Command used for data processing and novel miRNA discovery during soybean floral transition

Senhao Zhang[1], **Mohan B. Singh**[1], **and Prem L. Bhalla**[1]

[1]Plant Molecular Biology and Biotechnology Laboratory, School of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC 3010, Australia

Please note: $ means that command was executed in a Terminal.

1. Quality check and reads trimming

Quality check using FastQC (v0.11.8)[1]

```
$fastqc Raw.fq
```

Remove low-quality reads and retain reads between 18 to 26 nt using cutadapt (v1.18)[2]

```
$cutadapt -m 18 -M 26 -q 30 -o Raw.trim.fq Raw.fq
```

Quality check after reads trimming using FastQC (v0.11.8)

```
$fastqc Raw.trim.fq
```

The sequencing reads were aligned to soybean tRNA[3], rRNA, snRNA and snoRNA[4] to filter reads using bowtie (v1.2.2)[5] without mismatch.

```
$bowtie-build RMtrsno.fa RMtrsno
$bowtie -t RMtrsno -S -v 0 -a --un Clean.fq -q Raw.trim.fq > Raw.trim.sam 2>
    Raw.trim.log
```

RMtrsno is a bowtie index was built using bowtie-build. RMtrsno.fa (2,803 sequences) was combined from four files (soybean tRNAs, rRNAs, snRNAs and snoRNAs) which were downloaded on 10 Apr 2018.

Map to soybean genome (Gmax_275_v2.0.fa from Phytozome v12)

```
$bowtie-build Gmax_275_v2.0.fa Gmax275
$bowtie -t Gmax275 -S -v 1 -a -m 20 --best --strata --un Clean.unaligned.fq --
    al Clean.aligned.fq -p 16 -q Clean.fq > Clean.sam 2> Clean.log
```

2. Known miRNA identification

Only reads mapped to soybean genome in last step (e.g. Clean.aligned.fq) was used for identifying known soybean miRNAs deposited in miRBase (v22)[6]

```
$bowtie-build --threads 16 mature_gma22.fa gmamiRB22
```

(mature_gma22.fa, mature soybean miRNAs deposited in miRBase v22)

```
$bowtie -t gmamiRB22 -S -v 0 -a --al $o2 --un $o3 -p 16 -q $n > $o4 2> $o5
$fastq_to_fasta -v -i Clean.fq -o Clean.redun.fasta
```

(fastq_to_fasta, part of FASTX Toolkit v0.0.14[7])

3. Length statistics

Redundant read length

```
$bioawk -c fastx '{print $name, length($seq)}' Clean.redun.fasta | cut -f 2 |
    sort -n |uniq -c | awk '{ print $2 "\t" $1}' | sed '1i Length\tCount' >
    Clean.redun.fasta.length
```

Unique read length

```
$fastx_collapser -v -i Clean.redun.fasta -o Clean.uniq.fasta
```

(fastx_collapser, part of FASTX Toolkit v0.0.14)

```
$bioawk -c fastx '{print $name, length($seq)}' Clean.uniq.fasta | cut -f 2 |
    sort -n |uniq -c | awk '{ print $2 "\t" $1}' | sed '1i Length\tCount' >
    Clean.uniq.fasta.length
```

4. Novel miRNA prediction

By using miRDeep-P2 (v1.1.1)[8] with a published pipeline[9]

Combine all unaligned reads for novel miRNA identification

```
$cat LSD0_1.unaligned.fq LSD0_2.unaligned.fq ... SLD16_3.unaligned.fq >
    combined.fq
```

Convert fastq to redundant fasta file

```
$cat combined.fq | paste - - - - | sed 's/^@/>/g' | cut -f1-2 | tr '\t' '\n' >
    combined.fasta
```

Collapse redundant fasta to unique fasta file

```
$fastx_collapser -v -i combined.fasta -o unique_tags.fasta
```

Change header of the fasta file to meet miRDeep-P2's requirement

```
$sed 's/-/_x/' unique_tags.fasta > modified_tags.fasta
```

Align combined reads to soybean genome

```
$bowtie -a -v 0 -p 32 -f Gmax275 modified_tags.fasta aligned_reads
```

Convert alignment file to blast format

```
$convert_bowtie_to_blast.pl aligned_reads modified_tags.fasta Gmax_275_v2.0.fa
    > aligned_reads.bst
```

Filter reads

```
$filter_alignments.pl aligned_reads.bst -c 15 > filtered_reads_c15.bst
$overlap.pl filtered_reads_c15.bst CDS_annotation.gff3 -b > overlap_ids_c15
$alignedselected.pl filtered_reads_c15.bst -g overlap_ids_c15 >
    annotation_filtered_c15.bst
$filter_alignments.pl annotation_filtered_c15.bst -b modified_tags.fasta >
    filtered_tags_c15.fasta
```

Sort retained alignments

```
$sort +3 -25 < annotation_filtered_c15.bst > annotation_filtered_sorted_c15.bst
```

Extract potential miRNA precursor sequences

```
$excise_candidate.pl Gmax_275_v2.0.fa annotation_filtered_sorted_c15.bst 250 >
    precursors_c15.fasta
```

Using RNAfold (v2.4.9)[10] to predict secondary structure

```
$RNAfold -v --jobs=0 --noPS --infile=precursors_c15.fasta --outfile=precursors.
    structures
```

Build bowtie index for precursor sequences

```
$bowtie-build --threads 12 -f precursors_c15.fasta precursors_index
```

Align sequences to precursors

```
$bowtie -a -v 0 -p 12 -f precursors_index filtered_tags_c15.fasta >
    tags_aligned_to_precursors.aln
```

Convert to blast format

```
$convert_bowtie_to_blast.pl tags_aligned_to_precursors.aln filtered_tags_c15.
    fasta precursors_c15.fasta > tags_aligned_to_precursors.bst
```

Sort signature file

```
$cat tags_aligned_to_precursors.bst | sort +3 -25 > All.signatures
```

Using miRDeep-P2 for novel miRNA discovery (miRdeep-P2 incorporates the latest plant miRNA annotation criteria[11])

```
$mod-miRDP.pl All.signatures precursors.structures > miRNA_predictions
```

Filter miRNAs by incorporating latest annotation criteria for plant miRNAs

```
$mod-rm_redundant_meet_plant.pl Chr_length precursors_c15.fasta
    miRNA_predictions total_reads miR_nr_predictions miR_filter_P_predictions
```

5. Known and novel miRNA read counts retrieval

Use featureCounts[12] for reads quantification

```
$featureCounts -t miRNA -g Name -O -s 1 -M -a known_novel.gff3 -o LSD1_1.counts
     LSD1_1.sam
```

Acknowledgements

SZ thanks Melbourne Bioinformatics (The University of Melbourne) for providing the High-Performance Computing (HPC) facility for small RNA sequencing data analysis (project number punim0550).

References

1. Fastqc - a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
2. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12, DOI: 10.14806/ej.17.1.200 (2011).
3. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189, DOI: 10.1093/nar/gkv1309 (2015).
4. The RNAcentral Consortium. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* **45**, D128–D134, DOI: 10.1093/nar/gkw1008 (2016).
5. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.* **10**, DOI: 10.1186/gb-2009-10-3-r25 (2009).
6. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. mirbase: from microrna sequences to function. *Nucleic Acids Res.* **47**, DOI: 10.1093/nar/gky1141 (2018).
7. Fastx-toolkit - fastq/a short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit/.
8. Kuang, Z., Wang, Y., Li, L. & Yang, X. mirdeep-p2: accurate and fast analysis of the microrna transcriptome in plants. *Bioinformatics* **35**, 2521–2522, DOI: 10.1093/bioinformatics/bty972 (2018).
9. Ilnytskyy, S. & Bilichak, A. *Bioinformatics Analysis of Small RNA Transcriptomes: The Detailed Workflow*, 197–224 (Springer US, Boston, MA, 2017).
10. Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for Mol. Biol.* **6**, DOI: 10.1186/1748-7188-6-26 (2011).
11. Axtell, M. J. & Meyers, B. C. Revisiting criteria for plant microrna annotation in the era of big data. **30**, 272–284, DOI: 10.1105/tpc.17.00851 (2018).
12. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930, DOI: 10.1093/bioinformatics/btt656 (2013).