1 miRcat2 clustering algorithm

This algorithm is part of the step for filtering of the sequences, that checks whether the candidates have a miRNA-like alignment of incident reads. It is used to determine if the selected sequences present evidence of precise processing of the pre-miRNA by Drosha (animals) and Dicer (plants and animals) (Kim, 2005; Bartel, 2004; Chen, 2005) i.e. the presence of one or two peaks corresponding to the miRNA/miRNA* location. This filtering step ensures that the majority of reads aligned to the miRNA/miRNA* location have a high overlap (are variants of each other), and have the same genomic orientation (Kim, 2005; Bartel, 2004; Chen, 2005).

We define a cluster as all sequences that map to the same genomic location, having the start and the end of the mapping position within clear_cutnts of each other. We chose clear_cut to be 3 nts, because during the miRNAs biogenesis, many isomiRs are also generated within these bounds (sequences that have variations of a few nts with respect to the reference miRNA sequence) (Morin et al., 2008).

The rules for an sRNA sequence s (with sRNA beginning position b_s and ending position e_s) to be considered to belong to a cluster C (with cluster beginning position b_c and ending position e_c) are:

- if b_s ≥ b_c and e_s ≤ e_c, then the sRNA location is completely inside the cluster boundaries, therefore s ∈ C;
- if b_s ≥ b_c and e_s ≤ e_c, then the sRNA location is completely covering the cluster location, therefore s ∈ C;
- if (b_s − b_c ≤ clear_cut and b_c − b_s ≥ 0) or (e_s − e_c ≤ clear_cut and e_c − e_s ≥ 0), then the sRNA location is with clear_cutnts to the right or left from the cluster location, therefore s ∈ C;
- we define mid_s to be mid_s = \frac{(e_s + b_s)}}{2}, and mid_c to be mid_c = \frac{(e_c + b_c)}}{2}; if (mid_c − mid_s) ≤ clear_cut, then the middle of the sRNA location is with clear_cutnts to the right or left from the middle of the cluster location, therefore s ∈ C;
- otherwise s ∉ C.

2 Fold change computation

To validate miRNA predictions, we estimate fold changes between wild type and mutants in the miRNA biogenesis pathway. To do this, we consider only the genome mapping reads. To compare datasets with different sequencing depths, we normalize all abundances using the reads per million (RPM) method (Mortazavi et al., 2008) to the median total count (MTC) of each experiment (McCormick et al., 2011; Dillies et al., 2013). Briefly, we sum the abundances of genome mapping reads in each sample to obtain the total for each library; next we calculate the median of total counts for each comparison (the MTC value). We normalize the abundance of each read using: normalized_count = \frac{count}{median\ total\ count + MTC} (McCormick et al., 2011).

To create a control dataset, as suggested in (Friedlander et al., 2014), containing reads whose abundances are unlikely to be affected by the mutations on the miRNA pathway, the reads in each experiment were mapped to a file containing tRNAs and snoRNAs of the respective species, using PatMan, full length, with 0 gaps and 0 mismatches. The control file was created downloading tRNAs and snoRNAs from the rFAM database (Navrotsky et al., 2014) through the RNAcentral web service (RNAcentral Consortium, 2014) (http://maccentral.org/). For each RFAM transcript, we compute its abundance as the algebraic sum of the normalized abundances of mapped reads, for each condition.

We then calculate the log2 fold change using the normalized abundances of all predicted miRNAs, from each tool. The log2 fold change for each miRNA is calculated for each set of replicates as the ratio between the mean value of normalized abundances from the mutant samples to the median value of normalized abundances from the wildtype (control) samples. We use an offset approach, adding a count of 10 to both numerator and denominator, to avoid divisions by zero and cases where lowly expressed sequences appear to be differentially expressed (Mohorina et al., 2011). Next, we compare the percentage of reads that are significantly down regulated in the mutant samples (≥ 2-fold downregulated).

3 Data processing

All samples downloaded from GEO (Barrett et al., 2013) or SRA (Leinonen et al., 2010) databases were processed as follows: files were transformed to fasta format (using the star tool kit). In 2.2.4 (Leinonen et al., 2010) 3’ adapters were trimmed and sequences longer than 16nt were kept for the subsequent steps. All A. thaliana samples that we sequenced were transformed from fastq to fasta format, then the 3’ adapters and the HD tags were trimmed (using the UEA small RNA Workbench); sequences longer than 16 nt were kept for the subsequent steps. Next, all files were collapsed into non-redundant format (for each sRNA we kept the sequence and its abundance). Then the files were aligned full length, with 0 gaps and 0 mismatches to the respective reference genome using PatMan (Prüfer et al., 2008). For miRDeep2, the reads were mapped using mapper.pl, to make the results comparable with the other methods used, the sequences which mapped with mismatches were discarded. All software was run on the processed datasets with their default parameters.

4 A. thaliana wildtype and DCL1 mutant small RNA sequencing

4.1 Plant materials

Seeds of Arabidopsis thaliana cv. Columbia dcl1-7 (sm1-1) mutants from NASC The European Arabidopsis Stock Centre (stock ID N6953) were stratified at 4°C for 48h and germinated under long-day conditions with 16h daytime at 22°C and 8h night time at 18°C. The plants were grown under the same conditions. Leaves were harvested and genomic DNA were extracted as previously described (Amani et al., 2011) for genotyping. Heterozygous individuals were identified through PCR-sequencing for seed increase.

The seeds from wild type and heterozygous individuals were germinated and the plants were grown under the same conditions as above. The homozgyous plants were identified by the phenotypes (Lang et al., 1994). When the flower buds in both wild type and mutant plants were visible but without opening flowers, the leaves of nine plants were harvested and stored at -80°C for three biological replicates. One replicate consists of a pool of leaves from three individual plants.

4.2 Genotyping analysis

Dcl-1 specific primers: 5’-TGTCACCAACCGTCTAGAAAT-3’ (forward) and 5’-GCAGCTTTGTCATACTCGACG-3’ (reverse) were used for PCR. The final concentrations for the reaction components are: 1X GoTag® Reaction Buffer, 0.2mM dNTP each, 0.5mM primer each, 1.25U GoTag® DNA Polymerase and 5ng/ul genomic DNA. PCR was performed with the following programme: 2min at 95°C, 35 cycles of 30 sec at 95°C, 30 sec at 60°C and 30sec at 72°C, and a final extension for 5min at 72°C. The PCR products were purified using the Zymoclean™ Gel DNA Recovery Kit following the manufacturer’s protocol, and sequenced. Heterozygous samples contain both wild type and mutant sequences.
4.3 RNA extraction and small RNA library construction

Total RNA of both wild type and mutant leaf tissue was extracted using TRI-Reagent (Ambion) following manufacturer’s protocol. Small RNA (sRNA) of each sample was further enriched using the mirVana miRNA isolation kitTM (Ambion) according to the manufacturer’s instructions. The sRNA libraries were constructed using 1 µg of enriched sRNA fraction based on the previously published protocol by Xu et al., 2015 (Xu et al., 2015). The libraries were barcoded, pooled and sequenced on a single HiSeq2500 lane. The raw fastq files and processed csv files are publicly available on Gene Expression Omnibus (GEO) (Barrett et al., 2013) under accession number GSE90771 (GSM2412286 to GSM2412288 are the wild type samples and GSM2412289 to GSM2412291 are the DCL1 mutant samples).

References


