noncoding region encompassing the smithRNA was selected in the case of smithRNA that were found within a noncoding region, while the entire tRNA, plus neighboring small noncoding regions, was selected in the case of smithRNAs that were found within a tRNA. This is a sensible option in the Manila clam, whose mitochondrial genome is characterized by multiple large noncoding regions, and has been convincingly justified by the hypothesis that noncoding regions, released during the splicing of the initial mitochondrial transcript, may actually act as pre-smithRNAs [26]. However, this approach does not seem feasible in general, as metazoan mitochondrial genomes are almost invariably devoid of intergenic noncoding regions of sufficient size. As such, we revert to a more standard working definition of putative pre-smithRNAs that does not rely on the presence of large intergenic spacers.

Results

Fourteen candidate smithRNAs were identified in the Manila clam [19]. Subsequently, the activity of two of these genes was validated in vivo by [31], thus establishing this species as the reference organism for smithRNA studies. Given the absence of alternative software for smithRNA detection, evaluating the reproducibility of the results obtained by [19] for the Manila clam, with special attention given to the two validated smithRNAs, appears to be the most effective way to assess and present the functionalities of the SmithHunter pipeline.

We reproduced the analysis performed by [19] using the same data, with minimal modifications due to specific SmithHunter functionalities and parameters. Specifically, raw reads of the small transcriptome of *R. philippinarum* were downloaded from the NCBI Short Read Archive (SRA; accession numbers SRR3662624-SRR3662629), while male and female mitochondrial genomes, as well as the nuclear genome of the species, were recovered from GenBank (accession numbers AB065375.1, AB065374.1, and GCA_026571515.2, respectively). The original, unannotated transcriptome used in [19] was downloaded from GenBank (accession numbers JO101212-JO124029; [56]). In the absence of the original UTR annotations, updated annotations, produced by the same authors, were retrieved from [57]. The SmithHunter analysis was performed, independently and in parallel, on the male and female small transcriptome and mitochondrial genomes (see Commands in Supplementary Materials). Note that the species name

Venerupis philippinarum, used in some records, is a synonym of *Ruditapes philippinarum* (WORMS database; [58]).

Additional testing was performed on data from *P. streckersoni*, focussing on a male-related smithRNA that, while not biologically validated in strict terms, received substantial support as a masculinizing agent and regulator of the nuclear transcript for *GCNT1* [37]. Raw reads of the small transcriptome were obtained from SRA (accession numbers SRR23195578-SRR23195582, SRR23195559), while the male and female mitochondrial genomes, as well as the nuclear genome, were recovered from GenBank (accession numbers ON881148, MW413895, and JAEAOA01, respectively). Transcriptome UTR sequences were in turn retrieved from Supplementary Materials to [37] (GitHub: https://github.com/raqmejtru/mitonuclear-sd). The SmithHunter analysis was performed, as above, independently and in parallel on the male and female small transcriptome and mitochondrial genomes, under parameters designed to replicate [37] as well under a more stringent parameter set stemming from our parameter optimization (see below).

Trimming and remapping

A total of approximately 69 and 72 million reads belonging to male and female individuals of *R. philippinarum*, respectively, (21 to 27 million in individual replicates) were analyzed using the procedure implemented in the first module of SmithHunter. Reads were trimmed in SE-mode (option `-T SE`) and aligned to the mitochondrial genomes of both sexes, as well as to the nuclear genome of the species. There were 207,149 and 276,748 reads mapped on the mitochondrial genome for the male and female, respectively, representing 0.31% and 0.39% of the trimmed reads. Among these, 187,735 and 259,719, representing 0.28% and 0.36%, respectively, of the trimmed reads, were identified as mitochondrial-unique reads: i.e., they did not align to the nuclear genome (Table S1, Fig. 2).

The addition of a remapping step to the nuclear genome, which was not available to [19], confirmed what was hypothesized herein, i.e., that the source of the vast majority of mitochondrion-remapping reads is the mitochondrial genome and not the nuclear genome. In fact, only 9.37% of the mitochondrial reads mapping to the male mitochondrial genome, and 6.15% of reads mapping to the female genome, map to the nuclear

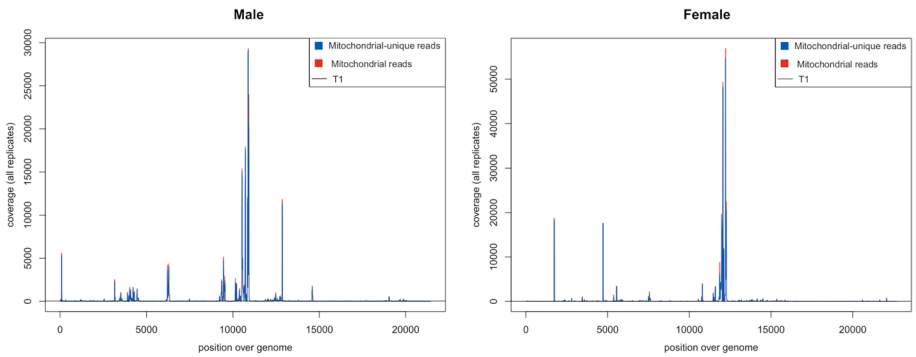

Fig. 2 Remapping of the small transcriptome over the male and female mitochondrial genomes. Coverage along the mitochondrial genome (all replicates, combined) is shown based on mitochondrial reads (red) and mitochondrial-unique reads (blue). The black horizontal line represents the T1 coverage threshold (`-S 0.50`). Image from the standard SmithHunter output

genome as well. Despite suggestive evidence in [59], these reads are cautionary considered to be of uncertain origin in the SmithHunter pipeline, possibly originating from nuclear mitochondrial pseudogenes (Numts; [45]), and are discarded.

For each of the six sequencing libraries and for the two sexes, the number of raw reads, the number/percentage of reads after trimming, the number/percentage of trimmed reads mapping to the mitochondrial genome and the number/percentage of trimmed reads uniquely mapping to the mitochondrial genome (i.e., not mapping to the nuclear genome), is reported in Table S1.

Clustering

Mitochondrial-unique reads were clustered, and each cluster was filtered for size using the procedure implemented in the first module of SmithHunter. In line with [19], clustering was performed at 99% identity (option `-I 0.99`). The stringency was set as the 50th percentile of unique cluster depth (option `-S 0.50`), which corresponds to 125 and 116 reads for males and females, respectively. The minimum number of replicates was not enforced (option `-M 1`). A total of 89 and 97 clusters were observed that passed all the thresholds for the male and female genomes, respectively. All the clusters appear to be on the plus strand, the same strand from which all the mitochondrial genes are transcribed in the species (Fig. 3; [60]).

Among these clusters, 42 and 32 exhibited conserved 5′ transcription ends in the male and female genomes, respectively, and their centroid sequences were considered

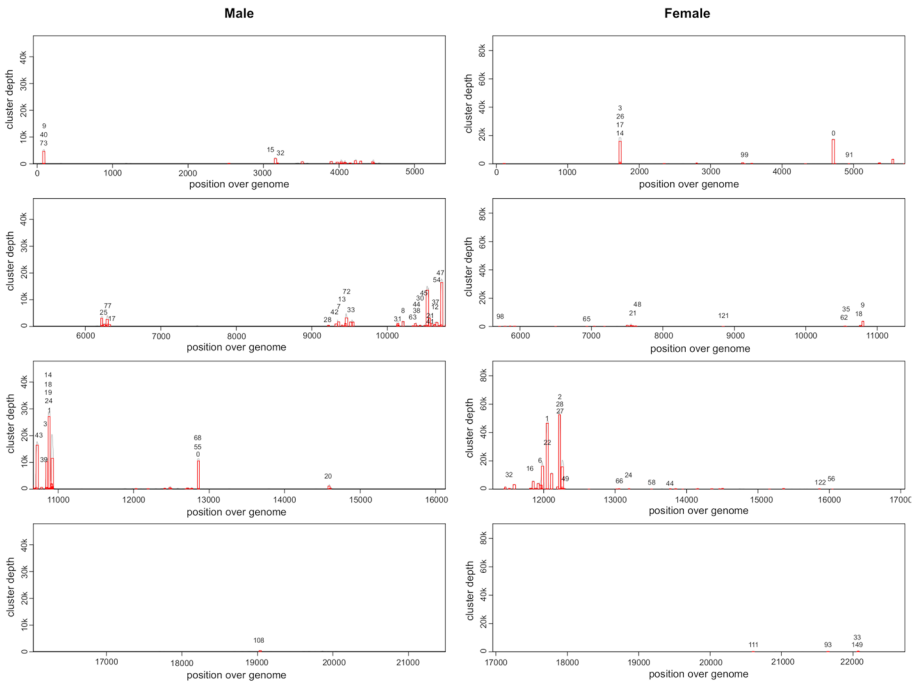

Fig. 3 Distribution and depth of clusters along the male and female mitochondrial genomes. Grey peaks in the background represent total coverage. The horizontal black line represents the T1 coverage threshold (`-S 0.50`). All clusters are observed on the plus strand and are shown in red (clusters on the minus strand would be shown in blue). To improve readability, only clusters giving rise to candidate smithRNAs are numbered. Image from the standard SmithHunter output

presumptive smithRNAs. All 14 smithRNAs selected by [19] (hereafter referred to as reference smithRNAs) were found to be presumptive smithRNAs in our analyses, and their coverage at the 5′ and 3′ transcription ends was almost identical to that previously reported (Fig. 4).

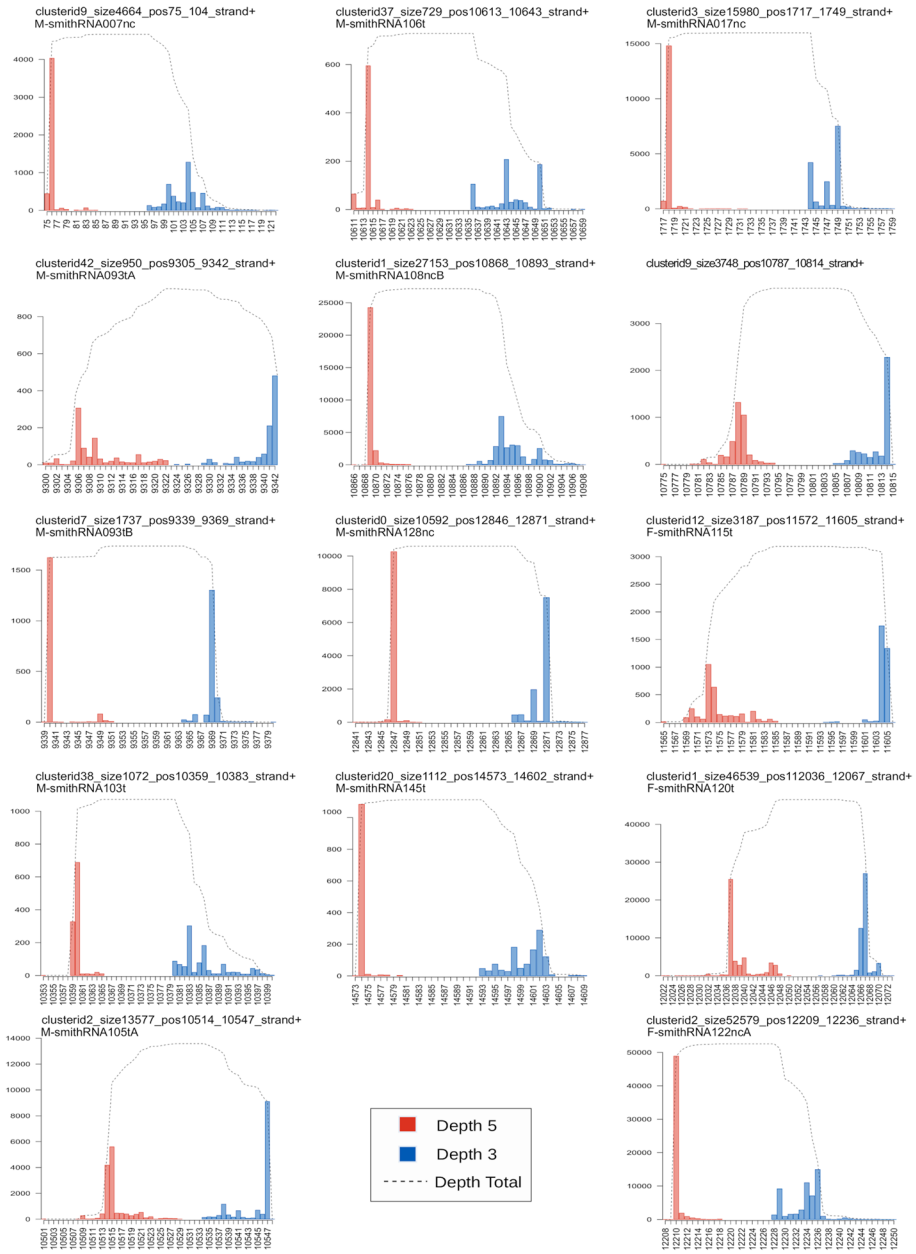

Fig. 4 Transcription end conservation of the 14 reference smithRNAs. Names of smithRNAs are shown following the nomenclature of SmithHunter as well as following [19]. Red and blue bars represent the distribution of unique start/ends of reads mapping on the mitochondrial genomes, respectively. The black dotted line represents overall, per base, coverage. Genome positions on the horizontal axis refer to sequences AB065375.1 (male mitochondrial genome) and AB065374.1 (female). Image from the standard SmithHunter output

Table 1 Reference smithRNAs-target pairs from [19] recovered using SmithHunter

SmithRNA	Target	dG PITA	dG RNAhybrid	Protein name
M-smithRNA108ncB	Locus_4366	− 26.5	− 29.4	Dynein 1 Heavy Chain 1
M-smithRNA145t	Locus_2953	− 22	− 24.5	DNA polymerase epsilon
M-smithRNA106t	Locus_14	− 15.2	− 17	Histone-lysine N-methyltranferase SETD2
M-smithRNA103t	Locus_1096	− 18.4	− 22.7	Microsomal triglyceride transfer protein large subunit
M-smithRNA103t	Locus_6539	− 18.4	− 19.4	Centrosomal protein of 131 kDA
M-smithRNA103t	Locus_62148	− 19.4	− 19.8	Kinein-like protein KIF21A
M-smithRNA093tA	Locus_15177	− 16.4	− 15.8	Elongator complex protein 5
M-smithRNA093tB	Locus_2534	− 15.9	− 17.6	Trifunctional enzyme subunit alpha, mitochondrial-like
M-smithRNA007nc	Locus_5815	− 17.2	− 18.2	U3 small nucleolar RNA-associated protein 6
M-smithRNA007nc	Locus_3650	− 19.2	− 18	eukaryotic translation initiation factor 3 subunit I
F-smithRNA120t	Locus_472	− 20.2	− 19.6	Serine/arginine repetitive matrix protein 2-like
F-smithRNA122ncA	Locus_15925	− 17.6	− 27.9	Nuclear Receptor Subfamily 0
F-smithRNA107t	Locus_31245	− 18.5	− 23.5	Cullin-5-like

The first column reports smithRNA names as in the original paper. Additional columns indicate the nuclear target, Gibbs Free Energy (dG, in kJ) of RNA-RNA hybrids calculated using PITA and RNAHybrid and the human homolog of the target according to UniProt

Target prediction

Presumptive smithRNAs were searched against UTR regions of the *R. philippinarum* transcriptome using the procedure implemented in the second module of SmithHunter. A total of 39 and 30 presumptive smithRNAs from the male and female genomes, respectively, were hypothesized to target at least one nuclear transcript and are here referred to as candidate smithRNAs. Of the total of 69 candidates, 12 were reference smithRNAs (see [19]), and 10 of these were associated with the same 13 nuclear target hypothesized herein (Table 1). Most importantly, the two reference smithRNAs that were validated in vivo by [31] (M_smithRNA106t and M_smithRNA145t) passed through all the filters and were associated with the previously documented targets. In particular, M-smithRNA106t was hypothesized to target the Manila clam homolog of the human Histone-lysine N-methyltransferase and M-smith145t was hypothesized to target the Manila clam homolog of the human polymerase epsilon (Table 1). Incidentally, 11 out of 24 smithRNA-target pairs reported by [19] were not identified in the current analyses. These sequences were considered individually, and the smithRNA sequence was searched over the UTR regions of the *R. philippinarum* transcriptome using BLAST. A total of 10 out of the 11 smithRNAs actually matched the previously reported target in the plus-plus orientation, suggesting a problem with a directionality filter in the original analysis; these were subsequently excluded from the analysis.

In its basic implementation, SmithHunter was therefore shown to largely duplicate the results of [19]. In the following section we aim at additional testing and the deployment of additional features implemented in SmithHunter.

Parameter optimization

The effects of user choice on selectivity related parameters, a new implementation of SmithHunter that was not available to [19] and [31], were evaluated by repeating the

analysis under 27 different parameter combinations as well as allowing 0 to 2 mismatches in the seed region (see below).

All combinations of cluster identity (option `-I`; tested: `0.99`, `0.95`, `0.90`), stringency (option `-S`; tested: `0.2`, `0.5`, `0.8`) and minimum number of replicates (option `-M`; tested: `1`, `2`, `3`) were studied on both the male and female data. Predictably, the number of presumptive smithRNAs identified by the first module was directly correlated with the identity (I) parameter and inversely correlated with the replicates (M) and stringency (S) parameters. By adopting the less selective parameter combination (`-I 0.99 -S 0.2 -M 1`) a total of 71 and 65 presumptive smithRNAs were identified for males and females, respectively. Out of this set, 67 candidate smithRNAs, targeting 320 nuclear genes, were identified for males and 59 candidate smithRNAs, targeting 323 nuclear targets for females (Table S2, Fig. 5). In contrast, by using the most selective parameter combination (`-I 0.90 -S 0.8 -M 3`), eight and six presumptive smithRNAs were found in male and female individuals, respectively. Six and five of these were identified as candidate smithRNAs and were associated with 25 and 16 nuclear targets, respectively, in males and females (Table S2, Fig. 5).

Taking the number of presumptive smithRNAs produced at the end of the first module as a proxy for selectivity, it was possible to visualize the effects of different parameters. Stringency (S) had the most substantial effect, leading to the filtering, in the range examined, of 74.5% of clusters. The number of replicates (M) had a more limited effect, with the filtering of 48.9% of clusters. The cluster identity parameter

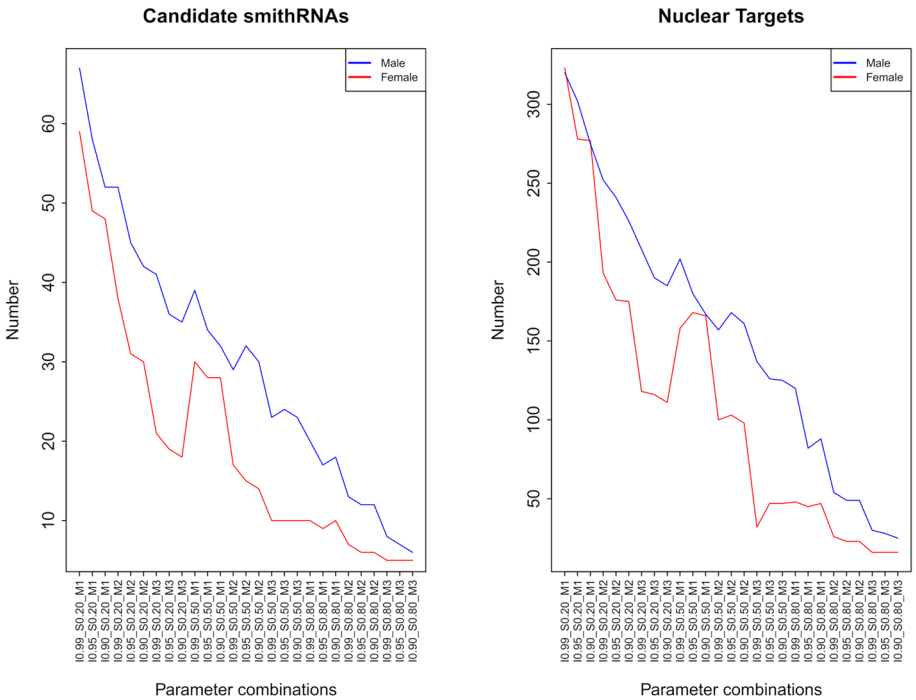

Fig. 5 Number of candidate smithRNAs and of their associated nuclear targets recovered under different parameter combinations. Parameter combinations are indicated on the horizontal axis as follows: I[identity parameter]_S[stringency parameter]_M[replicates parameter]

(I), on the other hand, displayed a minimal effect, at least on the Manila clam data, leading to the filtering of 11.5% of clusters (Figure S1).

Notably, reference smithRNAs from [19] were often identified even under the most selective criteria. Eight out of twelve reference smithRNAs were identified under all combinations tested, while the remaining four were identified under a minimum of 21 parameter combinations (Fig. 6). Moreover, candidate smithRNAs identified under the most selective criteria were also identified under all the tested combinations. This finding suggested that the most selective parameter combinations were optimal starting points and that the resulting candidate smithRNAs, in turn, constitute a reasonably reliable group and starting point for subsequent validation analyses.

Loosening of the selectivity criteria, in turn, leads to an expansion of the results from the highly confident core. Given the possibility of customizing selectivity parameters, users have the flexibility to balance the number of resulting candidates with their level of confidence by acting on the parameters. However, the level of selectivity should be consciously evaluated case by case, depending on the biological system under scrutiny, the data, the number of biological replicates and, in turn, the intended use of the results.

In the end, smithRNAs and their associated nuclear targets identified under the most selective parameter combination (`—I 0.90 —S 0.80 —M3`) are reported in Table 2 and Table S3.

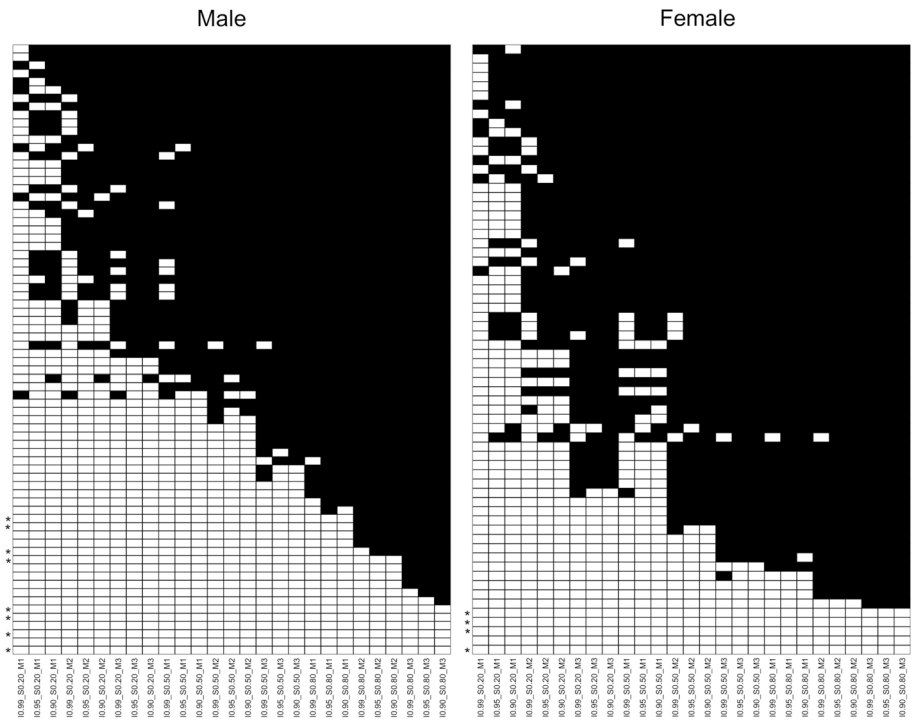

Fig. 6 Candidate smithRNAs recovered under different parameter combinations for the male and female genomes. Rows represent all candidate smithRNAs recovered under all parameter combinations. Reference smithRNAs are indicated by an asterisk. Columns represent parameter combinations and are labelled as follows: I[identity parameter]_S[stringency parameter]_M[replicates parameter]. White color in cross cells means that the candidate smithRNA was recovered under the specific parameter combination, black color means that it was not recovered

Table 2 SmithRNAs and their targets identified under the most stringent parameter combination in males and females

Sex	SmithRNA	Target	dG PITA	dG RNAhybrid	Protein name
M	clusterid0	Locus_690	− 17.1	− 15.0	Microtubule-associated protein RP/EB family member 1-like isoform X2
M	clusterid0	Locus_1177	− 12.39	− 15.2	Hydrocephalus-inducing protein homolog isoform X19
M	clusterid0	Locus_1711	− 13.7	− 22.2	Uncharacterized protein LOC123531196 isoform X12
M	clusterid0	Locus_4783	− 16.9	− 21.4	Integrator complex subunit 2-like
M	clusterid0	Locus_9761	− 14.3	− 16.2	Matrix-remodeling-associated protein 7-like
M	clusterid0	Locus_3140	− 14.23	− 15.0	Dynein axonemal heavy chain 2-like isoform X3
M	clusterid0	Locus_4916	− 21.7	− 20.7	Dynein heavy chain domain-containing protein 1-like
M	clusterid0	Locus_16697	− 13.29	− 16.2	WD repeat-containing protein 1-like
M	clusterid15	Locus_7476	− 23.4	− 27.2	Uncharacterized protein LOC123561027
M	clusterid15	Locus_7478	− 15	− 16.5	Hypoxia-inducible factor 1-alpha
M	clusterid15	Locus_3651	− 12.9	− 18.7	Protein polybromo-1-like isoform X13
M	clusterid1	Locus_4366	− 26.5	− 29.4	Cytoplasmic dynein 1 heavy chain 1-like isoform X6
M	clusterid3	Locus_1455	− 22.6	− 20.9	Uncharacterized protein LOC123556528
M	clusterid3	Locus_3369	− 15.35	− 17.7	Pericentriolar material 1 protein-like isoform X5
M	clusterid3	Locus_34514	− 22.2	− 23.6	Uncharacterized protein LOC123532869
M	clusterid3	Locus_3229	− 23.7	− 19.4	MAM and LDL-receptor class A domain-containing protein 2-like
M	clusterid7	Locus_2534	− 15.9	− 17.6	Trifunctional enzyme subunit alpha, mitochondrial-like isoform X1
M	clusterid7	Locus_2153	− 12.25	− 16.3	ATP-dependent translocase ABCB1-like isoform X2
M	clusterid7	Locus_17431	− 14	− 16.1	Putative inhibitor of apoptosis
M	clusterid7	Locus_1312	− 11.9	− 15.4	Testis-expressed protein 45-like
M	clusterid9	Locus_5815	− 17.2	− 18.2	U3 small nucleolar RNA-associated protein 6 homolog
M	clusterid9	Locus_3650	− 19.2	− 18.0	Eukaryotic translation initiation factor 3 subunit I-like
M	clusterid9	Locus_9171	− 16.1	− 17.3	Peripheral-type benzodiazepine receptor-associated protein 1-like isoform X3
M	clusterid9	Locus_5848	− 14.2	− 15.0	E3 ubiquitin-protein ligase rnf213-alpha-like isoform X2
M	clusterid9	Locus_3816	− 17.9	− 17.3	DNA-directed RNA polymerases I, II, and III subunit RPABC3-like
F	clusterid0	Locus_1177	− 15.1	− 15.2	Hydrocephalus-inducing protein homolog isoform X19
F	clusterid0	Locus_7826	− 12.3	− 21.3	Zinc finger and BTB domain-containing protein 17-like isoform X1
F	clusterid0	Locus_1982	− 12.51	− 15.1	Uncharacterized protein LOC123551470
F	clusterid0	Locus_4867	− 15.7	− 19.2	Zinc finger and BTB domain-containing protein 17-like isoform X1
F	clusterid0	Locus_1117	− 9.9	− 18.3	Dynein axonemal heavy chain 3-like
F	clusterid1	Locus_472	− 20.2	− 19.6	msx2-interacting protein-like isoform X2
F	clusterid1	Locus_25748	− 25.91	− 16.9	SH3 and multiple ankyrin repeat domains protein 2-like isoform X2
F	clusterid1	Locus_2909	− 9.5	− 18.3	Syntenin-1-like
F	clusterid1	Locus_1117	− 18.6	− 30.1	Dynein axonemal heavy chain 3-like
F	clusterid2	Locus_15925	− 17.16	− 27.9	Uncharacterized protein LOC123527919
F	clusterid3	Locus_13268	− 15.3	− 16.4	Transcription initiation factor TFIID subunit 1-like isoform X1

Table 2 (continued)

Sex	SmithRNA	Target	dG PITA	dG RNAhybrid	Protein name
F	clusterid3	Locus_1747	− 14.1	− 17.7	Dynein axonemal heavy chain 6-like isoform X2
F	clusterid3	Locus_6770	− 17.9	− 17.5	Acetyl-CoA acetyltransferase, mitochondrial-like
F	clusterid9	Locus_3552	− 26.8	− 20.2	Cilia- and flagella-associated protein 47-like isoform X8
F	clusterid9	Locus_31245	− 18.5	− 23.5	Cullin-5-like isoform X1
F	clusterid9	Locus_1096	− 25.1	− 16.1	Apolipophorins-like

Name of candidate smithRNA (simplified to cluster id) is given in the second column. Other columns report the targeted transcript, Gibbs Free Energy (dG, in kJ) of RNA–RNA hybrids calculated using PITA and RNAHybrid, as well as the targeted gene name. See Table S3 for extended information

Their length distribution, compared with those of the total reads mapped to mitochondrial genomes, is shown in Figure S2. The length distribution compares well with the results of [19] (compare to Fig. 1 therein) over the 22–35 bp range. This finding further suggested that, unlike miRNAs, smithRNAs may exhibit broad variation in length, with substantial peaks in the 20–34 bp range. Noteworthy, this may have a relevance in the context of smithRNA maturation and AGO2 binding, as proposed by [61].

At least six candidate smithRNAs obtained under the most selective parameter combination were predicted to form *bona fide* pre-miRNA-like harpins (Fig. 7). Their free energy is generally lower than that presented in the original study and their shape, in most cases, better conforms to the expectation of a long hairpin. This, in turn, supports the tentative identification of pre-smithRNAs based on position rather than on the span of tRNA and unassigned region annotations in the mitochondrial genome.

In order to evaluate the effects of allowing non-perfect alignments in the seed region in the target identification step, target identification was repeated under all parameter combinations allowing no mismatch in the seed region (as above) as well as allowing 1 or 2 mismatches (option `-m`).

The number of candidate smithRNAs did not increase significantly when mismatches were allowed in the seed region. With one mismatch allowed, up to two additional candidate smithRNAs were identified under the less selective parameter combinations, and no additional candidate was identified by allowing two mismatches. Conversely, the number of targets identified for each candidate smithRNA increased significantly if nonperfect alignments in the seed region were considered (Fig. 8). Compared to the case in which no mismatch was allowed, allowing for one mismatch led to the identification of almost twice as many targets (average 1.76 × in male data and 2.03 × in female data across different parameter combinations). Allowing two mismatches, in turn, did not lead to a further increase in the number of identified targets (average 1.01 × and 1.05 ×, respectively; Fig. 8). According to this evaluation, allowing mismatches in the seed region appears to have a marginal effect on candidate smithRNA identification. On the other hand, in the context of target identification, allowing mismatches in the seed region leads to a—possibly unwarranted—increase in targets that are less, or marginally, supported. As such, our advice is not to use this option in standard applications of SmithHunter, i.e. where

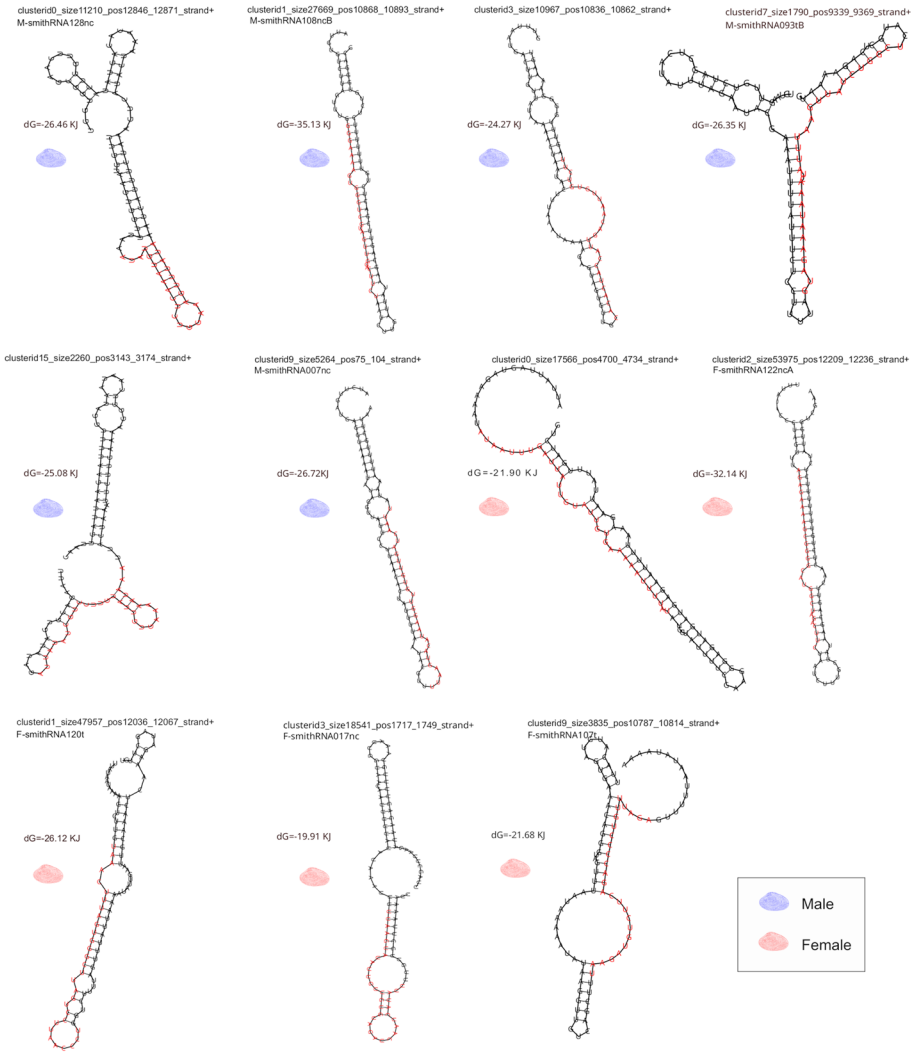

Fig. 7 Putative secondary structure of pre-smithRNAs for smithRNAs obtained under the most selective parameter combination. Structure labels report the name of the gene, dG of the folded structure, sex and, if available, the corresponding name in [19]. SmithRNA sequences are highlighted in red. Image modified from the standard SmithHunter output

the purpose is to identify a restricted number of high confidence candidate smithRNAs and their targets. This option, in turn, may be used in specific contexts, such as the evaluation of imperfect alignments in identifying one specific target that has previously been validated based on external evidence, or a study set of known targets for parameter optimization. Moreover, given that no difference was observed between one or two mismatches allowed, we suggest using one mismatch, if any, to reduce the run time of the second module.

Additional testing on *P. streckersoni*

Additional testing in *P. streckersoni* entailed the use of two different sets of analytical parameters and was conducted on both male and female data, focusing on the retrieval of smithRNA M-9, that received substantial support in [37]. The first parameter

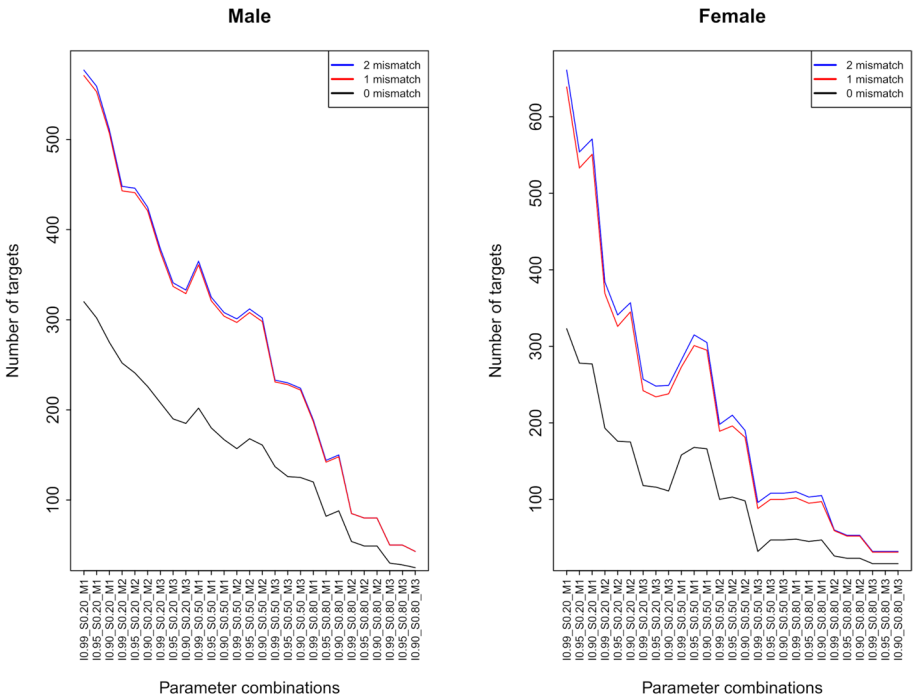

Fig. 8 Number of targets identified as a function of the number of mismatches allowed in the seed region. Parameter combinations, on the horizontal axis, are labelled as follows: I[identity parameter]_S[stringency parameter]_M[replicates parameter]

combination was designed to be grossly similar, in terms of stringency, and acknowledged the differences among the SmithHunter procedure and the procedure applied in [37], to the analysis performed in [37]. Stringency parameters were applied as: `-I 0.99 -S 0.5 -M 1`, the nuclear filter was not applied, and end conservation was evaluated following the guidelines in [37], that appear to be more relaxed than in our protocol. Compared to our testing regime, this corresponds to a medium/low stringency. A total of 28 out of the 33 smithRNAs identified in [37], namely 8/9 in males and 20/24 in females, were recovered by SmithHunter as candidate smithRNAs.

The second parameter combination applied (`-I 0.90 -S 0.8 M 2`) corresponds the most stringent parameter set in our testing regime. Here, and at variance with [37], the nuclear filter was also applied to exclude small RNAs of uncertain (nuclear or mitochondrial) origin, and end conservation was evaluated following the guidelines in [19]. A total of 10 out of the 33 smithRNAs identified in [37], 5/9 in males and 5/24 in females, were recovered as candidate smithRNAs by SmithHunter. Most importantly, the focal smithRNA M-9, that was singled out in [37] as the best candidate based on differential, sex related, expression, as well as its putative role in the sex determining pathway, was recovered by SmithHunter even under this most stringent parameter combination in association with its proposed target (*GCNT1*) according to [37].

Based on a comparison among the two runs, the reduction of the number of candidate smithRNA from the medium/low stringency to high stringency was due to the nuclear filter (5 candidates), coverage thresholds and replicate filter (4) and end conservation (4) or a combination thereof (5).

Discussion

As outlined above, the starting point of this work was the analysis of Manila clam data performed by [19] who, in turn, led to the first proposal of the existence of smithRNAs and to the identification of the only two smithRNAs that have been biologically validated [31]. While the analytical procedure is described in sufficient detail in the original paper, the actual bioinformatic implementation is no longer usable because some tools have been updated/surpassed, and more generally the script has been further modified and improved through time [31]. The purpose of this work was therefore to retain the original implementation whenever possible, modify outdated tools with current implementations whenever needed, and increase the level of customizability as appropriate for its intended use across different animal species. Incidentally, we share the view in [19] that at this early stage of smithRNA research it is preferable to avoid machine learning algorithms specifically trained on different types of small non coding RNAs and rely on more transparent parameters such as binding energy and sequence matching.

Testing, *ex post*, of the potential of the new analytical pipeline to duplicate the results of [19] in *R. philippinarum* gave positive results. Most smithRNAs and smithRNA/target pairs hypothesized in the original study were recovered. Some differences were nevertheless observed. Among the 14 candidate smithRNAs identified in the original study, 12 were confirmed in the present study. The two unconfirmed reference smithRNAs were found to be associated with nuclear targets previously reported by [19], but in plus/plus direction. Additionally, our analyses identified 13 of 24 reference smithRNA-target pairs. Once more, the associations not detected in this study mainly exhibited a plus/plus orientation, suggesting a possible failure in a directionality filter in the original implementation. Incidentally, both smithRNA/target pairs described in [19], whose biological function has been confirmed experimentally in [31], were recovered in the correct orientation.

Additional testing on *P. streckersoni* confirmed this view. SmithHunter was in fact capable to identify almost all smithRNAs considered in [37] at medium/low stringency and, most significantly, to recover the focal M-9 smithRNA, in conjunction with its target, even under the most stringent parameter combination.

Concerning the possibility, in SmithHunter, to customize different analytical steps and thresholds, we propose some a priori considerations of their possible applicability, and, at the same time, we tested the behavior of the script across different parameter combinations. The minimum identity for clustering, stringency and minimum number of replicates were under scrutiny, as well as the possibility to allow mismatches in the seed region.

Clustering identity reflects the minimum identity of reads that are combined in a cluster. The purpose of this parameter is to reflect the expected level of genetic variability across replicates. In an ideal situation where the genetic background is actually identical (i.e., all libraries come from one single individual and/or a single mitochondrial background), the parameter may be safely set to 0.99, whereas if libraries come from genetically different individuals some flexibility (0.95 or 0.90, grossly corresponding to 1 or 2 mutations out of approximately 20–25 bases) is to be allowed to avoid excess splitting of clusters across replicates.

The stringency parameter relates to the coverage (i.e., cluster size) threshold. In fact, higher stringency results in a higher threshold for cluster size. This definition of the cluster size threshold was preferred to a hard-coded indication of a minimum ad hoc cluster size (as in the original implementation) to foster comparability across libraries with unequal sequencing output and different species. We envision that, if the user is interested in a limited number of highly expressed clusters, this parameter may be set to 0.8 or higher. If, in turn, the operator is willing to expand its exploration to smithRNAs characterized by medium–low expression, thus maximizing the discovery rate, this parameter may be decreased. A medium level of filtering should nevertheless be retained at this step to avoid the background noise that is generally observed in sequencing libraries, possible associated to the widespread presence of degradation fragments [35], and becomes evident in the coverage plot of mitochondrial reads.

The intended use of the replicate parameters, i.e., the minimum number of replicates where a cluster reaches the minimum coverage threshold, in turn relates to the nature of the replicates in the study system and the expectations of the operator concerning the nature of the observed variability. If replicates come from different tissues/conditions/developmental stages and the operator is interested even in smithRNAs that may be expressed in one condition only, thus maximizing the discovery rate, this parameter may be set to 1 (in fact mimicking the original study). If, in turn, the operator is interested only in smithRNAs that are consistently expressed across conditions, thus reducing the number of false positives, the parameter may be set to the number of replicates or to the number of replicates minus one.

The effect of parameter customization was assessed analytically by exploring multiple parameter combinations. The most influential parameter was stringency (S), whereas replicates (M) had a more limited effect and identity (I) a minimal effect (Figure S1). Notably, while we expect that the stringency parameter will be influential regardless of the study system, the more limited effect of the replicate and cluster identity parameters observed here may be related to specific features of the Manila clam data. More specifically, tissue uniformity, as only gonads were used, may have led to a reduction of the relevance of the replicate parameter, and the substantial genetic uniformity, as all the samples came from one and the same location/sampling date, may have led to a reduction of the relevance of the clustering parameter. These two latter parameters may, in turn, become more influential in different study systems characterized by higher genetic and tissue diversity.

Allowing of mismatches in the seed region led to minimal differences in the number of candidate smithRNAs identified but, in turn, to a large, possibly unwarranted, expansion of the number of targets identified. Based on this observation, allowing of mismatches is recommended only in specific contexts (see above).

Relevant for target identification within the second module is the availability of a high quality transcriptome sequence, inclusive of reliable 5′ and 3′ annotations. While this is not always the case, especially in non-model organisms, it can be noted that the most common inaccuracies relate to missing and/or incomplete transcripts, as well as a poor identification of splicing variants, while the occurrence of supernumerary transcripts is more of an unlikely occurrence. As such, using a low-quality transcriptome sequence is liable to obscure the presence of some targets, but may not generally lead to the inclusion of nonexistent targets.

Conclusions

At present, smithRNA research is in its infancy, and information on smithRNA is arguably limited. SmithRNAs have been described bioinformatically in the Manila clam [19] and two have been functionally validated via RNA injection experiments [31]. Initial bioinformatic observations are available for some additional metazoan species [31, 37]. Although their nature, biosynthesis and mode of action are generally assumed to be similar to those of other, better known, families of small noncoding RNAs such as miRNAs, piRNAs and siRNAs, the actual molecular mechanisms of smithRNA production, maturation and biological action are largely unknown. Similarly, while one or two smithRNAs have been confirmed experimentally, a sufficiently large study set of true positive and true negative smithRNAs is not available to actually assess the performance of the pipeline analytically.

As such, we are not currently in the position to devise an analytical pipeline to actually predict functional (and not 'candidate', as defined here) smithRNAs with minimal standards of efficiency. Nevertheless, considering that multiple research groups are moving in the direction of searching for smithRNAs in different animal species, with the long-term aim of assessing whether smithRNAs could be a more widespread feature of Metazoa, and characterizing smithRNAs structure and mode of action in more detail, a unified, though preliminary, analytical pipeline is a necessary and timely addition to the available toolbox.

In our view, the proposed analytical pipeline will be of substantial interest in two different contexts. On the one hand, it will allow to produce initial data about the presence and characteristics of smithRNAs in different Metazoan species. While such data will be necessarily limited to the description of the microRNA transcriptome of mitochondrial origin and, at the very best, to the identification of putative smithRNA/target interactions, this approach is liable to produce data that are comparable across different species. These data will, in turn, serve as a basis for a comparative overview of smithRNAs across Metazoa. On the other hand, by focusing on one or a few individual species that may be targeted in functional studies, the proposed pipeline will allow the identification of candidate smithRNAs for biological validation (as in [31]).

In the end, we currently see the first SmithHunter module as solid and efficient. We do not foresee modifications in the short term apart from a) a possible length filter on clusters, if future studies suggest that functional smithRNAs display, in line with miRNAs, a tighter length distribution; and b) full incorporation of the end conservation filter. On the other hand, we consider the second module to be more experimental. Notably, the foreseeable availability, in the medium term, of a study set of true positives/negatives following in vivo experiments, will allow us to better gauge run parameters and to analytically evaluate the performance of different options and thresholds.

Availability and requirements

Project name: SmithHunter

Project home page: https://github.com/ESZlab/SmithHunter; https://sites.google.com/unisi.it/mitomicro/smithhunter

Operating system: Linux

Programming language: bash, R

Other requirements: conda, PITA.

License: GPLv3

Any restrictions to use by non-academics: freely available for non-commercial purposes.

Abbreviations

dG	Delta G
DUI	Doubly uniparental inheritance
kJ	Kilo joule
miRNA	Micro RNA
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
numts	Nuclear mitochondrial pseudogenes
PE	Paired-end
piRNA	Piwi-interacting RNA
pre-miRNA	MiRNA precursor
pre-smithRNA	SmithRNA precursor
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
rRNA	Ribosomal RNA
SE	Single-end
siRNA	Short interfering RNA
smithRNA	Small mitochondrial highly-transcribed RNA
sncRNA	Small non coding RNA
SRA	Short read archive
tRF	Transfer RNA fragment
tRNA	Transfer RNA
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05909-0.

Additional file 1. Title of data: SmithHunter distribution folder (compressed). Description of data: SmithHunter distribution as available on GitHub at the time of publication (https://github.com/ESZlab/SmithHunter).

Additional file 2. Trimming and remapping statistics of the small transcriptome.

Additional file 3. Presumptive and candidate smithRNAs identified under different parameter combinations.

Additional file 4. SmithRNAs and their targets identified under the most selective parameter combination (extended information).

Additional file 5. Effect of clustering threshold, stringency and replicate parameters on the number of smithRNAs identified.

Additional file 6. Length of mitochondrial reads from candidate smithRNAs.

Additional file 7. Commands used to perform the analyses described.

Acknowledgements

The authors wish to thank all colleagues from the research groups at the University of Siena and at the University of Bologna for their daily contributions and useful discussions.

Author contributions

MP, FP and FN conceived the study. GM and DC wrote the main scripts. CC, FP and FN contributed code. GM analyzed the data. GM, FN and DC wrote the first manuscript draft. CC, AC, FP, FF and MP contributed to the final manuscript. CC, FN and GM created the GitHub repository and the web-site. MP and FN coordinated the research (scientific). FN coordinated the research (administration). All authors read and approved the final manuscript.

Funding

The study was funded by the Italian Ministry of University and Research under the program PRIN2020 (Progetti di Ricerca di Rilevante Interesse Nazionale) to M.P. and F.N. (project MitoMicro; 2020BE2BC3).

Availability of data and materials

The ***R. philippinarum*** data analysed during the current study are available in: SRA, accession numbers SRR3662624-SRR3662629 (small transcriptome raw data); GenBank, accession numbers AB065375.1 and AB065374.1 (male and female mitochondrial genomes); GenBank, accession number GCA_026571515.2 (nuclear genome); GenBank, accession numbers JO101212-JO124029 (transcriptome); Osfhome, https://doi.org/10.17605/OSF.IO/CDKB9 (transcriptome annotations). The ***P. streckersoni*** data are available in: SRA, accession numbers SRR23195578-SRR23195582, SRR23195559 (small

transcriptome raw data); GenBank, accession numbers ***ON881148*** and ***MW413895*** (male and female mitochondrial genomes); GenBank, accession number ***JAEAOA01*** (nuclear genome); GitHub https://github.com/raqmejtru/mitonuclear-sd (i.e. the supporting materials to [37]; UTR sequences).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interest.

Received: 20 February 2024 Accepted: 21 August 2024
Published online: 02 September 2024

References

1. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10:94–108.
2. Formaggioni A, Cavalli G, Hamada M, Sakamoto T, Plazzi F, Passamonti M. The evolution and characterization of the RNA interference pathways in Lophotrochozoa. Genome Biol Evol. 2024. https://doi.org/10.1093/gbe/evae098.
3. Biswas K, Jolly MK, Ghosh A. First passage time properties of miRNA-mediated protein translation. J Theor Biol. 2021;529:110863.
4. Bartel DP. Metazoan microRNAs. Cell. 2018;173:20–51.
5. Szczepanek J, Pareek CS, Tretyn A. The role of microRNAs in animal physiology and pathology. Transl Res Vet Sci. 2018;1:13–33.
6. Moran Y, Agron M, Praher D, Technau U. The evolutionary origin of plant and animal microRNAs. Nat Ecol Evol. 2017;1:27.
7. Bofill-De Ros X, Yang A, Gu S. IsomiRs: expanding the miRNA repression toolbox beyond the seed. Biochim Biophys Acta Gene Regul Mech. 2020;1863:194373.
8. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 2008;23:P578–87.
9. Kim SS, Lee S-JV. Non-coding RNAs in *Caenorhabditis elegans* aging. Mol Cells. 2019;42:379–85.
10. Riggs CL, Summers A, Warren DE, Nilsson GE, Lefevre S, Dowd WW, Milton S, Podrabsky JE. Small non-coding RNA expression and vertebrate anoxia tolerance. Front Genet. 2018;9:230.
11. Wang M, Jiang S, Wu W, Yu F, Chang W, Li P, Wang K. Non-coding RNAs function as immune regulators in teleost fish. Front Immunol. 2018;9:2801.
12. Larriba E, del Mazo J. Role of non-coding RNAs in the transgenerational epigenetic transmission of the effects of reprotoxicants. Int J Mol Sci. 2016;17:452.
13. Jiao Y, Zheng Z, Du X, Wang Q, Huang R, Deng Y, Shi S, Zhao X. Identification and characterization of microRNAs in Pearl Oyster *Pinctada martensii* by Solexa deep sequencing. Mar Biotechnol. 2014;16:54–62.
14. Li P, Jiao J, Gao G, Prabhakar BS. Control of mitochondrial activity by miRNAs. J Cell Biochem. 2012;113:1104–10.
15. Paramasivam A, Vijayashee PJ. MitomiRs: new emerging microRNAs in mitochondrial dysfunction and cardiovascular disease. Hypertens Res. 2020;43:851–3.
16. Fan S, Tian T, Chen W, Lv X, Lei X, Zhang H, Sun S, Cai L, Pan G, He L, Ou Z, Lin X, Wang X, Perez MF, Tu Z, Ferrone S, Tannous BA, Li J. Mitochondrial miRNA determines chemoresistance by reprogramming metabolism and regulating mitochondrial transcription. Cancer Res. 2019;79:1069–84.
17. Ro S, Ma HY, Park C, Ortogero N, Song R, Hennig GW, Zheng H, Lin YM, Moro L, Hsieh JT, et al. The mitochondrial genome encodes abundant small noncoding RNAs. Cell Res. 2013;23:759–74.
18. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AM, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, et al. The human mitochondrial transcriptome. Cell. 2011;146:645–58.
19. Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M. SmithRNAs: could mitochondria "bend" nuclear regulation? Mol Biol Evol. 2017;34:1960–73.
20. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999;27:1767–80.
21. Formaggioni A, Luchetti A, Plazzi F. Mitochondrial genomic landscape: a portrait of the mitochondrial genome 40 years after the first complete sequence. Life (Basel). 2021;11:663.
22. Ghiselli F, Gomes-Dos-Santos A, Adema CM, Lopes-Lima M, Sharbrough J, Boore JL. Molluscan mitochondrial genomes break the rules. Philos Trans R Soc Lond B Biol Sci. 2021;376:20200159.
23. Passamonti M, Plazzi F. Doubly uniparental inheritance and beyond: the contribution of the Manila clam *Ruditapes philippinarum*. J Zool Syst Evol Res. 2020;58:529–40.
24. Zouros E, Rodakis GC. Doubly uniparental inheritance of mtDNA: an unappreciated defiance of a general rule. Adv Anat Embryol Cell Biol. 2019;231:25–49.
25. D'Souza AR, Minczuk M. Mitochondrial transcription and translation: overview. Essays Biochem. 2018;62:309–20.
26. Plazzi F, Le Cras Y, Formaggioni A, Passamonti M. Mitochondrially mediated RNA interference, a retrograde signaling system affecting nuclear gene expression. Heredity. 2023. https://doi.org/10.1038/s41437-023-00650-5.
27. Passamonti M, Scali V. Gender-associated mitochondrial DNA heteroplasmy in the venerid clam *Tapes philippinarum* (Mollusca Bivalvia). Curr Genet. 2001;39:117–24.
28. Milani L, Ghiselli F, Passamonti M. Mitochondrial selfish elements and the evolution of biological novelties. Curr Zool. 2016;62:687–97.

29. Plazzi F, Puccio G, Passamonti M. Comparative large-scale mitogenomics evidences clade-specific evolutionary trends in mitochondrial DNAs of Bivalvia. Genome Biol Evol. 2016;8:2544–64.
30. Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M. A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. Trends Genet. 2014;30:555–64.
31. Passamonti M, Calderone M, Delpero M, Plazzi F. Clues of in vivo nuclear gene regulation by mitochondrial short non-coding RNAs. Sci Rep. 2020;10:8219.
32. Shaukat A-N, Kaliatsi EG, Stamatopoulou V, Stathopoulos C. Mitochondrial tRNA-derived fragments and their contribution to gene expression regulation. Front Physiol. 2021;12:729452.
33. Mesguer S. MicroRNAs and tRNA-derived small fragments: key messenger in nuclear-mitochondrial communication. Front Mol Biosci. 2021;8:643575.
34. Chen Z, Sun Y, Yang X, Wu Z, Guo K, Niu X, Wang Q, Ruan J, Bu W, Gao S. Two featured series of rRNA-derived RNA fragments (rRFs) constitute a novel class of small RNAs. PLoS ONE. 2017;12:e0176458.
35. Xu X, Ji H, Jin X, Cheng Z, Yao X, Liu Y, Zhao Q, Zhang T, Ruan J, Bu W, Chen Z, Gao S. Using pan RNA-seq analysis to reveal the ubiquitous existence of 5′and 3′end small RNAs. Front Genet. 2019;10:1–11.
36. Jun X, Cheng Z, Wang B, Yau T, Chen Z, Barker SC, Chen D, Bu W, Sun D, Gao S. Precise annotation of human, chimpanzee, rhesus macaque and mouse mitochondrial genomes leads to insight into mitochondrial transcription in mammals. RNA Biol. 2020;17:359–402.
37. Smith CH, Mejia-Trujillo R, Breton S, Pinto BJ, Kirkpatrick M, Havird JC. Mitonuclear sex determination? Empirical evidence from bivalves. Mol Biol Evol. 2023;40:msad240.
38. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
39. Anaconda Software Distribution. Anaconda documentation. Anaconda Inc.; 2020. https://docs.anaconda.com/.
40. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39:1278–84.
41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. https://doi.org/10.14806/ej.17.1.200.
42. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
45. Bensasson D, Zhang D, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 2001;16:314–21.
46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
47. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
48. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–7.
49. Wu S, Manber U. Agrep—a fast approximate pattern-matching tool. In: 1992 Winter USENIX Conference. San Francisco, California. CiteSeer^X 1992. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.5424.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
51. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003;115:787–98.
52. Tan GC, Chan E, Molnar A, Sarkar R, Alexieva D, Isa IM, Robinson S, Zhang S, Ellis P, Langford CF, et al. 5′ isomiR variation is of functional and evolutionary importance. Nucleic Acids Res. 2014;42:9424–35.
53. Shin C, Nam J-W, Farh KK-H, Chiang HR, Shkumatava A, Bartel DP. Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell. 2010;38:789–802.
54. Krüger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 2006;34:W451–4.
55. Lorenz R, Bernhart SH, Siederdissen CHZ, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26.
56. Ghiselli F, Milani L, Chang PL, Hedgecock D, Davis JP, Nuzhdin SV, Passamonti M. De Novo assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. Mol Biol Evol. 2012;29:771–86.
57. Ghiselli F, Iannello M. A transcriptome annotation pipeline for non-model organisms. 2023. Osfhome. https://doi.org/10.17605/OSF.IO/CDKB9.
58. Ahyong S, Boyko CB, Baylli N, Bernot J, Bieler R Brandao SN, et al. Word register of marine species. 2023. https://www.marinespecies.org. Accessed 21 Dec 2023.
59. Pozzi A, Dowling DK. The genomics origins of small mitochondrial RNAs: are they transcribed by the mitochondrial DNA or by mitochondrial pseudogenes within the nucleus (NUMTs)? Genome Biol Evol. 2019;11:1883–96.
60. Ghiselli F, Milani L, Guerra D, Chang PL, Breton S, Nuzhdin SV, Passamonti M. Structure, transcription, and variability of metazoan mitochondrial genome: perspectives from an unusual mitochondrial inheritance system. Genome Biol Evol. 2013;5:1535–54.
61. Pozzi A, Dowling DK. New insights into mitochondrial-nuclear interactions revealed through analysis of small RNAs. Genome Biol Evol. 2022;14:evac023.

Publisher's Note