Analysis of small RNA-seq for characterising piRNA

What you need:
*Requires installation

- terminal
- Python (might need to install into your computer)
- Perl
- Snakemake*
- Fastqc*
- Cutadapt*
- Bowtie 1*
- Samtools*
- Bedtools*

Procedures:

Part 0. Quality check of sequencing file
Run fastqc program. You can use compressed fastq file.
Usage:
fastqc [sample].fastq.gz

Part 1. Processing fastq file into fasta file before mapping
Usage:
snakemake -s [filename].py

Snakemake filename	Description
01_remove_adapter.py	Remove the sequence from raw reads after NEB adapter sequence (use **bold** part) '**AGATCGGAAGA**GCACACGTCT'. Discard reads that are shorter than 16 nt in length and reads that are not trimmed at all.
02_get_sequence.py	Retrieve only sequence from fastq file output as plain text.
03_count_seq_make_fasta.py	Measure the length of sequences, collect and count duplicated sequences then transform them into fasta format. >[sequence]:[length]:[no. of sequence] sequence
04_count_seq_metrics.py	Not necessary. Check the length distribution of adapter removed sequences. You should see a strong peak at 22 nt which is from miRNAs. There will be also a peak at 25–30 nt.

Part 2. Mapping of sequence to known non-coding RNAs, retrotransposons, and genome

Snakemake filename	Description
05_remove_known_ncRNA.py	Map sequences to mature miRNAs, hairpin miRNAs, snRNAs, snoRNAs, rRNAs, tRNAs, DNA transposons, simple repeats without allowing mismatches. Mapped sequences will be discarded to avoid false positive annotation for retrotransposons (TEs) and coding genes.
06_map_TE_v3_genome_v3.py	Map the unmapped sequences from previous step to retrotransposon sequences from Repeatmasker **allowing up to 3 mismatches.** If there are several matches, annotation with best mapping score will be reported. Those unmapped sequence will be further mapped to genomic sequence **allowing up to 3 mismatches**.
07_annotate_TE_family_v3.py	Acquire full classification of TEs including strand information. Output as [sample]_teRNA_family.txt. This file will be used further analysis in **Part 4**.
08_annotate_genome_v3.py	Classify sequences that were mapped to genome into TEs or genic or others (no annotation) using bedtools.

Part 3. Generate master table using Rstudio and make figures for general information

File / command	Description
09_merging_v3.Rmd	Combine all annotation information from .sam files for each sample and output as [sample]_table.txt.
cat *_table.txt > sRNA_full_table_v3.txt	Combine above txt files to make 1 master data frame. This table is deposited to GEO and can be downloaded.
10_views_v3.Rmd	Input: sRNA_full_table_v3.txt Visualise: • Length distributions (all annotated sRNAs, TE-derived sRNAs, LINE-derived sRNAs, and IAP-derived sRNAs) • Composition of annotated 25–30 nt small RNAs • Pairwise plot of each 25–30 nt small RNAs that have more than 10 count

Part 4. Ping-pong analysis & TE-derived piRNA focused analysis

File	Description
11_views_TE_v3.Rmd	Input: [sample]_teRNA_family.txt Visualise: • Relative amount of Top 10 LINE/IAP piRNA in bar chart Output: • Spefcific piRNA sequence in dab-delimited txt for step 12-14
12_table_to_fasta_v3.py	Make fasta file out of tab delimited txt file.
13_map_consensus_v3.py	Map fasta format sequences to consensus sequence of TEs (up to 3 mismatches) and calculate the distances of 5’ ends of sense/antisense piRNAs.
14_count_nuc_IAP_L1_v3.py	Count first and 10th nucleotide from 5’ end of piRNA.
15_views_pingpong_v3.Rmd	Visualise: • Frequency of overlapped nucleotide length • Frequency of 1U & 10A