sRNA data cleaning script

The script “sRNA_clean.pl” is used to process raw sRNA file in fastq format by

1) trimming the adaptor sequences;
2) removing reads that do not contain adaptor sequences;
3) removing reads that are in low quality (containing “N”);
4) removing reads that are short (e.g., <15 nt) after trimming.

```
perl sRNA_clean.pl -s adaptor_seq –l 15 input1.fastq input2.fastq
```

```
-s sequences of the adaptor
-l minimum length of sRNAs after trimming
```

Multiple sRNA files can be used as input for the script. The script will process the files one by one (sequentially). We have collected six most commonly used sRNA adaptors from the files we have processed. Only the first 11 nt are used for removing adaptors (so providing the 11 nt of the adaptors will be good enough).

```
CTGTAGGCACCATCAAT
CAGATCGGAAGAGCACA
TCGTATGCCGTCTTCTG
TGGAATTCTCGGGTGCC
ATCTCGTATGCCGTCTT
GTACCTCGTATGCCGTC
```

The script will generate the following files:

1. A fastq file containing the cleaned reads for each of the input file.
2. A report file (***report_sRNA_trim.txt***) including the statistics on the sequence processing.

sample	total	unmatch	null	match	baseN	short	clean
test.fq	1000	132	0	868	0	6	862

- **sample**: input file name
- **total**: total number of raw reads
- **unmatch**: number of reads that do not contain the adaptor sequence
- **null**: number of reads containing empty adaptors
- **match**: number of reads containing both sRNA and adaptor sequence
- **baseN**: number of “match” reads containing undetermined base (N)
- **short**: number of “match” reads that are short
- **clean**: number of final cleaned sRNAs

3. A file (***sRNA_length.txt***) containing the sRNA length distribution.