smRNA, degradome, 5'-RATE analysis from a biological point of view.

Quality control

Number of reads should be counted after removal of artificial clone reads:

Longer stretches of insert nucleotides, identical to the known sequence of the 3' linker or 5' linker. Sequencing services do not remove cloning artifacts and it should be done in house: they constitute from a few to more than 10% reads in one library. It is necessary to perform a scan of insert sequences at least using all possible 8nt sequence stretches which can derived from both cloning primers (ALL primer sequences must be provided by a person who did the cloning).

Cloning primers used till now:

Structure of smRNA libraries:

AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATC**N**$_{(15\text{-}28)}$TCGTATGCCGTCTTCTGCTTG

N$_{(15\text{-}28)}$ represents inserts, then all possible 8-mers like: CGACGATC, CCGACGAT,….should be used to clean such a smRNA sequencing data from cloning artifacts.

Degradome and 5'-RATE structure is:

AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGAC**N**$_{(20)}$TCGTATGCCGTCTTCTGCTTG

A similar approach has to be used for a preliminary data cleaning.

Check whether such sequences can be mapped to the investigated genome. If this is the case a downstream data user should know that abundance of such tags can be biased.

Optional: blast sequences of rRNA from other organisms which differ from rRNA from A. thaliana – test for sample contamination or unwanted parasites, symbionts etc.

Quality descriptors provided by company (symbols or number showing the strength of a peak) like ATTCGTT/1,10,10,10,2,10,3 in my opinion should be at least placed in the first general table containing: sequence/number of reads/ mean values of quality descriptors.

Information behind is:

a) If "weaker base peaks" occur in a similar sequence context – either base modification or a structure makes sequencing or PCR difficult. RT mistakes are stably copied by PCR into dsDNA.

b) Other problems with sequencing, like chemistry, data acquisition or sample quality will probably generate more random distribution of "weaker" peaks. On the other hand all methods to get DNA sequence information based on copying of a DNA strand to a complementary one give weaker peaks of nucleotides depending on sequence (structure) context. Thus, weak peaks can be neglected if the 3'- linker sequence is of a good quality (actually this is tested by sequencing service providers).

Actually it seems difficult to interpret the data quality, but it seems reasonable to keep these data from MPSS to be able to go back to them if some further predictions or analyses will give rather unexpected results. However this is rather a valuable information for sequencing facilities.

Normalization of sequence sets to compare abundance of small RNAs in different samples

TPM – for mapped sequences if the sequence of a reference genome is known. Number of reads mapped to rRNA 100% + 90% should be good enough for other samples (Brassica). The normalization should be checked: at least several northern blots should be performed, using probes for miRs of stable and changed expression / high, and low abundance.

Consider removal of sequences of extreme lengths: usually for smRNA a region from 19-29 nt based on DNA marker is excised from a gel, from the same amount (10 µg) of total RNA (rRNA) for all samples. 17-mer of phosphorylated RNA migrates as fast as 20-mer DNA (RNA 16-28). My suggestion is to remove from normalization all inserts of size smaller than 18 and longer than 26nt. RNA yield from samples to be compared should be maximal and similar (to be shown by a biologist).

Extremely short inserts: can change the number of reads, due to imprecise electrophoresis or excision of a band, too long also, and …the sequence of 3'-linker should be recognized: minimum of 6 nt of a good quality sequence.

Limits: A precise comparison of abundance of two different RNA sequences, for example miR399 and 156 is rather not possible.

a) Different sequence = different structure= different ligation efficiency.

b) Not all 5'ends of smRNA fraction are phosphorylated . Ligation of 5'-linker is dependent on 5'P on small RNA , thus in libraries we have only active miRNAs which can be bound to AGO. Then the result of a northern can be a bit different

c) MiRs are methylated at 3'end- this modification lowers the ligation efficiency.

d) Some small RNAs have tendency to circularize during ligation– then these molecules are eliminated from a library produced by linker ligation. Polyadenylation is more robust, but the information about the small RNA 3'-end sequence is lost.

GENERAL DATA ANALYSIS

Number of reads in a sample, number of artificial clones

Size distribution, overall nucleotide distribution, and in all size categories.

Fraction of single reads, double, triple 4 and 5…?

Overlap between libraries to be compared.

Number of reads which can be mapped to:

a) to nuclear genome(100% and 95%*)

b) to plastid DNA (100% and 95%*)

c) to mitochondrial DNA(100% and 95%*)

d) mapped to annotated transcripts: TAIR9 (Arabidopsis) or to EST data (Brassica, …)

e) use a set of sequences not mapped to all above sequence categories for further mapping to:

f) tRNA sequences without introns and with CCA at the 3'-end (100% and 95%*) I can send you a file for A. thaliana if needed.

g) forward and reverse strand mapping to all gene categories (100 and 95%*)

h) for the remaining group allow mismatches only at 3' or 5'-end, map to all above.

(*smRNA sequences mapped with 100% identity are never mapped again)

From mapping with mismatches we can extract an additional information: U residues at 3'-ends are added mainly or exclusively to miRs which interact with argonautes, then it is possible to predict which molecules from a library can also be repaired like miRs and function like miRs.

Sequences which cannot be mapped neither to genome nor to transcriptome. In small RNA libraries usually 70-80% can be mapped with 100% identity– if less it can be due to bad sequencing quality – then sequencing should be repeated. From our side: library must be a highly purified dsDNA. The size of fragments must fit the sequencing capability: i.e. If one orders 36nt sequencing, the library cannot contain inserts longer than 28-30 nt. If it contains longer DNA: one can get information from a company that let say 20% of reads were rejected because the 3'linker sequence was not detected. 5% -10% seems to be acceptable in the case of small RNA libraries. In degradome libraries is close to 0% because all inserts are of an uniform size (20 or 21nt). Of course longer sequencing can be ordered, but it is a bit more expensive.

Not mapped sequences can appear also due to contamination: they fit to bacteria, other organism, plant viroids (de novo assembly?), fungi, come from other plant cultivar than expected? Come from another organism? Possible. One known example: plant library contaminated with C. elegans sequences in a company during library preparation. Option: mapping to a few ribosomal sequences, characteristic for other groups of species, other species. Small RNA sequences allow for a de novo assembly of viroids..

If not:

Base modification or editing or 3'end addition of nucleotides: C, CC, CCA –tRNAs, U, UU, UUU… in the case of miRNA – uridylation as a mark for degradation, AAAA, other homopolymeric or even a bit random polymers coming from plastids and mitochondria.

If not:

Bad sequencinq quality (quality descriptors from a company can be then helpful) or other, "unknown" contamination.

If one can observe base modifications which look like SNPs – one mismatch allowed in any position:

1. Map this unmapped set to mature miRs and miR* sequences – change of target set possible,

2. Map to abundant tRNA derived small RNAs, rRNA derived small RNAs changes in modifications can be important for regulation of protein biosynthesis in a changing cell environment, mapping should be done to mature and precursor sequences.

3. Remaining: map to mRNAs – possible editing events, esp. for organellar transcripts, but not only – then detection of physiologically determined changes in protein coding regions can be informative, also in UTRs…….silencing complexes…

In some cases single reads are also good reads. To calculate how deep is sequencing of a sample one can check: can one observe a reduction of single read fraction if 2 or 3 libraries are added? How many reads are necessary to have a full representation of small RNAs from a sample, a cell? Another question appears: does it make sense to work with single reads? They usually constitute a big fraction of all reads. They seem rather useless for comparisons and predictions and make data processing time consuming. 2 and 3 reads can be informative only if they represent known, previously characterized molecules. They should not be rejected from further analysis if present in higher abundance in other libraries which have to be compared.

Always leave unmapped as a separate group which can be tested later

The first general table useful for a biologist is (sequence/read number libr. 1/read number libr.2/ ……absolute, normalized)You have prepared for Brassica napus smRNAs: (smallRNAReadDatabase). The size of database can be selected by the downstream user: for example from 5 reads in at least 1 library or from 4 reads in at least 3 libraries etc. It would be useful to add annotations to smRNAs at this stage, like column (AT1G11110), next: mRNA, rRNA18S, tRNA Ala, miR166 or 166a etc/ next column : sense or antisense chr + or - strand.

In the case of unknown genome: Blast smRNA sequences to miRBase annotated mature and star strand miR sequences and add a column “100%mature” with the names of known miRs. If a sequence is identical to more than one miR from miRBase- then a cell should contain more general description than ath-miR399d or miR399 or miR. Next columns: “95%mature”, “90% mature” – these are important to: a) find possible base modifications, b) in the case of a plant which genome sequencing is not yet finished it will give a possibility to find quickly all possible conserved miRs even if their transcripts have not yet been sequenced (example miR 827 in Brassica napus and many others which differ in mature miR sequences between closely related plants by a mismatch, two mismatches, an insertion or deletion), c) changed 3’-end – allows for CCA or other nucleotides, like Us added to miRs in plants – decay/repair information.

Comparing libraries by analysis of gene categories of small RNA sequences:

General:

Mapping of library specific **hotspots** of smRNA in a genome and transcriptome. Mapping to genome: new transcripts, maybe participating in siRNA pathway (21-25nt) can be found in “noncoding regions”.

Search for all possible targets for library specific (comparison to 2 other libraries) smRNAs with PITA or another algorithm. Check of specifity of smRNA/mRNA interaction: microinspector or a similar tool to find if this one smRNA has the strongest interaction with this specific mRNA. These "best pairs" can be useful for a further degradome analysis.

Check the available transcriptomic (array) data whether the interaction can be a general response observed in all tissues of the plant. If not – a local action can be expected. Important to data interpretation and to plan further experiments.

Mapping to mRNAs:

Do plants produce tiRNAS? If yes, this can be maybe helpful in TSS analysis.

Sense and antisense mapping: how many siRNAs, tasiRNAs are induced, How many siRNAs are generated in a sample (tissue, cell or condition dependent manner)?

Mapping to promoters:

Is it (library) tissue or condition specific?

Mapping to rRNA:

We have a new rRNA unit transcript to annotate (CHO dependent).

Maps to rRNA are so dense that de-novo assembly of rRNA seems to be possible (Brassica and other non-model plants). Maybe changes in base modification pattern can be found.

Mapping to snoRNAs:

Detection of new expressed snoRNAs is still possible in Arabidopsis and if so it is much more promising in plants like Brassica napus, Medicago truncatula or Lotus japonicus and all other.

Mapping to tRNAs:

Remove introns, add CCA ends, map smRNAs with 100%, 95% and 90% - this is the number of smRNAs mapped to tRNA (the same for rRNA, snoRNA etc.). Since tRNA derived smRNAs associate with AGOs, and one 5'-end fragment has been experimentally shown to be excised in a Dicer dependent manner, these fragments can constitute a new, yet unexplored category of regulatory molecules.

Expression of mature miRNAs : if multiple libraries clustering would be useful for a fast screen for similar responses to the same stimulus.

Conserved miRNA processing and new miRNA prediction:

Conserved: mature miR and hairpin similarity search,…blast mismatches…indels.

List of expressed genes, how many smRNAs are specifically excised from each precursor (a list for degradome analysis)? Let say more than 10% of all reads mapped to a hairpin in at least one library.

List of smRNAs which seem to be miRNA processing products but hairpins cannot be predicted, esp. those highly abundant and similar to conserved miRs. We can sequence their precursors by RACE.

On 100% identity based map smRNAs having more than 1 hit to gemome should be shown (g) or more than one to miRNA hairpins (mi). Just a technical remark: not wrapped lines of hairpin sequences in mapping files seem to be better.

Categories:

Conserved in all plant kingdom, conserved in a genus, species specific.

Analysis of miR processing:

Processing graph like a degradome plot. It can be combined with degradome data to confirm processing of a hairpin by a dicer like protein.

GFF files containing up to 2000 annotations can be opened in Vector NTI software, which is in my opinion the most user friendly tool to view and analyze sequences, to design primers and constructs. The useful database should be constructed like a GenBank – one can trim sequence of interest (on a webpage) from a genome or transcriptome with all annotations – it is easy to share data and find marked motifs of interest.

Small RNA database should be connected with a microarray or other transcriptomic database:….

A small RNA in a dataset has its all possible targets in A. thaliana transcriptome and genome (unknown/heterogenous transcript ends or unknown transcripts) to omit overloading, only of better hybridization parameters than the weakest of the experimentally proven pair of miRNA/mRNA.

A library of such pairs if once generated can be useful in maybe faster degradome analysis.

1. Classification of reads :

    A) similar or identical to rRNA (similar will come from modified nucleotides)

    B) similar or identical to tRNA(similar will come from abundant modified nucleotides)

    C) snoRNA

    D) miRNA hairpins or transcripts

    E) identical and complementary to ORFs, exon, intron, different splicing forms, if possible

    F) identical and complementary to 5'UTRs

    G) identical and complementary to 3'UTRs

    H) identical to overlapping transcripts (any of two): 5' and 3' overlaps cis and trans – overlap database needed

    I) repetitive elements (dispersed, long, short tandem repeats)

    J) Is the ratio ALL/rRNA/miRNA or All/rRNA/tRNA mapped reads the same in each sample?

K) Consider that not all 5’ and 3’-ends of transcripts are properly annotated – EST data can be helpful, also in the case of degradome data maybe flexible transcript ends should be used …if thereis a tag, extend mapping for next 50nt.

L) Graph or table showing diversity of lengths/position of 5' and 3' as a function of abundance of a specific miRNA

M)

2. Size distribution for all identified miRNAs in 8 libraries and in each separately – possible differences in DCL 1,2,3,4 activities, then unique miRNAs - for each treatment.

Degradome, which will be continued…

1. Coverage of transcriptome in %.
2. New transcripts? Combined with 5’-RATE or should be RACE confirmed.
3. Splicing forms? Which splicing forms are present in a sample? List of genes .
4. From degradome one can extract information about 5’-ends of most of mitochondrial and plastid transcripts. I have checked that t-elements are well represented – this is why mapping of smRNA to plastid and mitochondrial genomes can be interesting. Can we confirm known and find new editing sites?
5. Mapping to miRNA genes: processing - a combined plot would be nice: degradome and most abundant smRNA from a hairpin for all miRNA precursors including newly predicted ones.
6. A difficult thing: search for targets of all small RNAs which are expressed in a sample and a parsimonious option: which also can be attached to argonautes (limited data: Mi 2008).
7. Earlier: a list of all (how many?) condition specific mRNA cleavages can be found in compared libraries. Maybe only this group should be used for a first prediction of mRNA/sRNA interactions. Table: it is important to include a tag sequence, and an extended tag towards the end of a complementary strand of miRNA. Next important information: is the tag unique to a specified gene or not, if not: how many hits to genome does it have (not only to other annotated genes).
8. For how many obvious tags (miRNA/mRNA interaction is already well documented) can you find additional cleavage 10nt upstream? I see some cases and seem to be very interesting because it seems that if mRNA is cleaved, its further exonucleolytic degradation from 5’-to 3’ indicates interaction with a downstream miRNA. Just need more such examples.
9. Unique mapping is important to plan further RACE experiments towards validation of potential targets.