Methods 126 (2017) 66–75

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Deep sequencing and high-throughput analysis of PIWI-associated small RNAs

Yuka W. Iwasaki [1], Kyoko Ishino [1], Haruhiko Siomi *

Department of Molecular Biology, Keio University School of Medicine, Tokyo 160-8582, Japan

ARTICLE INFO

Article history:
Received 28 March 2017
Received in revised form 8 May 2017
Accepted 21 May 2017
Available online 24 May 2017

Keywords:
RNA silencing
PIWI-piRNA
Small RNA-seq
Directional RNA-seq

ABSTRACT

Small RNAs are now known to be major regulatory factors of gene expression. Emerging methods based on deep-sequencing have enabled the analysis of small RNA expression in a high-throughput manner, leading to the identification of large numbers of small RNAs in various species. Moreover, profiling small RNA data together with transcriptome data enables transcriptional and post-transcriptional regulation mediated by small RNAs to be hypothesized. Here, we isolated PIWIL1 (MIWI)-associated small RNAs from mouse testes, and performed small RNA-seq analysis. In addition, directional RNA-seq was performed using *Piwil1* mutant mouse testes. Using these data, we describe protocols for analyzing small RNA-seq reads to obtain profiles of small RNAs associated with PIWI proteins. We also present bioinformatic protocols for analyzing RNA-seq reads that aim to annotate expression of piRNA clusters and identify genes regulated by piRNAs.

1. Introduction

Regulating differential gene expression is fundamental to the functioning of an organism. RNA interference or RNA silencing is a key gene regulatory pathway that is conserved in a wide range of species [1,2]. In this pathway, small non-coding RNAs of 20–35 nucleotides (nt) associate with AGO- or PIWI-clade proteins of the Argonaute family, forming an effector complex termed the RNA-induced silencing complex (RISC), to regulate expression of their target genes. Small RNAs act to guide the RISC to complementary target sequences. Small RNAs can be divided into three distinct classes: small interfering RNAs (siRNAs), microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs). siRNAs and miRNAs associate with AGO-clade proteins, whereas piRNAs associate with PIWI-clade proteins, to transcriptionally or post-transcriptionally regulate various target genes. piRNAs mediate gene silencing to maintain integrity of the genome in animal gonads [3,4]. In mice, piRNAs are most abundantly expressed in male germ cells, and form the piRNA-induced silencing complex (piRISC) with three PIWI-clade proteins of the Argonaute family: PIWIL1 (MIWI), PIWIL2 (MILI), and PIWIL4 (MIWI2). piRISC is guided by piRNAs, which pair with complementary RNA targets to promote the inactivation of homologous sequences, either by cleaving the target transcript or by inducing specific chromatin modifications, such as DNA methylation and histone modifications. Depletion of individual *Piwi* genes causes male-specific sterility because of severe defects in sperm formation. This suggests non-redundant functions of *Piwi* genes in the testes [5–7]. piRNAs are 24–35 nt long, have characteristic features of uridine (U) at their first nucleotide or adenine (A) at their 10th nucleotide, and 2′-O-methylation at their 3′-end. They mostly originate from transposable elements (TEs) [3,4]. In contrast to the other small RNAs, piRNAs are produced in a Dicer-independent manner from discrete genomic regions called piRNA clusters, which mainly consist of TEs and intergenic unannotated genomic regions. Long, single-stranded precursor transcripts derived from these clusters are processed to produce mature piRNAs, which are then loaded onto the PIWI-clade proteins [8]. Most of these primary piRNAs are thought to function by guiding piRISCs to their target genes, but some are further amplified by a feed-forward amplification loop, the so-called ping-pong cycle, in which sense and antisense TE transcripts are reciprocally cleaved by the endonuclease (slicer) activity of PIWI proteins [9–11]. The ping-pong amplified secondary piRNAs have characteristic features with 1st U and 10th A partners and a 10 nt 5′ overlap (ping-pong signature).

Mouse piRNAs are expressed from piRNA clusters in distinct patterns, in the fetal pre-pachytene and adult pachytene stages of meiosis. Pre-pachytene piRNAs associate with PIWIL4 and PIWIL2 proteins, and they are known to be enriched in TE- and other repeat-derived sequences [9,12], suggesting that they play

* Corresponding author.
E-mail address: awa403@keio.jp (H. Siomi).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2017.05.020
1046-2023/

a role in silencing TEs. In contrast, pachytene piRNAs associate with PIWIL2 and PIWIL1, and include a higher proportion of intergenic, unannotated sequences, with a lower contribution from TE-derived sequences [13–16]. Although loss of genes, such as *Piwil1*, required to generate pachytene piRNAs blocks the production of mature sperm and results in TE deregulation [4,5,17–19], a biological role for pachytene piRNA clusters has yet to be identified. Meanwhile, some piRNA clusters are located within protein-coding genes [20,21]. piRNAs derived from these clusters are referred to as genic piRNAs, although their functions are largely unknown due to the absence of highly complementary genes.

There are thousands of types of piRNA that can target the majority of genes through complementarity. Deep-sequencing followed by computational analyses are, therefore, frequently applied to analyze novel piRNAs and piRNA biogenesis, targets, and the mechanism by which piRNAs regulate their targets. Most of the analyses are performed using a combination of different sequencing techniques, such as small RNA-seq (including small RNA-seq followed by immunoprecipitation of PIWI proteins), RNA-seq, ChIP-seq, global run-on sequencing (GRO-seq) and cap analysis gene expression (CAGE). Because different sequencing techniques are frequently combined, analysis is performed using a combination of software and tools that are available to the community. At the same time, piRNA research in individual laboratories has been complemented by in-house custom scripts. This manuscript precisely describes an example of such a pipeline for piRNA analysis, and can be referred to by those who wish to develop their own custom pipeline. Meanwhile, piPipes has been recently introduced [22]. piPipes is a prepared pipeline that utilizes a set of available software and tools and can be used to perform the analysis in much simpler steps.

The PIWI-piRNA pathway is present in a diverse range of eukaryotes, including humans, and total small RNA-seq is commonly performed in these species to detect novel piRNAs. However, total small RNA of the size range 24–35 nt may contain not only small RNAs but also contaminants of abundant endogenous non-coding RNAs (such as tRNAs or rRNAs), and degradation products of longer RNAs. Also, total small RNA-seq reads provide information of the entire profile of expressed small RNAs, but it is difficult to determine which Argonaute protein a sequence associates with. By definition, piRNAs are "PIWI-interacting RNAs". Therefore, it is critical to combine small RNA-seq data derived from small RNAs co-immunoprecipitated with PIWI proteins together with information from total small RNA-seq. We have previously performed a precise analysis of PIWIL1-associated RNA using testes from the adult common marmoset (*Callithrix jacchus*) [23]. This revealed that PIWIL1 is highly expressed and that the most abundant class of small RNAs is piRNAs. These piRNAs show characteristics of mouse pachytene piRNAs, and are mostly derived from clustered regions in the genome consisting of intergenic regions and a smaller number of TEs. The precise analysis of piRNAs revealed that TE-derived small RNAs are likely to regulate TEs expressed in common marmoset testes. Additionally, some piRNAs were identified to originate from tRNAs or the antisense strand of pseudogenes, suggesting a possible role of marmoset PIWIL1 in the regulation of targets other than TEs.

Here, we describe a protocol for performing small RNA-seq using PIWI-immunopurified samples and for analyzing the sequence data. Using an antibody raised against common marmoset PIWIL1 [23], which cross-reacts with mouse PIWIL1, we extracted small RNAs co-immunoprecipitated with PIWIL1 from mouse testes and performed small RNA-seq. Additionally, we performed directional RNA-seq analysis using *Piwil1* mutant mouse testes, and we describe the steps to analyze the data (Fig. 1). The methodology described here can be used for various animals expressing PIWI proteins. Furthermore, the sequence data provided here can be used as a resource for analyzing mouse PIWIL1-piRNAs, and their effect on the transcriptome.

2. Methods for analyzing PIWI-associated small RNAs

Analysis of PIWI-associated small RNAs consists of two major parts. The first part is the generation and sequencing of a small RNA library. The second part is analyzing the sequence data. PIWI-piRNA complexes are immunoprecipitated using a specific monoclonal antibody. Small RNAs are then isolated from PIWI-piRNA complexes, and libraries are generated using a commercially available library preparation kit. After deep sequencing of the library, the data are passed through a pipeline to characterize the reads. Briefly, adapters are removed from the reads and reads are then size selected. The reads are mapped to the reference genome and annotated to ascertain where in the genome they map. The mapped reads are further analyzed to obtain information, such as nucleotide variation and sequence length variation (Fig. 1A).

2.1. Immunoprecipitation of PIWI proteins

The first step in generating a small RNA library is to obtain a high quality small RNA sample. One powerful way to do this is to immunopurify small RNA-associated complexes in harsh buffer conditions, as using different types of detergent (refer to [24] for further information). Previously, we have generated monoclonal antibodies for PIWI proteins of various species, including *Drosophila*, silkworm, mouse and marmoset [11,23,25–27]. Using these antibodies, we were able to specifically immunopurify PIWI proteins, enabling us to obtain associated small RNAs for the generation of sequencing libraries. The step-by-step protocol for immunopurification has been described previously [24,27]. The critical point of this immunopurification step is to use a combination of antibody and buffer conditions that yields abundant yet specific PIWI proteins for further analysis.

For the immunopurification of PIWIL1 described here, testes tissues from a 10-week-old C57BL/6J mouse were homogenized in binding buffer [30 mM HEPES, pH 7.3, 150 mM potassium acetate, 2 mM magnesium acetate, 5 mM dithiothreitol (DTT), 0.1% Nonidet P-40 (Calbiochem), 2 mg/mL Leupeptin (Sigma), 2 mg/mL Pepstatin (Sigma) and 0.5% Aprotinin (Sigma)]. An antibody against marmoset PIWIL1 [*anti*-PIWIL1/MARWI (1A5)], which could also recognize mouse PIWIL1 [23] (Fig. 2A), was immobilized on Dynabeads protein G (Life Technologies). Reaction mixtures were incubated at 4 °C for 3 h and beads were rinsed twice with binding buffer then five times with wash buffer (binding buffer with 0.5 M NaCl). The immunopurified PIWIL1 protein was analyzed using silver staining to check the efficiency and specificity of immunopurification (Fig. 2B).

2.2. Isolation of small RNAs associated with PIWI proteins

After immunopurification of PIWI proteins, associated piRNAs can be isolated by phenol-chloroform extraction. We have previously described a detailed protocol for extraction of small RNAs associated with PIWI proteins [27]. Briefly, wash buffer from immunopurification is removed from PIWIL1-bound beads, and an equal volume of buffer-saturated phenol is added. It is recommended to use buffer-saturated phenol, because beads tend to accumulate at the interface between organic and aqueous phases when unsaturated phenol is used. Alternatively, RNA extraction reagents, such as Isogen-LS (Nippon gene), can be used for this step. Ethanol precipitation is performed to concentrate RNA. It is likely that the RNA at this step is not visible; therefore, use of a co-precipitant, such as Pellet Paint NF (Millipore), is highly recom-

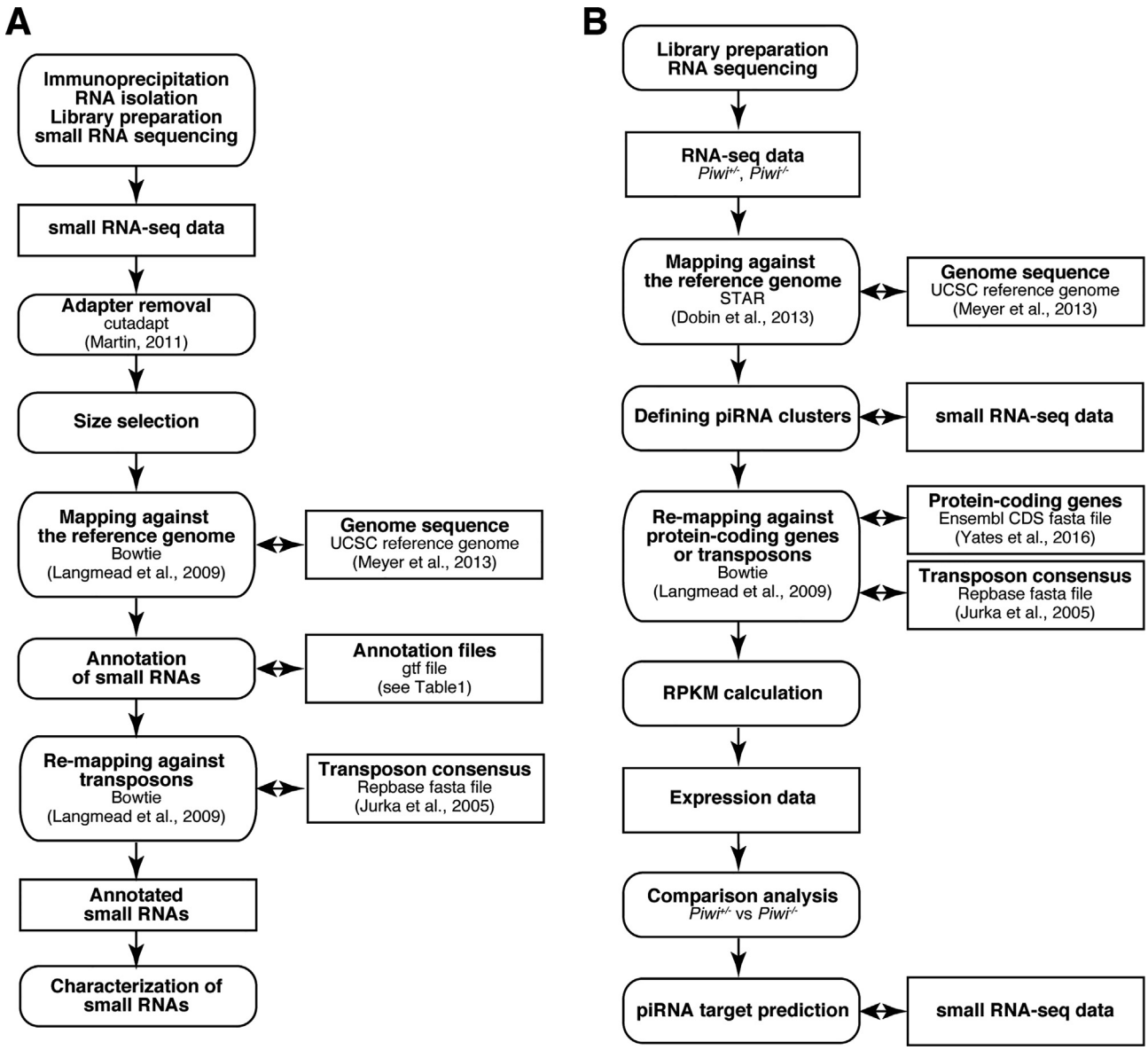

Fig. 1. Overview of PIWI-associated small RNA analysis. Examples of workflows are described for small RNA-seq (A) and RNA-seq (B) analysis.

mended. Also, because small RNAs with low GC contents are sometimes washed out by 70% (v/v) ethanol, it is recommended to rinse the pellet in 80% (v/v) ethanol. Small RNAs can be treated with calf intestinal alkaline phosphatase (CIP) and labeled with [γ-^{32}P]ATP, followed by electrophoresis on a 12% (w/v) acrylamide/6M urea-denaturing gel (Fig. 2C). The co-immunoprecipitated small RNAs can either be used directly for small RNA library preparation, or those in the size range 20–30 nt can be gel-extracted to decrease background signals. The PIWIL1-associated small RNA samples in this study were prepared as described above, and gel extracted before library preparation.

2.3. Preparation and deep sequencing of a small RNA library

Preparation of small RNA libraries can be performed easily using commercially available kits. We normally use NEBNext Small RNA Library Prep Set for Illumina (New England BioLabs), following the manufacturer's protocol. Other kits can also be used. For example, the SMARTer smRNA-Seq Kit for Illumina (Clontech) may be used for small starting amounts of RNA. For library generation with the NEBNext kit, we normally perform the 3′ adapter ligation step at 16 °C for 18 h, because of the low efficiency of adapter ligation to piRNAs that are 2′-O-methylated at the 3′ end. Also, before performing PCR amplification of the libraries, we divide the samples into aliquots and perform PCR on a small scale to determine the minimum number of PCR cycles necessary to obtain a sufficient amount of library. The number of PCR cycles needed can be assessed by analyzing small scale check samples from the obtained library using gel electrophoresis and a bioanalyzer, and checking for the absence of PCR artifacts and a sufficient library concentration. The minimum library concentration depends on which instrument is used for sequencing, how many times the library needs to be sequenced, and whether the samples are sequenced in multiplex. For example, Illumina MiSeq requires 5 μl of the library at a concentration of 4 nM per run. For preparation of the PIWIL1-associated small RNA library described here, 13 PCR cycles were applied, obtaining a library concentration of ~140 nM. Small RNA sequencing often requires fewer reads compared with other sequencing assays. Therefore, libraries can be multiplexed (barcoded) using different index sequences upon library preparation,

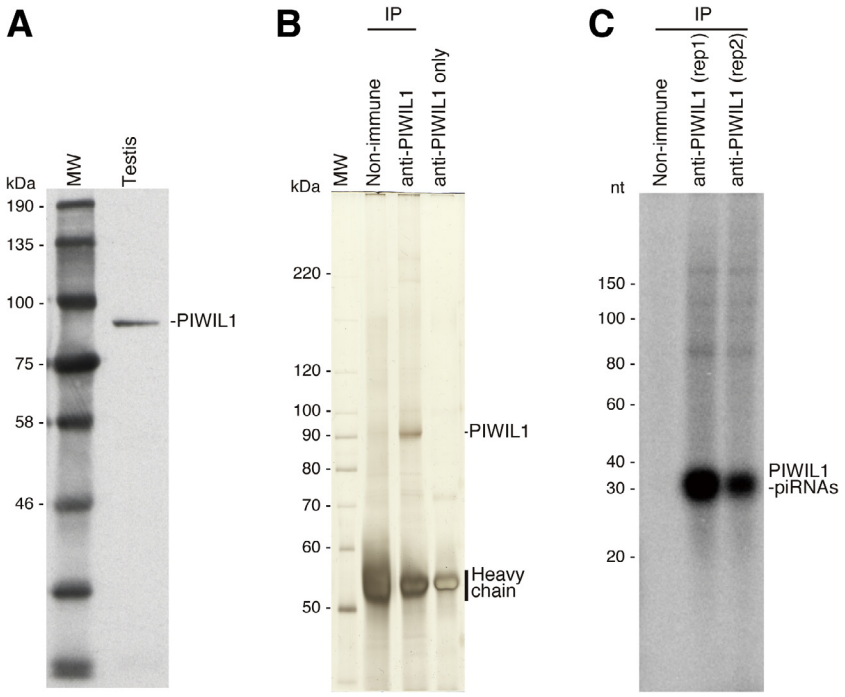

Fig. 2. Immunopurification of mouse PIWIL1-associated small RNAs. (A) Western blotting was performed on mouse testes with an anti-PIWIL1 antibody (1A5). Mouse PIWIL1 protein is detected as a single band in the testes lysate. (B) Silver staining of mouse testes proteins in anti-PIWIL1 (1A5) antibody immunoprecipitant identifies a ~90 kDa band for PIWIL1. PIWIL1 is exclusively immunopurified, without protein contaminants. (C) Isolated RNAs from PIWIL1 immunoprecipitates were ^{32}P-labeled, and separated on a denaturing polyacrylamide gel. Mouse PIWIL1 protein associates with ~30 nt-long piRNAs.

and several samples can be combined and sequenced in one lane. The number of libraries that can be multiplexed depends on how many reads need to be obtained from each library. For example, if using the MiSeq v3 kit, which is capable of sequencing ~25,000,000 reads, two libraries can be combined if ~10,000,000 reads are required from each library.

We frequently use MiSeq (Illumina) for sequencing because a relatively small number of reads is required for small RNA-seq. The amount of required reads varies depending on various factors, such as the analysis performed, the number of replicates performed, and the genome size of the species being analyzed. We sequence at least 10,000,000 reads for the analysis of small RNAs from mammalian species. The reads from replicate samples are combined when reproducibility has been confirmed between samples (see Section 2.5). Owing to the short sequence length of small RNAs, we analyze 50 nt in single end reads, which is sufficient to read through the whole sequence of small RNAs. The PIWIL1-piRNA sequencing described here was performed in two replicates from different samples, yielding a total of 11,245,071 reads (4,836,797 and 6,408,274 reads) using a MiSeq instrument.

2.4. Adapter trimming and size selection of reads

After obtaining the sequenced reads, adapters are trimmed from the reads using tools, such as Cutadapt [28]. Reads obtained from libraries generated using the NEBNext kit do not contain an adapter sequence at their 5′ ends, so only the adapter sequence at their 3′ ends are removed using the −a option. For reads with a 5′ end adapter sequence, the −g option can be used to remove the adapter. It is essential to completely remove adapter sequences for small RNA-seq analysis because most reads contain adapter sequence when sequencing of 50 nt is performed. Also, because of the short length of small RNAs, extra nucleotides originating from adapters would have significant effects on mapping and characterization. Adapter sequence information is often available with the library manufacturer's protocol. Otherwise, sequences can be analyzed by software such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to obtain potential adapter sequences or over-represented sequences, which are likely to be adapters. For the PIWIL1-associated small RNA-seq experiment described here, the adapter sequence "AGATCGGAAGAGCACACGT CTGAACTCCAGTCAC" was used for adapter trimming. The length of reads after adapter removal can be analyzed to check the size distribution of the sequenced reads. The length of small RNA sequences depends on which protein they associate with. For PIWIL1-associated piRNAs, ~90% of reads are in the size range of 24–32 nt with the peak at 30 nt. Therefore, reads of 24–32 nt are selected to avoid contamination from library preparation artifacts. When examining piRNAs of unknown-size that are associated with novel PIWI family members, it is especially important to carefully check the size distribution and decide whether size selection is beneficial or not. This is to avoid eliminating small RNAs that may actually bind PIWI proteins. For Cutadapt, −m and/or −M options can be used to set minimum and the maximum read lengths for downstream analysis.

2.5. Comparison of replicate samples

It is recommended to generate biological replicates for a sample. The number of required replicates is highly dependent on the analysis to be performed, but at least two replicates were obtained in most recent studies. It is also important to validate results using different experimental methods, such as northern blotting or quantitative PCR (see Section 3.3). The replicate samples are compared to check for the reproducibility of the data, and if the samples are highly similar, they are merged for analysis to be performed using the larger dataset. To check for reproducibility, read counts of small RNA sequenced reads are calculated for each of the libraries and the Pearson correlation coefficient is calculated. If the sequence is only present in one of the libraries, the

read count is considered to be zero for the other library. The number of reads obtained from each unique sequenced read can be plotted as a dot plot for direct comparison. For the PIWIL1-piRNA libraries sequenced here, the biological replicate libraries generated using RNA originating from different animals were highly correlated to each other (R^2 = 0.94, Fig. 3A) and, therefore, the reads were merged for the further analysis.

2.6. Analyzing the characteristics of small RNAs

The reads obtained as described above, are from RNA sequences associated with a specific PIWI protein; therefore, it is highly likely that general piRNA characteristics can be determined by analyzing the sequenced reads prior to mapping them on the genome. The length of the small RNA reads can be checked to see if the PIWI protein associates with small RNAs of a specific size. Also, sequence logos, which are graphical representation of the sequence conservation of nucleotides, can be produced by using software, such as weblogo [29] or the motifStack R package (http://www.bioconductor.org/packages/release/bioc/html/motifStack.html). To generate sequence logos, small RNA sequences are aligned at the 5′ end, and searched for nucleotide bias. Because the length of small RNA sequences varies, the sequences must be aligned to either the 5′ or 3′ end, depending on the analysis.

The PIWIL1-piRNAs sequenced in this study were distributed in the range of 26–33 nt and the peak was observed at 30 nt (Fig. 3B). When total small RNA reads were aligned to their 5′ ends and analyzed for nucleotide bias, 88.05% had U at the first nucleotide. Consistently, a sequence logo calculated using motifStack indicated a strong 1st U bias (Fig. 3C).

2.7. Mapping small RNA reads to the genome

After removal of adapter sequences and size selection, reads are mapped to the reference genome using software such as bowtie [30]. An example bowtie command line is: "bowtie [options] <ebwt> {−1 <m1> −2 <m2> | −−12 <r> | <s>} [<hit>]". The index file can be set at <ebwt>, file(s) containing sequenced reads can be set at <m1> and <m2>, <r>, or <s>, and output is written in the file defined by <hit>. Details are described in the bowtie manual. Different options can be set at [options] depending on the analysis you wish to perform. Here, we describe examples of frequently used options. The −v 0 option can be used to extract the small RNA reads that are perfectly mapped to the genome. Also, some mismatches (e.g. 2–3) can be permitted to complement the difference between the reference genome sequence and the genome sequence of the analyzed samples, but since piRNA lengths are as short as ~30 nt, this may increase false positives or the ratio of reads aligned to multiple positions on the genome. For analyzing piRNAs that frequently map to TEs, it may be necessary to consider how to map reads that map to more than one locus of the genome. To preserve the information of reads mapped to several genomic locations, we set a large number (e.g. 1,000,000) for the −k option to obtain all of the mapped positions. This information is important for read annotation, because assignment of annotation features should be prioritized, to avoid any conflict of assignment (see Section 2.9). For further analysis of reads that map to more than one position, the reads mapped to each position are normalized by the number of positions mapped. Similar results may be obtained without setting a −k option because the default parameter selects one random location when reads are mapped to several different locations. When the analysis must be performed with the reads mapped to a single genomic position, the −m option is set to 1. This may lose the majority of the reads mapped to TEs because TEs frequently originate from more than one locus. However, this option is useful for detecting loci that definitely produce piRNAs. The −p option can be used for multiple parallel search threads to speed up the mapping procedure. Reference sequences can be downloaded from databases, such as the UCSC Genome Browser, and indexed for mapping. Otherwise, pre-indexed files are provided at Illumina iGenomes (http://support.illumina.com/sequencing/sequencing_software/igenome.html). For the analysis described here, the PIWIL1-piRNA reads were mapped to the UCSC mouse genome (mm10) using –k 10,000,00 and –v 0 options for annotating the reads, and only the –v 0 option for visualizing the reads (also see Section 2.10.).

2.8. Mapping small RNA reads to TE sequences

The genome mapped reads can be extracted and mapped again (re-mapped) to the consensus sequences of TEs, which can be obtained from databases, such as Repbase [31]. To extract genome mapped reads as fastq files, sam files obtained from genome mapping can be processed using the sort feature of samtools followed by the bam2fastq feature of bedtools. It is important to use genome mapped reads, including those mapped to multiple loci, for this analysis because it is possible to lose many reads that map to TEs by restricting the mapped sequences to unique mappers. Using the obtained fastq file, re-mapping genome mapped reads to TE consensus sequences can be performed using the −m 1 option, which eliminates reads mapped more than once, to avoid false positive mapping of the reads. This option may not be used in case obtaining larger number of piRNA reads is preferred than avoiding false positives. Also, when small RNA reads are mapped to TE consensus sequences, some mismatches may be permitted by setting a higher number (e.g. 2 or 3) in the −v option or by using the default setting because TE consensus sequences do not cover the exact sequence of every TE in the genome.

If a large fraction of reads map to TE consensus sequences, features such as frequency of the reads mapped to each TE, bias of the strand (sense or antisense) that reads are mapped to, and the ping-pong signature (10-base binding frequency) can be calculated. Small RNA reads mapped to each TE consensus sequence are counted and the frequency of the reads mapped to each TE is calculated to estimate which TE is highly targeted by piRNAs, and the ratio of reads mapped to each direction (sense or antisense) is calculated to observe strand-bias of the mapped reads. If piRNAs are mapped to antisense TEs, they are likely to be produced from piRNA clusters and regulate TEs, whereas piRNAs mapped to TE sense strands indicate TEs that may serve as the source of piRNA production. Additionally, frequency of piRNA production and target regulation by the ping-pong pathway can be analyzed. The ping-pong signature is defined as the likelihood that piRNA sequences have a partner with a 10 nt offset [10]. The 52 end nucleotide position for an antisense piRNA partner against a sense piRNA is calculated for piRNA sequences mapped to TE consensus sequences, and is shown as a frequency relative to the other possible positions (e.g. positions 1~25). Frequency of sense-antisense overlap should be higher at 10-base pairs if ping-pong amplification is occurring. The calculations of these characteristics are performed using custom scripts.

2.9. Annotation of reads

Annotation of genome mapped reads is determined by examining the overlap between read-mapped genomic regions and the feature track data from a database, such as the UCSC Genome Browser [32] or Ensembl [33] databases. Reads are assigned to a feature when the length of its overlap is longer than 90% of the small RNA. The priority of the feature assignment is defined to avoid any conflict of assignment. Upon mapping the reads with bowtie [30], a large number for the −k option is used to extract

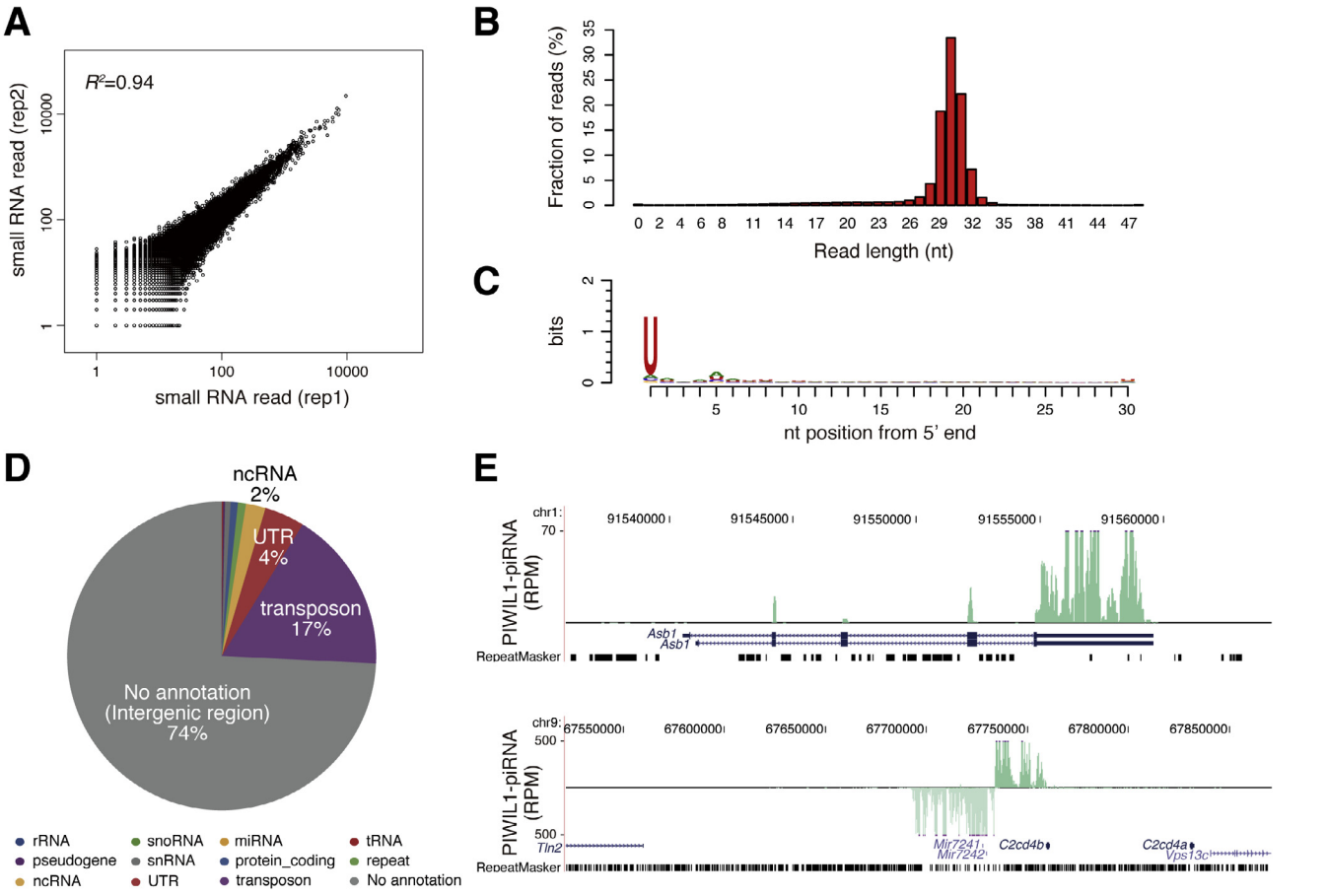

Fig. 3. Analysis of reads obtained from mouse PIWIL1-piRNA sequencing. (A) PIWIL1-piRNA read sequences were plotted as a pairwise comparison between two individual libraries from different animals. R^2 = 0.94 indicates that individual libraries are closely correlated to each other. (B) The size distribution of PIWIL1-piRNA reads, which peak at ~30 nt. (C) Nucleotide bias of PIWIL1-piRNA reads analyzed by motifStack. PIWIL1-piRNAs have a uridine bias at their 5′-end. (D) Annotation of genome-mapped PIWIL1-piRNAs. Most of the PIWIL1-piRNAs were mapped to unannotated genomic regions. (E) Examples of the genomic loci to which the PIWIL1-piRNA reads were mapped. PIWIL1-piRNA reads were mapped to *Asb1* coding gene (upper panel) and also to the unannotated region as a bi-directional piRNA cluster (lower panel).

all of the mapped positions for the reads that are mapped to multiple positions (see Section 2.7). Referring to this information, each read is assigned to one annotation category with higher priority. For the PIWIL1-piRNAs analyzed here, the priority was in the following order: miRNA, ncRNA, rRNA, tRNA, snRNA, snoRNA (to remove small RNA contaminants), followed by UTRs, protein coding genes, transposons, repeats, and pseudogenes (Table 1). The assignments to miRNA, ncRNA, rRNA, snRNA, snoRNA, UTRs, coding genes, and pseudogenes were performed according to Ensembl tracks [33]. UCSC Genome Browser tRNA tracks was used for tRNAs, and UCSC Genome Browser RepeatMasker (repeats) and Simple repeats tracks were used for repeats [32]. Transposons were defined by a combination of Ensembl retrotransposed tracks [33] and the RepeatMasker track (transposon) from the UCSC Genome Browser [32]. This annotation process was performed using custom scripts. Most of the PIWIL1-piRNAs analyzed here were mapped to unannotated intergenic regions, consistent with previous studies [13,15,17] (Fig. 3D).

Table 1
Sources for annotation datasets.

Priority	Annotation	Data source (category name: count of annotations)
1	miRNA	Ensemble gtf (miRNA: 2010)
2	ncRNA	Ensemble gtf (misc_RNA: 17661)
3	rRNA	Ensemble gtf (rRNA: 355)
4	tRNA	UCSC tRNA Genes (435)
5	snRNA	Ensemble gtf (snRNA: 1387)
6	snoRNA	Ensemble gtf (snoRNA: 1512)
7	UTR	Ensemble gtf (protein_coding: 43792)
8	Protein_coding	Ensemble gtf (protein_coding: 43792)
9	Transposon	Ensemble gtf (retrotransposed: 345) UCSC RepeatMasker (transposon: 4192320)
10	Repeat	UCSC RepeatMasker (repeats: 798203) UCSC Simple repeats (765468)
11	Pseudogene	Ensemble gtf (pseudogene: 2000)

2.10. Visualizing the distribution of genome-mapped reads

Distribution of the genome-mapped sequences can be checked using tools, such as Integrative genomics viewer (IGV) [34] or the UCSC Genome Browser [32]. The information of mapped reads is usually in sam or bam format, and can easily be transferred to bam format and indexed using samtools [35]. Using bam files as direct input, reads can be visualized using IGV. Also, reads can be visualized using custom track of the UCSC Genome Browser, which contains a large collection of data that can be overlaid on the small RNA reads. The input files for the UCSC Genome Browser, bedGraph formatted files, can be prepared using HOMER [36]. makeTagDirectory, makeUCSCfile with the –strand separate option, and makeMultiWigHub.pl with the –strand option in HOMER can be used. By using these options, sense and antisense mapped small RNA reads are displayed separately, which is important especially when analyzing small RNAs. Other types of graphics, such as overlaid reads from two different libraries, can also be generated using

these features with different options. Refer to the HOMER manual for further information about option choices. The PIWIL1-piRNA reads analyzed in this study were visualized using the UCSC Genome Browser following the steps described above (Fig. 3E). Examples from custom tracks of the UCSC Genome Browser show genic piRNAs originating from the 3′-UTR of the *Asb1* gene, and also from a bi-directional piRNA cluster located between *Tln2* and *Vps13c* genes. These piRNA clusters are consistent with those described previously by Chirn et al. [21]. The tracks from the UCSC Genome Browser can be exported as pdf or eps files, and used to generate figures.

2.11. Defining piRNA clusters

piRNA clusters were originally identified as genomic regions where a large number of piRNAs mapped [10,14,15]. Using the genome-mapped RNA-seq data, we defined piRNA clusters according to the definition described previously [15], with some minor modifications. A 5 kb window is slid for every 100 bp, searching for regions with the number of piRNA reads exceeding the threshold; bordering windows are merged as a single cluster. To define piRNA cluster threshold, we normalized the previously described piRNA cluster threshold of five piRNAs per 5 kb, according to the total number of genome mapped reads obtained in the study. For example, if the number of genome mapped reads was 10 times greater than that of Girard et al. [15], the threshold for piRNA clusters would be 50 piRNAs per 5 kb. Also, the threshold for the number of unique sequences per 5 kb was also determined, to avoid defining artifact reads as piRNA clusters. These thresholds vary upon analysis, and defined piRNA clusters should also be curated by eye, to confirm the determined threshold. Directional RNA-seq data can also be overlaid on this data, to confirm the transcription of piRNA clusters (see section 3). An example of a PIWIL1-piRNA bidirectional cluster is shown in Fig. 4A along with the directional RNA-seq reads obtained in this study. This confirms that longer RNAs are transcribed as precursor sequences of piRNAs. RNA-seq data indicate the transcription unit of piRNA precursors, but if promoter regions need to be defined precisely, additional data are necessary. CAGE or GRO-seq data can be used to more precisely define the 5′ ends of transcripts, and ChIP-seq data for histone modifications and/or transcription factors are necessary to identify promoters.

3. Methods for utilizing directional RNA-seq data for small RNA analysis

RNA-seq analysis can be performed alongside small RNA-seq analysis, and the results can be overlaid. Transcripts originating from piRNA source loci can be analyzed using RNA-seq data, and can be defined as candidates for piRNA precursor (or piRNA cluster) transcripts. Additionally, by comparing transcriptome data from *Piwi* mutants, the impact of piRNA regulation and the possible piRNA target genes can be estimated. The RNA-seq analysis can be performed using a pipeline commonly used for expression analysis. Here, we analyzed RNA-seq data from *Piwil1*$^{+/-}$ and $^{-/-}$ mice testes by mapping the reads to the genome, overlaying the data with small RNA-seq data, and analyzing the differences in expression profiles by calculating the expression of protein coding genes or TEs (Fig. 1B).

3.1. Preparation and sequencing of a directional RNA-seq library

RNA samples can be purified and libraries can be generated using commercially available library preparation kits. A larger number of reads are required for RNA-seq than for small RNA-seq; therefore, we normally use the Illumina HiSeq system. For the analysis described here, RNA samples were obtained from post-natal day 23 (P23) *Piwil1* mutant mice testes ($^{+/-}$ and $^{-/-}$) [5] (a kind gift from Dr. Kuramochi-Miyagawa), using Isogen (Nippon gene). Instead of polyA selection, a ribo-Zero Magnetic Gold Kit (Human/Mouse/Rat) (Illumina) was used to remove ribosomal RNAs (rRNAs), and the RNA was used for library construction. RNA-seq libraries were prepared according to the directional RNA-seq library prep protocol using a TruSeq Stranded mRNA Sample Prep Kit (Illumina). Libraries generated from *Piwil1*$^{+/-}$ and *Piwil1*$^{-/-}$ mouse testes RNA were paired end sequenced with a 100 nt read length, using the HiSeq2500 system, and 219,524,472 and 211,423,794 reads were obtained, respectively. The optimal protocol for performing RNA-seq analysis can differ. We performed directional RNA-seq to identify strand specific transcription of RNA populations, which is useful for annotating expression of piRNA clusters and non-coding RNAs. Also, we removed rRNAs using the ribo-Zero technique, to avoid limiting the RNA populations to those harboring polyA sequences.

3.2. Mapping RNA-seq data and visualizing reads

RNA-seq reads can be mapped to the genome using commonly used software, such as Bowtie [30], similarly to the small RNA read mapping described earlier (see Section 2.7). However, when analyzing longer paired end reads, an improved mapping rate may be obtained using other software, for example Bowtie2 [37]. Further, newly developed mapping software, Kallisto [38] and STAR [39], can be useful for obtaining accurate mapping of RNA-seq data more rapidly. Adapter trimming prior to genome mapping can be performed for RNA-seq reads, but this is not as critical as for small RNA-seq reads, because most of the reads do not contain an adapter sequence when only 100 nt is sequenced. We did not trim adapter sequences before mapping the RNA-seq reads obtained from *Piwil1* mutant testes, and we used STAR software with default options to map the reads to the UCSC reference genome (mm10).

Reads can be visualized using a viewer, such as IGV [34] or the UCSC Genome Browser [32], as described in the section for small RNA-seq reads (see Section 2.10). In the analysis described here, the UCSC Genome Browser was used to visualize reads. HOMER software was used to prepare input bigwig files for the UCSC Genome Browser. We performed directional RNA-seq, which enables knowledge of the strand the reads originated from to be preserved. This enabled us to use makeTagDirectory with the –sspe option for strand specific paired-end sequencing, followed by makeUCSCfile with the –strand separate option and makeMultiWigHub.pl with the –strand option, to separately display sense and antisense mapped reads. When performing strand specific sequencing in paired-end mode, it is important to check which read (either Read1 or Read2) corresponds to the sense direction of the sequenced transcript (or the 5′ end of the fragments from the transcript). This depends on the method used to prepare the libraries and the information should be available in the manufacturer's library preparation kit protocol. For the TruSeq Stranded mRNA Sample Prep Kit (Illumina) used in this study, Read2 serves as the determinant of the strand (or the read originating from the sequenced transcript). Examples of RNA-seq reads visualized using the UCSC Genome Browser are described in Fig. 4A. The information can be displayed together with small RNA-seq data, to enable comparison of transcript expression patterns and to determine the origin of RNAs.

3.3. Calculation and comparison of gene expression levels using RNA-seq data

Expression of each transcript can be determined using software, such as TopHat [40] and Cufflinks [41]. HTSeq [42] and DESeq [43]

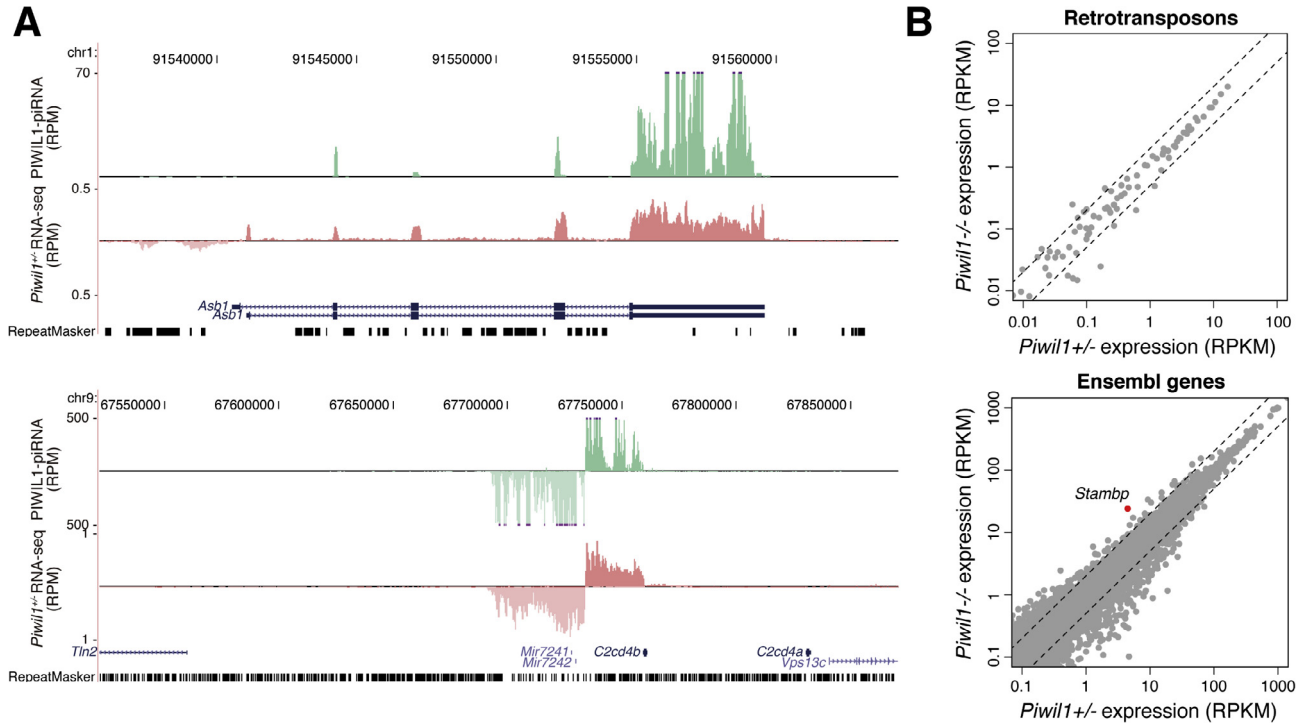

Fig. 4. Analysis of RNA-seq reads from *Piwil1* mutant mice testes. Visualization of RNA-seq reads mapped to the region of piRNA read origin. PIWIL1-piRNA reads are shown with RNA-seq reads from *Piwil1*$^{+/-}$ mouse testis. Transcripts are produced from the genomic regions where piRNAs are produced (described in Fig. 3E). (B) Comparison of RNA-seq data from *Piwil1*$^{+/-}$ and *Piwil1*$^{-/-}$ mouse testes, focusing on TEs (upper panel) and coding genes (lower panel). RPKM values are plotted in log scale, and dashed diagonal lines indicate two-fold change. *Stambp*, a known PIWIL1-piRNA target gene is plotted in red. Note that *Piwil1* is not plotted here, since expression was 0 RPKM in *Piwil1*$^{-/-}$ mouse.

can also be used for differential gene expression analysis, and can determine transcripts with significantly different expression between samples, especially when replicate samples are obtained. Also, genome mapped reads can be re-mapped to a dataset of interest (e.g. protein coding gene sequences or TE sequences), to calculate their expression levels (RPKM; reads per kilobase of transcript per million mapped reads). The optimal number of biological replicates for RNA-seq analyses has been described previously as over six replicates [44]. It is difficult to obtain a large number of replicates per sample; therefore, RNA-seq data registered in the database may be combined (see Section 3.5), or confirmation by a different experimental method (e.g. qPCR or northern blot analysis) may be performed. For the analysis described here, replicates are not performed, and we calculated the expression level by re-mapping the genome mapped reads to the protein coding gene sequence data from the Ensembl [33] database or to the TE consensus sequence data from the Repbase [31] database. As described in re-mapping of small RNA sequences to TEs (see Section 2.8), genome mapped reads were extracted as fastq files, and re-mapping was performed using bowtie [30]. The –m 1 option is used to preserve only uniquely mapped sequences, when mapping reads against TEs. The RPKM value for each gene was calculated by normalizing the number of reads mapped to the sense orientation of each gene (*value A*) by the length of the gene (*value B*) and the number of genome mapped reads (*value C*). Namely, RPKM value = (*value A* × 1000 × 1,000,000)/(*value B* × *value C*). When our RNA-seq data obtained from *Piwil1*$^{+/-}$ and *Piwil1*$^{-/-}$ mice testes are compared, the expression level of most retrotransposons remained unchanged. Meanwhile, when the analysis focused on protein coding genes, we detected de-silencing of the *Stambp* gene, which encodes a deubiquitinating enzyme that is essential in the nervous system [45,46], and has been previously described as a PIWIL1-piRNA target mRNA [47]. These results were consistent with previous reports [13,15,17,47] (Fig. 4B).

The use of software such as DESeq [43] is recommended with high numbers of replicates when the analysis aims to detect novel transcripts, and when accuracy is essential to calculate expression level and differential gene expression. At the same time, the expression level of protein coding genes can be calculated relatively easily using the above-mentioned software, but analyzing TEs would be difficult. The software calculates the expression level from the reads mapped to a specific genome location and, therefore, TEs located at different genomic positions would be treated as different genes. TEs of common origin often share parts of their sequence; therefore, it would be highly challenging to separately analyze TEs located at different genomic positions (many reads would be eliminated by uniquely mapping the reads). One possible solution for analyzing the expression level of TEs is treating the TEs from the same origin as one gene, and re-mapping the genome-mapped reads to TE consensus sequences using the RPKM calculation described above. The optimal method should be chosen depending on the aim of the analysis.

3.4. piRNA target prediction

After retrieving the list of differentially expressed genes in a mutant animal, the next question may be whether piRNAs are able to directly target those genes to regulate their expression. To determine whether piRNAs can target TEs and the other genes whose expression level was significantly affected in a mutant animal, or PIWI protein knockdown samples, searches for sequences complementary to piRNAs may be performed. If piRNAs can be mapped to be complementary to the transcriptional orientation of specific genes, it can be hypothesized that piRNAs are able to regulate those genes, possibly by direct association with the region. The

alignment can be performed using NCBI BLASTN or, when there is a large number of obtained sequences, aligners, such as bowtie [30], may be used. Base pairing at nucleotides 2–21 is required for efficient target cleavage by PIWIL1 proteins and mismatches at several 3′-terminal nucleotides of piRNAs are tolerated [17]; therefore, potential targets may be identified by searching for sequences with complementarity at position 2–21 of piRNAs. This analysis may identify candidate genes regulated by PIWI-piRNAs. However, sequence complementarity against piRNAs is not sufficient to confirm that a gene is a piRNA target, and further experimental data would be necessary to confirm the regulation.

3.5. Using publicly available sequencing data

Published sequence data are deposited in databases to serve as a resource for second-order analysis. The data described here (PIWIL1-associated small RNA-seq and directional RNA-seq data of *Piwil1*$^{+/-}$ and *Piwil1*$^{-/-}$ mice) have been deposited in the NCBI Gene Expression Omnibus (GEO) [48] under accession number GSE97195. The deposited data can be downloaded in SRA format and converted into fastq files by fastq-dump in the SRA Toolkit [49]. The quality of the downloaded data can be checked using software, such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For example, the following criteria can be inspected before using the data: per base sequence quality, sequence length distribution, overrepresented sequences, and adapter content. Refer to the FastQC manual for details.

Utilizing the publicly available dataset is useful to gain insights into small RNA regulation. Here we have used PIWIL1-piRNA and directional RNA-seq information as an example, but using other sequence data would provide further information on piRNAs and piRNA clusters. For example, Li et al. used small RNA-seq and RNA-seq data to identify piRNA clusters. Moreover, they overlaid the piRNA cluster information with CAGE, Poly(A) Site Sequencing (PAS-Seq), and H3K4me3 ChIP-seq data, and precisely identified transcription units and also the promoter regions of the clusters [50]. Also, as discussed earlier, replicates of RNA-seq experiments are important for differential expression analysis (see Section 3.3). Some datasets may be combined to obtain a larger number of replicates. The combination of datasets from different sources must be performed with caution because samples made in different laboratories might not be readily comparable because of specific technical biases that could override any biological signal in the data. Several approaches can be used to reduce batch effects on differential analysis to increase the likelihood of performing a meaningful analysis. For example, DESeq [43], described earlier (see Section 3.3), ignores highly expressed features to reduce the bias induced by skewed read count distribution caused by highly and differentially expressed features. Also, a wide variety of diagnostic plots are available in the NOISeq R package [51] to identify sources of bias in RNA-seq data and to apply appropriate normalization procedures in each case. Various methodologies to reduce biases in RNA-seq data have been reviewed in [52].

Small RNA-seq performed using RNA co-immunopurified with a PIWI protein can identify piRNAs associated with a specific PIWI protein with high confidence. However, a disadvantage of this method is that it is difficult to compare the expression levels of piRNAs between different samples. Only PIWI-associated piRNA reads are analyzed in the library and no internal control exists; therefore, reads are normalized to the total small RNAs associated with the PIWI protein. This method will, therefore, not be able to detect global changes in piRNA expression levels. To overcome this issue, it would be useful to combine total small RNA-seq information from the same sample. The combination of several different sequence data sets, and validation by biochemical analysis, will be important for the elucidation of novel small RNA regulation.

Acknowledgements

We are grateful to members of the Siomi laboratory for helpful discussions and support, especially Takamasa Hirano for experimental support. We also thank Dr. Satomi Kuramochi-Miyagawa for the *Piwil1* mutant mice testes. This work was supported by Grants-in-Aid for Scientific Research, the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) to Y.W.I. (No.15H05583) and H.S (No.25221003).

Appendix A

PIWIL1-associated small RNA-seq and directional RNA-seq data have been deposited in the NCBI GEO under accession number GSE97195. The custom scripts described in this publication can be provided upon request.

References

[1] H. Siomi, M.C. Siomi, On the road to reading the RNA-interference code, Nature 457 (7228) (2009) 396–404.
[2] M. Ghildiyal, P.D. Zamore, Small silencing RNAs: an expanding universe, Nat. Rev. Genet. 10 (2) (2009) 94–108.
[3] Y.W. Iwasaki, M.C. Siomi, H. Siomi, PIWI-interacting RNA: its biogenesis and functions, Annu. Rev. Biochem. 84 (2015) 405–433.
[4] R.S. Pillai, S. Chuma, piRNAs and their involvement in male germline development in mice, Dev. Growth. Differ. 54 (1) (2012) 78–92.
[5] W. Deng, H. Lin, miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis, Dev. Cell 2 (6) (2002) 819–830.
[6] M.A. Carmell, A. Girard, H.J. van de Kant, D. Bourc'his, T.H. Bestor, D.G. de Rooij, G.J. Hannon, MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline, Dev. Cell 12 (4) (2007) 503–514.
[7] S. Kuramochi-Miyagawa, T. Kimura, T.W. Ijiri, T. Isobe, N. Asada, Y. Fujita, M. Ikawa, N. Iwai, M. Okabe, W. Deng, H. Lin, Y. Matsuda, T. Nakano, Mili, a mammalian member of piwi family gene, is essential for spermatogenesis, Development 131 (4) (2004) 839–849.
[8] C.D. Malone, G.J. Hannon, Small RNAs as guardians of the genome, Cell 136 (4) (2009) 656–668.
[9] A.A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K.F. Toth, T. Bestor, G.J. Hannon, A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice, Mol. Cell 31 (6) (2008) 785–799.
[10] J. Brennecke, A.A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, G.J. Hannon, Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*, Cell 128 (6) (2007) 1089–1103.
[11] L.S. Gunawardane, K. Saito, K.M. Nishida, K. Miyoshi, Y. Kawamura, T. Nagami, H. Siomi, M.C. Siomi, A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*, Science 315 (5818) (2007) 1587–1590.
[12] S. De Fazio, N. Bartonicek, M. Di Giacomo, C. Abreu-Goodger, A. Sankar, C. Funaya, C. Antony, P.N. Moreira, A.J. Enright, D. O'Carroll, The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements, Nature 480 (7376) (2011) 259–263.
[13] A.A. Aravin, R. Sachidanandam, A. Girard, K. Fejes-Toth, G.J. Hannon, Developmentally regulated piRNA clusters implicate MILI in transposon control, Science 316 (5825) (2007) 744–747.
[14] A. Aravin, D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M.J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J.J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan, T. Tuschl, A novel class of small RNAs bind to MILI protein in mouse testes, Nature 442 (7099) (2006) 203–207.
[15] A. Girard, R. Sachidanandam, G.J. Hannon, M.A. Carmell, A germline-specific class of small RNAs binds mammalian Piwi proteins, Nature 442 (7099) (2006) 199–202.
[16] E. Beyret, N. Liu, H. Lin, piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism, Cell. Res. 22 (10) (2012) 1429–1439.
[17] M. Reuter, P. Berninger, S. Chuma, H. Shah, M. Hosokawa, C. Funaya, C. Antony, R. Sachidanandam, R.S. Pillai, Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing, Nature 480 (7376) (2011) 264–267.
[18] A.A. Aravin, G.J. Hannon, Small RNA silencing pathways in germ and stem cells, Cold Spring Harb. Symp. Quant. Biol. 73 (2008) 283–290.
[19] A. Vourekas, Q. Zheng, P. Alexiou, M. Maragkakis, Y. Kirino, B.D. Gregory, Z. Mourelatos, Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis, Nat. Struct. Mol. Biol. 19 (8) (2012) 773–781.
[20] N. Robine, N.C. Lau, S. Balla, Z. Jin, K. Okamura, S. Kuramochi-Miyagawa, M.D. Blower, E.C. Lai, A broadly conserved pathway generates 3′UTR-directed primary piRNAs, Curr. Biol. 19 (24) (2009) 2066–2076.
[21] G.W. Chirn, R. Rahman, Y.A. Sytnikova, J.A. Matts, M. Zeng, D. Gerlach, M. Yu, B. Berger, M. Naramura, B.T. Kile, N.C. Lau, Conserved piRNA expression from a

distinct set of piRNA Cluster Loci in Eutherian Mammals, PLoS Genet. 11 (11) (2015) e1005652.
[22] B.W. Han, W. Wang, P.D. Zamore, Z. Weng, piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing, Bioinformatics 31 (4) (2015) 593–595.
[23] T. Hirano, Y.W. Iwasaki, Z.Y. Lin, M. Imamura, N.M. Seki, E. Sasaki, K. Saito, H. Okano, M.C. Siomi, H. Siomi, Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate, RNA 20 (8) (2014) 1223–1237.
[24] K. Miyoshi, T.N. Okada, H. Siomi, M.C. Siomi, Biochemical analyzes of endogenous argonaute complexes immunopurified with anti-Argonaute monoclonal antibodies, Methods Mol. Biol. 725 (2011) 29–43.
[25] K. Saito, K.M. Nishida, T. Mori, Y. Kawamura, K. Miyoshi, T. Nagami, H. Siomi, M.C. Siomi, Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome, Genes Dev. 20 (16) (2006) 2214–2222.
[26] K.M. Nishida, Y.W. Iwasaki, Y. Murota, A. Nagao, T. Mannen, Y. Kato, H. Siomi, M.C. Siomi, Respective functions of two distinct Siwi complexes assembled during PIWI-interacting RNA biogenesis in Bombyx germ cells, Cell Rep. 10 (2) (2015) 193–203.
[27] T. Hirano, H. Hasuwa, H. Siomi, Identification of mouse piRNA pathway components using anti-MIWI2 antibodies, Methods Mol. Biol. 1463 (2017) 205–216.
[28] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.J. 17 (1) (2011) 10–12.
[29] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190.
[30] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (3) (2009) R25.
[31] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res. 110 (1–4) (2005) 462–467.
[32] L.R. Meyer, A.S. Zweig, A.S. Hinrichs, D. Karolchik, R.M. Kuhn, M. Wong, C.A. Sloan, K.R. Rosenbloom, G. Roe, B. Rhead, B.J. Raney, A. Pohl, V.S. Malladi, C.H. Li, B.T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R.A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B.M. Giardine, P.A. Fujita, T.R. Dreszer, M. Diekhans, M.S. Cline, H. Clawson, G.P. Barber, D. Haussler, W.J. Kent, The UCSC Genome Browser database: extensions and updates 2013, Nucleic Acids Res 41 (Database issue) (2013) D64–D69.
[33] A. Yates, W. Akanni, M.R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C.G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S.H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F.J. Martin, T. Maurel, W. McLaren, D.N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H.S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S.P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S.J. Trevanion, F. Cunningham, B.L. Aken, D.R. Zerbino, P. Flicek, Ensembl 2016, Nucl. Acids Res. 44 (D1) (2016) D710–D716.
[34] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (1) (2011) 24–26.
[35] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome, Project data processing the sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[36] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y.C. Lin, P. Laslo, J.X. Cheng, C. Murre, H. Singh, C.K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol. Cell 38 (4) (2010) 576–589.
[37] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods. 9 (4) (2012) 357–359.
[38] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (5) (2016) 525–527.
[39] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2013) 15–21.
[40] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (9) (2009) 1105–1111.
[41] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (5) (2010) 511–515.
[42] S. Anders, P.T. Pyl, W. Huber, HTSeq–a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169.
[43] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550.
[44] N.J. Schurch, P. Schofield, M. Gierlinski, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G.G. Simpson, T. Owen-Hughes, M. Blaxter, G.J. Barton, How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?, RNA 22 (6) (2016) 839–851
[45] N. Ishii, Y. Owada, M. Yamada, S. Miura, K. Murata, H. Asao, H. Kondo, K. Sugamura, Loss of neurons in the hippocampus and cerebral cortex of AMSH-deficient mice, Mol. Cell. Biol. 21 (24) (2001) 8626–8637.
[46] S. Suzuki, K. Tamai, M. Watanabe, M. Kyuuma, M. Ono, K. Sugamura, N. Tanaka, AMSH is required to degrade ubiquitinated proteins in the central nervous system, Biochem. Biophys. Res. Commun. 408 (4) (2011) 582–588.
[47] T. Watanabe, E.C. Cheng, M. Zhong, H. Lin, Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline, Genome Res. 25 (3) (2015) 368–380.
[48] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, A. Soboleva, NCBI GEO: archive for functional genomics data sets – update, Nucleic Acids Res 41 (Database issue) (2013) D991–D995.
[49] Y. Kodama, M. Shumway, R. Leinonen, C. International Nucleotide Sequence Database, The Sequence Read Archive: explosive growth of sequencing data, Nucleic Acids Res. 40 (Database issue) (2012) D54–D56.
[50] X.Z. Li, C.K. Roy, X. Dong, E. Bolcun-Filas, J. Wang, B.W. Han, J. Xu, M.J. Moore, J. C. Schimenti, Z. Weng, P.D. Zamore, An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes, Mol. Cell 50 (1) (2013) 67–81.
[51] S. Tarazona, P. Furio-Tari, D. Turra, A.D. Pietro, M.J. Nueda, A. Ferrer, A. Conesa, Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package, Nucleic Acids Res. 43 (21) (2015) e140.
[52] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szczesniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, Genome Biol. 17 (2016) 13.