A method for de novo miRNA annotation using Prost! output (Thomas Desvignes; December 2018)

Initial considerations

First of all, *Prost!* annotates a sequence only if:

- a sequencing read aligns perfectly and on its full length to a sequence present in the annotation database provided to Prost! (Forward annotation).
- a sequence present in the annotation database aligns perfectly and across its full length to a sequencing read (Reverse annotation).
  - ⇨ Practically, this means that if, in the specie considered for de novo annotation, a miRNA is even one nucleotide different from any sequence present in the annotation dataset it will not be annotated.
  - ⇨ However, miRNAs are usually very well conserved throughout evolution and in most cases, sequences will receive an annotation for a miRNA from related species. In addition, if one of the two strand gets annotated, the complementary strand can be found easily by its neighboring location.

This annotation method considered, to perform a *de novo* annotation of a new species it is important to:

- Use as primary potential annotation (used as "in_species" annotation) the sequences available in the most closely related specie available.
- Use an exhaustive dataset for other species annotation in case for some miRNAs the species subject to annotation differs in sequence from the closely related species but not from other most distantly related species.

Also, to increase confidence in annotation, it is recommended to annotate sequencing reads from both strands of a miRNA (5p and 3p strands). However, in cases of asymmetrically expressed miRNAs (a strand being consistently much more expressed than the other one), the dataset may not contain sequencing reads for the complementary strand. To minimize this, it could be wise to use a low minimum count for sequence retention ('MinCount' parameter in Prost! setup).

Extracting data from the *Prost!* output file

Important note: *The following method is the method that I, Thomas Desvignes, use to generate de novo miRNA annotations. The following series of step, filters, and sequence selection is a method that proved to work but it is certainly not the only possible way. Each user may want to adapt these steps or use a completely differently method depending on the situation.*

1. **Create a duplicate of the "by_genomic_location" tab.** For the annotation miRNAs, the important information will be found in the tab "by_genomic_location" because annotating mature miRNAs shouldn't be dissociated from annotating miRNA genes and therefore loci. Creating a duplicated tab allows you to modify it without affecting the original tab.

2. **Filter out any sequence that don't match perfectly the genomic reference.** Sequences that don't match perfectly the genome can't be used to annotate a locus so they should be removed. For that simply retain only the genomic location bins that have a "1" in the "Designations" column (Column E).
3. **Filter out any sequence that have too many locations (i.e. repeats).** miRNAs can only originate from a limited number of loci (for some miRNAs it may still be up to 6 or 8, e.g. let-7 miRNAs). For that simply remove any genomic location bins that have a "TML:" in the "Locations" column (Column C).
4. **Filter out any sequence that are known to be from another type of RNA.** In smallRNA sequencing dataset, there is always a diversity of small sequences originating from snoRNA, t-RNAs, rRNA, etc. These fragments are *de facto* not miRNAs and can be filtered out. This information may however not be available for the species considered for annotation.

After you have filter all sequences that can't be miRNAs, many sequences can now be directly annotated using the putative annotation provided in column F. Annotation should be done one mature miRNA at a time and for each mature miRNA, a series of step is required to provide a confident annotation.

In order to keep good track of the progress, I usually create a side file to report for each gene (locus) the sequences for the 5p miRNA (and its position) and the 3p miRNA (and its position) (cf Supp Table 3 of the Prost publication).

5. **Select a miRNA and filter the column F for only that annotation.** You may then have one or more genomic location bins remaining.

- If only one genomic location bin remains after the name filtering, it is therefore likely that the genomic location(s) of that bin corresponds to miRNA locus (loci). Each genomic location should therefore be considered as a putative miRNA locus and needs attention. To validate it, it is important to look for the complementary strand. Therefore, remove the filtering for the specific mature miRNA and, using one genomic location at a time, filter the dataset for that specific genomic location extending the location.

For example, if the miRNA I want to annotate reports a genomic location of 'ChrX:11100-11120', I will filter the 'Locations' column for 'ChrX:11' so that all the genomic location bins that would be from around the studied one will be shown. The analysis of the sequences retrieved that way may reveal the presence of:

- Additional genomic location bins for the miRNA considered. These bins would have overlapping positions with the miRNA considered for annotation or maybe additional genomic locations compared to the sequence selected for verification => the sequence kept for annotation should correspond to the bin starter displaying the highest count.
- Genomic location bins corresponding to the complementary strand of the miRNA considered. These bins would be about 10-20 nucleotides away from the considered miRNA and should be on the same strand => the sequence kept for annotation should correspond to the bin starter display the highest count.
- Loop sequences which would have genomic location(s) almost or perfectly contiguous with the considered miRNA and be located in between the two mature strands.
- moRNAs (miRNA-offset) which would have genomic location(s) almost or perfectly contiguous with the considered miRNA but would be located at the end of one of the two mature strands.

- Additionally, this filtering method of displaying a large window of locations, may reveal the expression of clustered miRNAs. In the example above, the location filter step allows any genomic location bin within a 1000nt window, so if two or more genes are present in that window and form a cluster, it is very likely that mature miRNAs for these clustered genes will show up.

Each time a mature miRNA is chosen for annotation, it should correspond to the most expressed isomiR for that genomic location.

- <u>If more than one genomic location bin remains after the name filtering</u>, two main situations are possible:
  I. Only part of the possible genomic location(s) bins corresponds to true miRNA loci. In this case, the additional possible loci are often associated with the shortest isomiRs which, because shorter, may align incidentally at more than one location on the genome. Each additional location should however be carefully examined.

     *For example, "**TACAG**TACTGTGATAACTGAAG" (22nt) (miR-101a-3p) aligns to only one genomic location in the stickleback genome, while "TACTGTGATAACTGAAG" (17nt) (also annotated as miR-101a-3p) aligns to two genomic locations, but inspection of each locus reveals that the additional possible genomic location for the short sequence corresponds to an artefactual match.*

  II. All genomic location corresponds to true miRNA loci. In this case, the various isomiRs would generally be associated with additional annotation names.

     *For example, in stickleback, "TAGCAGCACGTAAATATTGG**C**" (21nt) corresponds to miR-16a-5p and has one genomic location; "TAGCAGCACGTAAATATTGG**AG**" (22nt) corresponds to miR-16c-5p and also has one genomic location; but "TAGCAGCACGTAAATATTGG" (20nt) corresponds to both miR-16a-5p and miR-16c-5p and has two genomic locations. In such case, I usually tend to favor the annotation of mature miRNAs that are unique for each locus to preserve as much gene-specific information as possible. Therefore annotating the 21nt and 22nt sequences as miR-16a-5p and miR-16c-5p respectively. However, if the sequence that Prost! annotated with two names has many more reads than both the singled-annotated sequences, then it may be better to annotate the sequence that has the most reads and would therefore be common to both loci.*

6. **Repeat this search and annotation steps for each miRNA that *Prost!* annotates.**

7. **Search for missing strands.** When everything that *Prost!* had suggested has been checked, some strands may be missing for some loci because their sequences were different from anything present in the annotation database. For each of those, simply perform the same location filtering as the one presented above. The missing complementary strand may be found this way, but will have no "in_species" annotation.

8. **Search for missing genes.** Some genes that may be expected to be present in the genome of the species considered may not have been found by *Prost!* because their sequence differs from what is present in the annotation dataset, or they may simply be absent, but in each

case these genes should be looked for. There are several methods to search for these miRNAs: narrowing down the genomic region by synteny analysis and search for reads mapping this region in the *Prost!* output file; search the *Prost!* output for the miRNA seed which is likely conserved, etc.…

Conclusions:

The steps presented above outline the strategy I use to create a *de novo* annotation. The complete process, obviously, contains subtleties, especially when it comes to making sure appropriate names are given to paralogous miRNAs. Therefore, it is important to assess all the sequences originating from each locus and potentially perform a conserved synteny analysis to validate the annotation.