Regulatory RNA in ruminants

Small RNAs in goats: miRNA and piRNA

Dott. Ing. Ilaria Fojadelli – PTP Bioinformatic Team
Definition and classification of small RNAs conventionally relies on their biogenesis mechanism.

Two relatively well-defined classes of small RNAs include microRNAs (miRNAs) and small interfering RNAs (siRNAs).

The biogenesis mechanism for piRNAs is currently unknown, but some studies report that they are a class of small non-coding RNA primarily expressed in germ cells that can silence transposons at the post-transcriptional level.*

*Prediction of piRNAs using transposon interaction and a support vector machine. Kai Wang et al.
## Functions of non-coding RNA in mammals

**Table 1: Main classes and functions of mammalian non-coding RNAs**

<table>
<thead>
<tr>
<th>ncRNA</th>
<th>No. of known transcripts</th>
<th>Transcript lengths (nucleotides; nt)</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precursors to short RNAs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miRNA</td>
<td>1,756</td>
<td>&gt;1,000</td>
<td>Precursors to short (21–23 nt) regulatory RNAs</td>
</tr>
<tr>
<td>snoRNA</td>
<td>1,521</td>
<td>&gt;100</td>
<td>Precursors to short (60–300 nt) RNAs that help to chemically modify other RNAs</td>
</tr>
<tr>
<td>snRNA</td>
<td>1,944</td>
<td>1,000</td>
<td>Precursors to short (150 nt) RNAs that assist in RNA splicing</td>
</tr>
<tr>
<td>piRNA</td>
<td>89</td>
<td>Unknown</td>
<td>Precursors to short (25–33 nt) RNAs that repress retrotransposition or repeat elements</td>
</tr>
<tr>
<td>CRNA</td>
<td>497</td>
<td>&gt;100</td>
<td>Precursors to short (73–93 nt) transfer RNAs</td>
</tr>
<tr>
<td><strong>Long ncRNAs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antisense ncRNA</td>
<td>5,446</td>
<td>100–&gt;1,000</td>
<td>Mostly unknown, but some are involved in gene regulation through RNA interference</td>
</tr>
<tr>
<td>Enhancer ncRNA (eRNA)⁵</td>
<td>&gt;2,000</td>
<td>&gt;1,000</td>
<td>Unknown</td>
</tr>
<tr>
<td>Enhancer ncRNA (meRNA)⁶</td>
<td>Not fully documented</td>
<td>As variable as the length of miRNAs</td>
<td>Unknown, but they resemble alternative gene transcripts</td>
</tr>
<tr>
<td>Intergenic ncRNA</td>
<td>6,742</td>
<td>$10^2$–$10^5$</td>
<td>Mostly unknown, but some are involved in gene regulation</td>
</tr>
<tr>
<td>Pseudogene ncRNA</td>
<td>680</td>
<td>$10^2$–$10^4$</td>
<td>Mostly unknown, but some are involved in regulation of miRNA</td>
</tr>
<tr>
<td>3' UTR ncRNA</td>
<td>12</td>
<td>&gt;100</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Non-coding RNAs (ncRNAs) are functional RNA transcripts that do not translate into proteins.

**microRNA (miRNA) and piwi-interacting RNA (piRNA)** play important roles in post-transcriptional regulation and are implicated in many essential biological processes.

The Drosha enzyme cleaves the pri-miRNA, resulting in a shorter **hairpin structure**, called the precursor miRNA (pre-miRNA), but there are alternative non-canonical biogenesis pathways to produce pre-miRNA without **Drosha** (for example mirtrons are miRNAs formed within the introns of a protein coding gene).
miRNA and piRNA: regulators of spermatogenesis in the adult testis in sheep

Small RNAs including microRNA (miRNAs) and PIWI-interacting RNAs (piRNAs) are regulators of spermatogenesis.

1. **miRNAs are small (more or less 22 nt)** endogenous RNAs that **negatively regulate gene expression** by targeting mRNA 3’ untranslated and/or coding regions.

1. **piRNAs are longer (more or less 26-33 nt)** than miRNAs and can bind to PIWI, a spermatogenesis-specific protein belonging to Argonaute protein family: they guide PIWI protein to **repress the transportable elements that protect genomic integrity**; they have derived from mRNAs a role in the regulation of gene expression

Reference: Roles of small RNAs in the effect of nutrition on apoptosis and spermatogenesis in the adult testis Guan et al. 2014
piRNA: what is already known?

- piRNAs lack clear secondary structure motifs, and primary sequence conservation except for enrichment for the presence of a uridine nucleotide at the 5’ first position of the transcript.
- 24–35 nt of length, most of them are encoded in genome clusters ranging from 1 to >100 kb long.
- There are both monodirectional clusters encoding piRNAs on one strand, and bidirectional clusters whose halves encode piRNAs on opposite strands.
- in Drosophila, piRNAs have the tendency to be expressed near telomere and centromere regions on the chromosome.

Reference: http://bioinformatics.oxfordjournals.org/content/30/17/i364.full
miRBase: the miRNA online reference database

miRBase provides the following services:

- The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed miR in the database), with information on the location and sequence of the mature miRNA sequence (termed mir). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.
- The miRBase Registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the help pages for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the miRBase announcements mailing list. Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is managed by the Griffiths-Jones lab at the Faculty of Life Sciences, University of Manchester with funding from the BBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute.

http://www.mirbase.org/index.shtml
piRNA online databases

- http://pirnabank.ibab.ac.in/index.shtml
- http://regulatoryrna.org/database/piRNA
- http://www.smallrnagroup.uni-mainz.de/piRNAclusterDB.html
piRNAs length in other species

Review: Zhang et al., 2014,
(A) Percentage of unique piRNA from each species in piRBase.
(B) Sum of piRNA sequences obtained by different experimental methods.
(C) Distribution of piRNA sequence lengths in piRBase.
Collecting Data: Illumina Next Generation Sequencing

Step 1: cDNA introduced into flowcell

Complementary adapter sequences and primers are ligated to the surface of the flowcell.

The adapter at one end of a library fragment hybridizes to a complementary adapter sequence on the surface.

Step 2: Hybridization and synthesis

The anchored fragment then bends toward the surface and hybridizes to a second complementary sequence which contains a primer.

The primer allows DNA polymerase to replicate the fragment in place via DNA synthesis.
Collecting Data: Illumina Next Generation Sequencing

**Step 3: Denaturation**
The double-stranded DNA is denatured, leaving two complementary fragments attached to the flowcell.

This process of hybridization, DNA synthesis and denaturation is repeated many times to create a cluster of fragments.

**Step 4: Complement fragments removed**
All complementary fragments are removed from the surface.

The resulting cluster consists of single-stranded, identical copies of the original library fragment, and is called a clonal cluster.
Step 5: Single-nucleotide DNA synthesis

Primers are bound to the free fragments ends in the cluster

Fluorescently-labeled, reversibly terminated nucleotides are washed over the surface

Step 6: Imaging

The color of the bound fluorophores reveals the identity of the nucleotides that were added to the cluster.

Clonal clusters amplify the signal that would be generated by a single library fragment
Bioinformatics predictions

Bioinformatics predicting models can be implemented to better recognize classes of structures from RNA or DNA not already well known to predict similar sequences finding patterns of recognition.

- Bioinformatic predictions can involve:
  - sequence similarity searches
  - multiple sequence alignments
  - identification and characterization of domains
  - secondary structure prediction
  - solvent accessibility prediction
  - automatic protein fold recognition
  - constructing three-dimensional models to atomic detail model validation.
How does MirDeep2 predictor work?

The miRDeep2 software can predict **novel miRNAs using a probabilistic model of miRNA biogenesis** to score compatibility of the position and frequency of sequenced RNA with the **secondary structure of the miRNA precursor**.

(A) the miRDeep2 module **identifies known and novel miRNAs** in high-throughput sequencing data

(B) the Mapper module processes Illumina output and **maps it to the reference genome** and

(C) the Quantifier module sums up read counts for **known miRNAs** in a sequencing data set
How does ShortStack predictor work?

☑ ShortStack discovers small RNA ‘clusters’ de novo, based on user-set thresholds (such as clusterize incoming min and max of the dicer call) and annotates clusters with respect to small RNA size, orientation, and repetitiveness.

☑ ShortStack also discovers and annotates MIRNA genes following a score of probability to have found a MIRNA structure.

☑ It is able to flag a cluster corresponding to a structure in a given DB as image -> ex. Validated miRNAs in miRBase.
How does piRNAs online predictor work? (Zhang)

Without an annotated reference genome it finds putative piRNAs
The algorithm is based on the frequency of the positive k-mer found in the structure: a 1364 positive vector is calculated and the weight of each positive-kmer is evaluated by a probabilistic model.
GenHome project: goat dataset

- 3 pre-pubescent
- 3 pubescent
- 3 in test phase (adult)

X 9 pituitary (x 3 phases)

X 3 hypothalamus (test)

X 3 ovary (test)
GenHome project : reads length
GenHome project: reads length in hypothalamus from goats in different developmental stages

- It seems to have a significant **increment of the reads in the range 29-34 bp for pubescent goats**

*Note: the data has been normalized assuming to have had 1’000’000 of reads for each tissue*
GenHome project: reads length in hypothalamus from goats in different developmental stages

✧ It seems to have a significant increment of the reads in the range 29-34 bp for pubescent goats

Note: the data has been normalized assuming to have had 1'000'000 of reads for each tissue
GenHome project: reads length in different organs (adult goats)

- It seems to have a significant **increment of the reads in the range 28-34 in ovary**

*Note: the data has been normalized assuming to have had 1’000’000 of reads for each tissue*
GenHome project: reads length in different organs (adult goats)

- It seems to have a significant **increment of the reads in the range 28-34 in ovary**

*Note: the data has been normalized assuming to have had 1’000’000 of reads for each tissue*
Genhome project: novel miRNAs

566 cluster of miRNA miRDeep2 - percentage novel/known
- Novel miRNA: 82.05%
- known miRNA: 7.95%

* 45/265 cluster in miRBase recognized

682 cluster of miRNA ShortStack - percentage novel/known
- Novel miRNA: 68.77%
- known miRNA: 31.23%

* 213/265 cluster in miRBase recognized

Intersecting 566 vs 682 cluster: 277 clusters with a 100% overlap (even if ≠ length)
- 192/277 cluster in miRBase recognized by Shortstack
- 29/277 cluster in miRBase recognized by mirDeep2
Genhome project: putative miRNA (chromosome distribution)

The putative miRNA that can be proposed using the prediction of the different predictors.

Comparable predictions:

- miRNA secondary structure features are known
- miRDeep2 predictions are based on secondary structure recognition
- Shortstack predictions are mainly based on sequence length, orientation and clustering.
This predictor reveals a greater incidence of piRNAs in ovarian and pituitary tissues.

Note: the 102 piRNAs have been normalized assuming to have had 1’000’000 of reads for each tissue.
GenHome project: comparison in putative piRNA length

Fig. 1 putative piRNA from Zhang predictor (on line predictor) length distribution. Original dataset: GenHome goat reads between 26-33 bp in length

Fig. 2 putative piRNA length distribution from Shortstack predictor. Original dataset: all GenHome goat reads
GenHome project: putative piRNA (chromosome distribution)

- The two models perform differently, due to the different assumptions.
- Shortstack clusterizes reads and classifies them according to length.
- Zhang predictor works on a few known features of the primary structure.
- A lot of investigation is still to be done.
Similarity search versus piRBase sequences

BLAST parameters:
- Coverage $\geq 80\%$
- Identity $\geq 80\%$

VS VS
VS
VS

piRNAs are supposed to be species-specific

only 2 sequences
Intersect with bovine piRNAs from GenHome project (motile spermatozoa)

BLAST parameters:  Coverage >= 80%
                    Identity >= 80%

- Supporting more evidence that piRNAs are probably species-specific
New algorithms proposed from community in piRNA detection: 1. 

McRUM + CFS (Menor et al.)

- The correlation-based feature selection (CFS) method proposed by Menor et al. (Int J Mol Sci. 2015 Jan; 16(1): 1466–1481) avoids the need of reference genome and the computationally expensive pairwise folding of the reads required by existing models.

- It uses multiclass relevant units machine (McRUM) method for classification, to achieve compact models appropriate for age scale analysis.

- It uses correlation-based feature selection (CFS) to select a subset of features on which to build classifier models, considering 1389 features, including 1364 unique k-mers for k=1 to 5 of the nucleotide composition in the seed region (first 8 positions).
New algorithms proposed from community in piRNA detection: 1. McRUM + CFS (Menor et al.)

- The CFS algorithm selected 154 features (such as the four binary features representing A,C,G and U of the first nucleotide and the frequency of the two-mer CG)
- It has been more powerful in 60% of true positive detection of the online predictor.

Results of the method in characteristics detecting:

1. both miRNA and piRNA:
   - Tend to start with a U base

2. Only for piRNA:
   - CG frequency is biased toward low scores
It uses piRNAs-trasposons interaction information: the piRNAs were aligned to trasposons with a maximum of three mismatches.

Triplet elements combining structure and sequence information were extracted from piRNAs trasposons matching/pairing duplexes.

Support Vector Machine (SVM) is used on these features to classify real/pseudo piRNAs.

It is available online

Results:
it achieved to predict correctly human, mouse and rat piRNAs with an overall accuracy of 90.6%
Connection with epigenetic

- The piRNA complexes contribute to epigenetic regulation and post-transcriptional silencing of retrotransposition, particularly in the germ line cells, and to tumorigenesis.

- Like miRNA, piRNA molecules are associated with proteins of the Ago/Piwi family to execute sequence-specific gene silencing.

- piRNA molecules may fine-tune gene expression by mediating epigenetic modifications of heterochromatin.

- Recent data have suggested piRNA expression and biological activity in somatic cells as well.
Conclusions

1. We found some putative novel miRNAs

1. We scanned small RNAs with different predictors to obtain a list of putative novel piRNAs to improve knowledge on the goat genome.

2. We found that piRNAs seem to be:
   ✩ species-specific
   ✩ more expressed in the ovary
   ✩ dependent from developmental stage of the animal
Future aims

1. To provide **information on smallRNA position** to complete the annotation of the goat genome.

2. To try improved predictors for piRNAs in goats, i.e. the CFS/McRMs algorithm used in Menor *et al.* work without a reference genome and the algorithm ant the “Piano” algorithm.

3. Using MBD-Seq of goat data in the GenHome project we propose to study the **incidence of miRNAs and piRNA in the methylated region** to investigate their effect in the different position and/or tissues and/or developmental stage.

4. Following the 3’ point to study if the **presence of miRNAs or piRNAs** could be responsible of mediating gene expression of biological processes in goats (such as genes involved in lactation).

5. In general, to **improve the knowledge of the goat genome**
Thanks for the attention