Mármol-Sánchez, Emilio; Cirera, Susanna; Quintanilla Aguado, Raquel; [et al.]. «Discovery and annotation of novel microRNAs in the porcine genome by using a semi-supervised transductive learning approach». Genomics, Vol. 112, Issue 3 (May 2020), p. 2107-2118. DOI 10.1016/j.ygeno.2019.12.005

Discovery and annotation of novel microRNAs in the porcine genome by using a semi-supervised transductive learning approach

Emilio Mármol-Sánchez[1], Susanna Cirera[2], Raquel Quintanilla[3], Albert Pla[4], Marcel Amills[1,5]

[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.

[2]Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark.

[3]Animal Breeding and Genetics Program, Institute for Research and Technology in Food and Agriculture (IRTA), Torre Marimon, 08140 Caldes de Montbui, Spain.

[4]Department of Medical Genetics, University of Oslo and Oslo University Hospital, Oslo, Norway.

[5]Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

Corresponding author: Emilio Mármol-Sánchez. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. Phone: +34 935636600. Email: emilio.marmol@cragenomica.es

Highlights

- Motif search improved pre-miRNA reconstruction from mature microRNA sequences.
- Semi-supervised methods outperformed canonical supervised classification algorithms.
- The presence of multiple isomiRs in the porcine muscle miRNA repertoire was uncovered.
- A total of 47 novel microRNA genes were identified in the porcine genome.
- RT-qPCR analyses allowed us to confirm the existence of three novel porcine microRNAs.

Abstract

Despite the broad variety of available microRNA (miRNA) prediction tools, their application to the discovery and annotation of novel miRNA genes in domestic species is still limited. In this study we designed a comprehensive pipeline (eMIRNA) for miRNA identification in the yet poorly annotated porcine genome and demonstrated the usefulness of implementing a motif search positional refinement strategy for the accurate determination of precursor miRNA boundaries. The small RNA fraction from *gluteus medius* skeletal muscle of 48 Duroc gilts was sequenced and used for the prediction of novel miRNA loci. Additionally, we selected the human miRNA annotation for a homology-based search of porcine miRNAs with orthologous genes in the human genome. A total of 20 novel expressed miRNAs were identified in the porcine muscle transcriptome and 27 additional novel porcine miRNAs were also detected by homology-based search using the human miRNA annotation. The existence of three selected novel miRNAs (ssc-miR-483, ssc-miR484 and ssc-miR-200a) was further confirmed by reverse transcription quantitative real-time PCR analyses in the muscle and liver tissues of Göttingen minipigs. In summary, the eMIRNA pipeline presented in the current work allowed us to expand the catalogue of porcine miRNAs and showed better performance than other commonly used miRNA prediction approaches. More importantly, the flexibility of our pipeline makes possible its application in other yet poorly annotated non-model species.

Keywords: MicroRNA discovery; Motif search; Porcine skeletal muscle; Semi-supervised learning; Small RNA-seq.

Introduction

The accurate annotation of a comprehensive set of miRNAs in different species has been challenging since the first genome assemblies were published, although an ever-increasing amount of knowledge about miRNA diversity across species has been accumulating during the past years, being available in public databases [1-3]. Despite these advances, many commonly studied domestic species still lack a complete and reliable set of annotated miRNAs in their genomes [1].

The computational prediction of miRNAs in sequenced genomes initially relied on the strong conservation of mature miRNA sequences across closely related species [4,5], taking advantage of homology-based comparisons between well annotated genome assemblies and other poorly annotated organisms [6-8]. Other approaches focused on rule-based classification, integrating other sources of information such as sequencing data or structural features to identify novel miRNAs [9-12]. More recently, several Machine Learning (ML) approaches have been proposed for miRNA prediction. Different tools have addressed the problem of correctly classifying miRNAs by training ML algorithms with a set of positive (annotated miRNAs) and negative (other non-miRNA sequences) data sets. [13-16]. Nevertheless, despite the broad array of available tools for novel miRNA identification, their application to the discovery and annotation of novel miRNAs in domestic species is still limited [17-25]. Moreover, the majority of miRNA surveys carried out in domestic species do not generally take into account several issues regarding miRNA genes prediction that have recently emerged. For instance, the set of positive training annotated miRNAs often include misclassified sequences [26,27], whereas the negative class is sometimes not clearly defined, i.e. different types of sequences have been used as negative data sets (coding regions, pseudo-hairpins, non-coding hairpins or artificial randomized miRNA sequences). Despite some efforts [28], obtaining a truly representative negative class is still challenging and few approaches have critically addressed this important issue [29-31]. Besides, miRNAs are thought to encompass a small percentage of the total non-coding transcriptomic repertoire, with thousands of other non-miRNA hairpin-like RNA molecules that represent a major fraction of it. This circumstance contributes to create a high class-imbalance between positive and negative sequences. Different approaches have dealt with such phenomenon [32], but recent studies have shown that commonly used techniques for solving the high-class imbalance problem in microRNA prediction may not be suited to a real-case classification scenario [15].

In this study we present eMIRNA, a bioinformatics pipeline for miRNA discovery and annotation in sequenced genomes. The proposed pipeline implements a semi-supervised transductive learning approach to predict and annotate novel microRNAs in the porcine genome, overcoming several of the drawbacks outlined above. In order to validate the performance of our pipeline in a real-case scenario, we have applied it to the analysis of a data set comprising the small RNA fraction of *gluteus medius* skeletal muscle from a population of 48 Duroc gilts [33,34]. Furthermore, making use of the better annotated *H. sapiens* miRNAome, an additional set of novel porcine miRNA genes were identified based on a homology-based search approach. Finally, some of the identified novel porcine miRNA candidates were independently validated in a Göttingen minipig population, investigating their expression in skeletal muscle and liver tissues.

Materials and methods

A detailed flow chart depicting all steps described in the eMIRNA pipeline is shown in Figure 1. Additional instructions and modular scripts needed for the implementation of eMIRNA are available at: https://github.com/emarmolsanchez/eMIRNA/.

Positive and negative training data sets

To define the corresponding positive (annotated miRNAs) data set required for novel miRNA prediction, two approaches were considered:

1) The annotated pre-miRNA coordinates in Sscrofa11.1 genome assembly were obtained from Ensembl repositories, release version 97 (http://www.ensembl.org/info/data/ftp/index.html), and the corresponding sequences were extracted from the pig reference genome by using the BEDTools suite v2.27.0

software [35]. miRNA loci located in scaffolds were removed from further analyses, resulting in a total of 484 annotated porcine miRNA genes. Sequence repeats from pre-miRNA duplicated elements were removed from the retrieved positive data set by using the CD-HIT Suite [36] with a 0.9 sequence identity cut-off value (i.e. sequences showing a similarity ≥ 90% to each other were removed and only unique representative pre-miRNA candidates were retained). Moreover, to avoid the inclusion of miss-annotated miRNA loci, an additional filtering based on secondary structure folding was applied. To this end, the RNAfold tool from the ViennaRNA Package 2.0 [37] was used to select sequences with canonical pre-miRNA hairpin secondary structures (stem-loop conformation with one single terminal loop and two stems). Sequences that failed to comply with required folding structure pre-requisites were removed.

2) In the second approach, the curated miRNA annotation for Sscrofa11.1 available in the miRCarta database [2] was retrieved, and the same pre-filtering criteria based on sequence identity and secondary structure employed in the analysis of the Ensembl data set were applied. The miRCarta database [2] integrates one of the most comprehensive and curated databases for miRNA annotation and functional activity, aiming to overcome the limitations of other widely used miRNA databases such as miRBase [1].

Regarding the negative data set (other hairpin-like sequences), two different data sources were used. First, the annotated non-coding transcripts in Ensembl repositories were retrieved and non-miRNA sequences were retained. Analogously to what was implemented for the positive data set, identity by sequence and secondary structure pre-filters were applied, and non-miRNA non-coding hairpin-like unique sequences were obtained. Only sequences ranging from 50 up to 150 nucleotides (nt) were retained, thus removing hairpin-like long non-coding RNAs from the negative data set. Additionally, a set of unlabeled sequences within the porcine reference genome (Sscrofa11.1) were generated by extracting candidate pre-miRNA-like sequences from random blocks of 1 Megabase (Mb) in each of the chromosomes of the porcine assembly with the *HextractoR* package [38], and the previously described pre-filters for the negative class were subsequently applied.

Obtaining putative miRNA candidate sequences from the porcine genome

In order to test our method with pig transcriptomic data, a small RNA-seq data set was generated by sequencing the muscle transcriptome of 48 gilts used in two previous studies [33,34]. Upon collection, muscle samples were individually submerged in RNAlater and snap-frozen in liquid nitrogen. Samples were pulverized and homogenized in 1 ml of TRI Reagent (Thermo Fisher Scientific, Barcelona, Spain). Total RNA was isolated with the RiboPure kit (Ambion, Austin, TX). A Nanodrop ND-100 spectrophotometer (Thermo Fisher Scientific, Barcelona, Spain) was used to assess RNA concentration and quality. RNA integrity expressed in RNA Integrity Number (RIN) units was measured with a Bionalyzer-2100 equipment (Agilent Technologies Inc., Santa Clara, CA). High quality RNA samples were then submitted to Sistemas Genómicos S.L. (https://www.sistemasgenomicos.com) for small RNA sequencing. Library preparation for each individual sample was carried out with the TruSeq Small RNA Sample Preparation Kit (Illumina Inc., USA) and small RNA libraries were single-end sequenced (1 × 50 bp) in a HiSeq 2500 platform (Illumina Inc., CA).

FASTQ sequence files were subjected to a quality control check as reported by Cardoso et al. [33]. After preliminary quality-based filtering, sequencing adaptors were trimmed with the Cutadapt software [39] and an acceptance sequence window of 15–30 nt per read was established. Processed FASTQ files from all sequenced samples (N = 48) were pooled and collapsed to unique FASTA sequences with the FASTQ collapser tool from

FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Unique FASTA sequences represented by >10 reads-per-million (RPM) were considered to be significantly expressed above the background noise [40], and thus selected for further analyses (File S1). The CD-HIT Suite [36] was employed to build sequence clusters with >0.9 sequence identity.

Furthermore, the human mature miRNA coordinates were obtained from Ensembl repositories and the corresponding sequences were retrieved from the GRCh38.p12 assembly. Pre-filtering based on sequence identity was applied and a set of non-redundant human mature miRNAs was generated for homology-based search in the Sscrofa11.1 porcine assembly (File S2).

Pre-miRNA reconstruction by sequence elongation and motif search

Once putative mature miRNA candidate sequences from the small RNA-seq data set and the human mature miRNA sequences were retrieved, they were aligned against the porcine reference assembly (Sscrofa11.1) with the Bowtie aligner [41] and the following specifications for short reads: 1) allowing 2 mismatches within the entire aligned sequence with respect to the reference assembly, 2) removing reads with >50 putative mapping sites and 3) reporting first single best stratum alignment (*bowtie -n 2 -l 25 -m 50 -k 1 --best --strata*). Reported alignment genome positions for successfully mapped putative mature miRNAs were elongated upstream and downstream, thus ensuring an adequate pre-miRNA reconstruction. As no prior knowledge about the 3p or 5p identity of putative mature miRNA sequences was available for porcine small RNA-seq data, two candidate pre-miRNA structures were generated for each expressed sequence. The same procedure was applied to human mature miRNAs when 3p or 5p identity was not specified. Candidate sequences that were aligned and extracted from overlapping regions corresponding to other annotated non-miRNA non-coding loci were discarded from further analyses.

Elongation patterns were based on previously reported pre-miRNA favored size, with a stem length of ~35 ± 3 nt and an apical loop ≥10 nt [42,43]. With these specifications, we established two upstream and three downstream elongation pattern combinations: 1) from the starting genome position of each aligned sequence, 15 and 30 nt were added upstream, beginning from each mature miRNA sequence start position. 2) Additionally, 60, 70 and 80 nt were added from each miRNA end position, resulting in the following elongation pattern combinations for each candidate sequence: 15/60, 30/60, 15/70, 30/70, 15/80 and 30/80 added nt (i.e. we generated a total of 12 putative elongated pre-miRNA candidates per each aligned sequence). Besides, the presence of flanking microprocessor motifs was assessed for positionally correcting the elongated pre-miRNA candidate sequences. Downstream CNNC and upstream UG motifs were assessed within the 30/60, 30/70 and 30/80 elongated candidates for each sequence, as described in [44], whereas downstream mismatched GHG and upstream CHC motifs were searched in 15/60, 15/70 and 15/80 candidates [42].

To determine the most prevalent positional range of flanking processing motifs surrounding pre-miRNA sequences in the porcine genome, 30 and 15 nt were added at the flanking positions of annotated porcine pre-miRNAs available at the curated miRCarta database [2]. The presence of CNNC and UG motifs within flanking ±30 nt, as well as GHG and CHC motifs within ±15 nt was hence assessed. According to positional results (Figure 2A), the CNNC and UG flanking motifs appeared more prominently located 18 nt after miRNA gene ending and 12 nt before miRNA starting points, respectively. Therefore, when downstream CNNC or upstream UG motifs were found within ±30 nt flanking windows along pre-miRNA candidates, −18 and +12 nt positions were added from CNNC and UG motifs location, respectively, so as to establish accurate miRNA genes boundaries determined by the microprocessor machinery. In the event that none of the aforementioned motifs within flanking upstream and/or downstream defined regions were found, the original elongated pre-miRNA candidates with no motif-based positional refinement were kept.

Selecting putative pre-miRNA candidate sequences based on structural integrity

To better assess the optimal elongation pattern for each candidate sequence, the structural stability of the 12 pre-miRNA candidates per single sequence was determined based on the randfold algorithm [45]. This approach assumes the estimated minimum free energy (MFE) of the folded pre-miRNA hairpin to be consistently lower than that of other random sequences resembling hairpin-like folded structures [45]. Based on this property of pre-miRNA sequences, we implemented a Monte Carlo randomization test to select the most stable hairpin, i.e. those having the least folding minimum free energy (MFE) values among the 12 previously generated candidates during pre-miRNA elongation reconstruction for each of the analyzed sequences. To this end, we generated a total of 100 randomized sequences per candidate by shuffling their nucleotide distribution while maintaining k-let counts [46]. The corresponding MFE values for each shuffled and original hairpin-folded sequences were calculated with the RNAfold tool [37] and the structural integrity score (p) was defined as:

$$p = \frac{R}{N+1}$$

where R is the number of randomized sequences having an MFE value equal or smaller than that of the MFE value of the original sequence and N is the number of generated iterations (100 in this study).

Subsequently, the candidate sequence showing the higher structural integrity (i.e. the one showing the smallest p score) among all 12 generated pre-miRNA candidates per sequence was selected. The proportion of the most structurally stable sequences for each elongation pattern is shown in Figure 2B. When two or more sequences had equal p scores (i.e. they had equivalent structural stability irrespective of the elongation pattern) the reconstructed candidates belonging to the motif-corrected (if available) and shortest elongation pattern were retained. The proportion of each elongation pattern selected as the most structurally stable among all 12 tested patterns from expression-based and homology-based data is shown in Figure 2C and D, respectively.

Candidates classification with semi-supervised transductive learning

After defining training and candidate data sets, we selected a total of 100 features representing structural and statistical properties from each pre-defined sequence. These extracted features have been previously reported in other state-of-the-art methods and thoroughly reviewed in [47]. A complete list of all used features is shown in Table 1.

For pre-miRNA classification, the *miRNAss* algorithm proposed by Yones et al. [31] was applied. This method implements a semi-supervised transductive learning scheme by using well defined labeled cases, either positives (annotated pre-miRNAs) or negatives (comprising other annotated non-coding hairpin-like sequences and unlabeled cases with unknown hairpins), in order to draw a graph-based representation of each sequence based on input features. Each node in the graph represents a sequence, whereas the corresponding edges account for the expected similarities among them. In order to accurately represent the spatial distribution and connections of each node, the feature importance is obtained by applying the Relief-F algorithm [48,49], where k-nearest predictors are weighted based on conditional dependencies among all the considered features and the response vector of labels. This algorithm penalizes those predictor features giving different values to k-neighbors from the same label class and vice versa. After graph construction, a prediction score is assigned to each sequence node [31].

Sscrofa11.1 pre-miRNA sequences from Ensembl and miRCarta databases were evaluated and different imbalance ratios between positive (taken as reference) and negative data sets were applied to assess the performance of the classification algorithm for miRNA discovery in the porcine genome (i.e. 1:1, 1:2, 1:10, 1:20, 1:40, 1:60, 1:80, 1:100, 1:150 and 1:200 imbalance ratios were considered). Labeled sequences comprised annotated pre-miRNAs (+1) as positive sequences, while other non-coding hairpin-like transcripts (−1) were considered as negative. Genome-wide randomly extracted hairpins were assigned as unlabeled cases (0) within the negative data set.

Testing subsets were randomly assigned from all proposed imbalanced training data set combinations using a 0.25 ratio. The performance of the classification algorithm for miRNA identification was assessed with a total of 100 random Monte Carlo iterations and average performance measures based on Sensitivity (SE), Specificity (SP), Accuracy (Acc), F-1 score (F1) and Adjusted Geometric-mean (Agm) [50] were estimated (Figure 3A). Furthermore, we evaluated the performance for each imbalance scenario by computing the corresponding Receiver Operating Characteristics (ROC) curves and the Precision-Recall (PR) curves. PR curves can be more informative than ROC curves for highly imbalanced data sets [51]. ROC and PR curves as well as the corresponding Areas under the curve (AUC) estimates are shown in Figure S1 and Table S1. The ability of the algorithm to correctly classify the list of Ensembl and miRCarta annotated porcine miRNAs was also assessed by incorporating the positive data set as unlabeled candidate sequences during the classification process in each of the defined imbalance scenarios. Results for annotated porcine miRNAs assignment are shown in Table S2.

Finally, the reconstructed expressed candidate sequences from the porcine small RNA-seq data and *H. sapiens* homologous miRNAs detected in the porcine genome were used for identifying putative novel miRNAs. For this purpose, annotated pre-miRNAs from the Ensembl database were used as positive class and other hairpin-like sequences were considered as either negative or unlabeled sequences. Candidates classification was implemented with all previously proposed imbalance ratios. In order to reduce the false positive rate (i.e. reducing the misclassification of non-miRNA short hairpins as true miRNA candidates), the Ensembl miRNA data set was defined as the positive class, due to its higher overall reported specificity (Figure 3A and B). Prediction of novel miRNA candidates was carried independently with every defined imbalance ratio. Only candidates consistently reported as putative miRNAs in all imbalance scenarios were kept in order to minimize the number of false positive miRNA candidates, albeit probably at the expense of increasing the false negative rate.

Besides, for homology-based predicted novel pre-miRNA candidates, we calculated the proportion of shared neighboring genes (setting a 2 Mb window before and after each annotated human miRNA detected in the porcine genome) present in both *S. scrofa* and *H. sapiens* assemblies and expressed as a Neighborhood Score (N):

$$N = \frac{G_r \cap G_i}{G_r}$$

where Gr is the number of orthologous genes within the 4 Mb window in the model species (*H. sapiens*) and Gi is the number of genes within the same window in the species of interest (*S. scrofa*). Only homology-based novel pre-miRNA candidates with N > 0.1 were considered for further analyses, based on the assumption that microRNAs residing in genomic regions with surrounding and/or host genes phylogenetically conserved across species are more prone to be integrated in biologically relevant transcriptional networks [52].

Benchmarking for miRNA prediction performance

One of the most cited and used prediction miRNA algorithms is miRDeep. This tool was developed by Friedländer et al. [53], and further improvements were made in subsequent updates [11,54]. This algorithm implements a series of heuristics to compute a score for each miRNA candidate expressing the log-odds probability of a sequence being a true miRNA gene against the probability of being a miRNA-like pseudo-hairpin [53]. In order to benchmark the eMIRNA pipeline compared with the widely used miRDeep approach, we used the miRDeep2 algorithm [54] to identify novel and annotated miRNAs by using the same small RNA-seq data set employed for *de novo* miRNA identification with the eMIRNA pipeline. To ensure a fair comparison, the arf alignment file needed for running the miRDeep2 software was generated from the eMIRNA alignment pipeline using the bowtie tool (*bowtie -n 2 -l 25 -m 50 -k 1 --best --strata*) on pre-filtered expressed small RNA sequences generated in this study. After running the miRDeep2 algorithm, both novel and already annotated pre-miRNA candidates were compared with those obtained with the eMIRNA pipeline. The positional accuracy of the annotated pre-miRNA candidates concurrently identified with both approaches was then determined using the

Ensembl annotation available for the Sscrofa11.1 assembly. To further determine which of the two approaches provided a better positional annotation of predicted miRNAs, the deviation rate (dr) of each miRNA gene commonly detected was calculated for both eMIRNA and miRDeep2, expressed as the average number of upstream and downstream overhanging nucleotides compared with the latest porcine miRNA Ensembl annotation (v97). The differential deviation estimate (ΔD) was assessed separately for each predicted pre-miRNA candidate, as follows:

$$\Delta D = eMIRNA_{dr} - miRDeep2_{dr}$$

Additionally, the performance statistics of the semi-supervised transductive learning method [31] implemented in the eMIRNA pipeline was compared with other canonical widely used state-of-the-art supervised ML approaches for miRNA prediction, such as Support Vector Machine (SVM), Random Forest (RF), K-nearest Neighbors (KNN), Naïve Bayes (NB), Extreme Gradient Boosting Trees (XGB) and Light Gradient Boosting Trees (lGBM). Only labeled positive and negative data sets were used for comparison between semi-supervised and supervised algorithms. Training and testing subsets were randomly generated with a 0.25 ratio for testing data and commonly used with all the proposed methods. No imbalance correcting procedure was applied. The comparative performance of these tools was assessed on the basis of SE, SP, F1-score, ROC and PR curves obtained for each algorithm implementation. SVM, RF, KNN and NB algorithms were trained allowing 10 iterations for parameter tuning and a 10-fold cross-validation scheme, using built-in functions included in the *caret* R package [55]. The *xgboost* [56] and *lightgbm* (https://github.com/microsoft/LightGBM/tree/master/R-package) R packages with default parameters were employed for the training of XGB and lGBM classifiers, respectively.

Experimental confirmation of novel identified porcine miRNAs through the RT-qPCR analysis of an independent Göttingen minipig population

In order to investigate the existence of several of the novel putative predicted miRNAs in the porcine genome, three well established orthologous novel miRNA candidates detected by homology-based search and not previously annotated in the Sscrofa11.1 assembly were selected (hsa-miR-483-3p, hsa-miR-484-5p and hsa-miR-200a-3p). The existence of miRNA genes orthologous to hsa-miR-483-3p and hsa-miR-484-5p was supported by the identification of the corresponding expressed mature miRNA sequences in our small RNA-seq data set. Transcripts corresponding to hsa-miR-200a-3p were detected at very low expression levels (RPM < 10) in the porcine skeletal muscle transcriptomic data, so they were not considered as biologically relevant or functionally active in our experimental conditions. *Longissimus dorsi* muscle and liver RNA samples were collected from an independent Göttingen minipig population [57]. A total of 7 extracted RNA samples from muscle and liver tissues were randomly selected and cDNA synthesis was carried out as reported by Balcells et al. [58]. Primers for the qPCR amplification of miRNAs were designed with the miRprimer software [59] according to described protocols [60] and they are indicated in Table S3.

MiRspecific qPCR was performed on a MX3005P machine (Stratagene, USA). Briefly, 1 µl of cDNA diluted 8 fold, 5 µl of 2× QuantiFast SYBR Green PCR master mix (Qiagen, Germany) and 250 nM of each primer (Table S3) were mixed in a final volume of 10 µl. Cycling conditions were: 95 °C for 5 min followed by 40 cycles of 95 °C for 10 s and 60 °C for 30 s. Melting curve analyses (60 °C to 99 °C) were performed after completing amplification reaction to ensure the specificity of the assays. Data were processed with the MxPro qPCR associated software. Assays were considered successful when: 1) the melting curve was specific (1 single peak) and 2) the samples had Cq values <33 cycles (i.e. sufficiently expressed to be considered biologically functional). Finally, amplified products for muscle and liver samples were visually inspected by electrophoresis in a 3% agarose gel.

Results

Motif-based positional refinement enhances structural stability of pre-miRNA candidates

We have evaluated the usefulness of previously reported flanking motifs that enhance pre-miRNA processing [42,44] as possible novel determinants for pre-miRNA reconstruction from mature sequences. The presence of UG and CHC motifs in upstream flanking regions as well as of downstream CNNC and GHG motifs was assessed in the curated porcine miRNA annotation available in the miRCarta database [2] (Figure 2A). Consistent with data reported by Fang et al. [42] and Auyeung et al. [44], the most common flanking upstream positions for UG and CHC motifs from the 5′ start of the porcine pre-miRNA genes were −13/−12 and −7/−5, respectively, whereas for downstream CNNC and GHG motifs, the most common position from the 3’end of the pre-miRNA genes were +18/+21 and +4/+6 (Figure 2A).

Moreover, we determined the percentage of annotated porcine miRNAs that were surrounded by the aforementioned processing motifs, allowing ±2 nt of positional variation from their corresponding expected sites. From a total of 328 confidently annotated porcine pre-miRNAs in the miRCarta database [2], CNNC, UG, GHG and CHC flanking motifs were found in 53.05%, 42.68%, 30.79% and 33.54% of the sequences, respectively. The high frequency of the CNNC motif agrees well with its key role in the correct Drosha ribonuclease III (DROSHA) positioning through the recruitment of Serine and Arginine rich splicing factor 3 (SRSF3) at the basal junction of the processed pri-miRNA [61]. The proportion of the three other flanking motifs were also consistent with previously reported surveys [42,44].

To further elucidate the contribution of each motif to better delineate the boundaries of pri-miRNA processing, we compared the structural stability (i.e. the estimated p score of the hairpin secondary structure with the randfold approach [45]) for every pre-miRNA candidate in each of the 12 generated elongation patterns per sequence (15/60, 30/60, 15/70, 30/70, 15/80 and 30/80, with and without taking into account motif search positional refinement). As depicted in Figure 2B, predictions of candidate miRNA sequences based on positional information obtained through processing motif search showed a consistently increased structural stability compared with non-positionally corrected original sequences. This phenomenon was less evident for shorter elongation patterns, where the structural stability of the positionally corrected hairpins resembled that of non-corrected candidates (Figure 2B). In certain cases, both approaches resulted in equally stable secondary structures. Furthermore, shorter elongation patterns appeared to be more favored than their longer counterparts, showing higher overall structural stability both in small RNA-seq and homology-based derived candidate sequences (Figure 2C and D). This result highlights that the preferred length for pre-miRNA processed transcripts would be approximately in the range of 80 to 90 nt, with few cases showing longer stable hairpin structures. Interestingly, this favored pre-miRNA length interval coincides with that reported by Roden et al. [43], who determined a preferred 2×

stem length of 35 nt and a terminal loop of ~10 nt, accounting for a total pre-miRNA sequence length of ~80 nt. Indeed, the average length of annotated pre-miRNAs in the porcine genome after filtering for secondary structure and sequence similarity was 84.63 nt, also in accordance with results obtained after selecting the most structurally stable elongation pattern from all generated candidates per sequence.

Classifier performance and feature importance

For assessing the performance of transductive semi-supervised miRNA classification on the porcine transcriptome, Ensembl and miRCarta positive pre-filtered porcine miRNA data sets (415 Ensembl and 244 miRCarta non-redundant hairpin-like stable annotated miRNAs) were tested against selected non-coding hairpin-like sequences (252 annotated non-coding hairpin-like RNAs other than miRNAs) and different imbalance ratios were applied by incorporating genome-wide randomly extracted hairpins (unlabeled). Overall, SE and SP obtained with the Ensembl miRNA data set (Figure 3A) were slightly better than those inferred for the miRCarta data set (Figure 3B). Ensembl average SE and SP were 0.9199 and 0.9101 respectively, whereas results obtained with the miRCarta data set were slightly worse (SE = 0.8975, SP = 0.9019). Optimal performance was achieved by using a balanced ratio between positive and negative classes, with a slightly descending trend in the classifier performance when increasing the imbalance ratio (Figure 3A and B), a result that was also observed when analyzing the ROC and PR curves (Figure S1). When we compared the performance of the semi-supervised approach vs that of other supervised algorithms, the *miRNAss* algorithm [31] implemented in the eMIRNA pipeline outperformed the rest of supervised approaches, with the exception of lGBM, which showed similar performance results (Table 2). SP, as well as AUROC and AUPR estimates obtained with the *miRNAss* method [31] showed its high ability to discard false positives miRNA candidates, at the cost of a lower SE (Table 2). Additionally, after evaluating the ability of the algorithm to correctly identify the annotated porcine miRNA loci in all defined imbalance scenarios, a total of 399 (89.92%) and 213 (87.30%) annotated miRNAs were consistently classified as miRNA sequences using Ensembl (415) and miRCarta (244) positive databases, respectively.

The improved performance achieved with the Ensembl data set was expected because Ensembl annotation includes a more diverse and complete miRNA catalogue (415) than miRCarta (244). However, these differences are probably due to a more strict miRNA annotation procedure in the case of miRCarta database [2], which only includes manually curated bona fide miRNA genes. Nevertheless, the slight increase in overall performance observed in the Ensembl miRNA data set evidenced that even when reducing the set of positive sequences to a more stringent annotation, as that available in the miRCarta database [2], the ability of the eMIRNA pipeline to accurately distinguish miRNA sequences from other non-miRNA hairpins remained almost unaltered.

Besides, we determined the importance of the set of calculated features for classifying the miRNA candidates with the relief-F algorithm [48,49]. The estimated importance of the 30 most discriminant features is depicted in Figure 3C. The estimated impact of each feature on the accuracy of miRNA is shown in Table S4. Structural stability-related features accounted for the most important variables for classifying miRNAs correctly (MFEadj, EFEadj, MFE, EFE, MEAFE, MFEadj.GC and CFE). All of these parameters represented different hairpin structure folding statistics and they were highly intercorrelated (Figure 3D). The discriminant power of structural stability features is better exemplified in Figure 3E, where Ensembl annotated pre-miRNA sequences had an overall higher structural stability (i.e. lower MFEadj values) compared with that of other non-coding hairpin-like RNA sequences. These results clearly show the outmost importance of the structural folding configuration in order to discriminate true miRNA candidates from other hairpin-like sequences, hence supporting the need of a careful determination of pre-miRNA boundaries.

Novel porcine miRNA identified in the muscle transcriptome and by homology-based search

After microRNA identification from the porcine small RNA-seq data set, a total of 1,403 reconstructed pre-miRNA candidates from expressed transcripts were successfully identified as putative novel miRNAs in the porcine *gluteus medius* transcriptome, which corresponded to 160 unique miRNA loci after assigning clustered isomiRs to consensus single miRNA genes. Among these, 140 consensus candidates (87.5%) overlapped already annotated miRNAs in the porcine genome, whereas the 20 remaining ones (12.5%) were classified as novel miRNA candidates.

Regarding homology-based search miRNA discovery in the porcine assembly (Sscrofa11.1), a total of 310 annotated human miRNAs had orthologous miRNA genes in the porcine genome. The already annotated miRNAs in the porcine genome comprised 281 (90.64%) of the 310 homologous miRNAs detected with eMIRNA (File S3), and the 29 (N > 0.1) remaining candidates were classified as novel non-previously annotated homologous miRNAs in the porcine assembly (Table 3). The miR-483 and miR-484 genes were also identified as novel expressed miRNA candidates in the *gluteus medius* muscle transcriptome generated in our small RNA-seq experiment. A complete list of the novel miRNA candidates obtained with *de novo* and homology-based approaches is shown in Table 3. The full list of detected miRNAs that had been already annotated and all isomiRs associated with novel miRNA sequences can be found in File S3. The existence of multiple isoform candidates for single predicted miRNA loci, either displaying polymorphisms within the mature miRNA sequence or corresponding to 5′ or 3′-trimming variations (File S3), evidenced the wide variety of isomiR sequences expressed at significant levels in our *gluteus medius* muscle transcriptomic data set.

The eMIRNA pipeline accurately recalls miRNA loci

The same *gluteus medius* skeletal muscle transcriptomic data from the small RNA-seq experiment employed for de novo miRNA discovery with the eMIRNA pipeline was used for running the miRDeep2 algorithm [54]. A total of 148 transcripts belonging to 134 unique annotated miRNA loci were identified with miRDeep2. These numbers were slightly smaller than the 140 annotated porcine miRNAs recovered as expressed transcripts by the eMIRNA pipeline. Among these, 126 annotated miRNAs (85.14%) were consistently recovered with eMIRNA and miRDeep2, 14 (9.46%) were only reported by eMIRNA, and 8 (5.41%) were exclusively predicted by miRDeep2 (Table S5).

Regarding novel candidates, miRDeep2 was able to recover a total of 11 putative novel candidates belonging to 10 unique loci (Table S6). Seven of these candidates displayed an estimated probability of being a true positive miRNA above 19% (miRDeep2 score ≥ 4, Table S6). Noteworthy, two of the putatively true miRNAs detected by miRDeep2 spanned other previously annotated non-coding RNAs in the porcine assembly and were hence considered as miRNA-like false positives (Table S6). Among the 5 remaining candidates, 4 of them (miR-193a, miR-26a, miR-106b and miR-17) spanned other already annotated miRNAs in the porcine assembly and were thus wrongly classified as novel miRNAs by miRDeep2. The remaining candidate corresponded to miR-483, which had already been identified with the eMIRNA pipeline (Table 3, Table S6).

When comparing the accuracy of miRNA loci boundaries determined by the eMIRNA pipeline and miRDeep2, the eMIRNA approach demonstrated an overall better capability to accurately assign miRNA boundaries according to data from porcine miRNA loci annotated in the Ensembl database. A total of 103 out of 126 (81.74%) annotated miRNA genes detected by both eMIRNA and miRDeep2 showed reduced ΔD values (Table S7). This result implies that genomic positions of miRNA precursors predicted with the eMIRNA pipeline were more concordant with the annotation of the Sscrofa11.1 assembly than those predicted with miRDeep2. This outcome illustrates the effectiveness of motif search positional correction for reconstructing pre-miRNA candidates with a higher reliability than the fixed elongation patterns strategy used by miRDeep2 [54]. Three of the miRNA candidates showed no differences in positional accuracy between both approaches, while the positions of the remaining sequences (15.87%) were more accurately predicted with miRDeep2 (Table S7).

Experimental confirmation of the existence of three novel miRNAs in the muscle and liver tissues of Göttingen minipigs

The RT-qPCR analyses allowed us to detect the expression of the novel ssc-miR-483, ssc-miR-484 and ssc-miR-200a candidates in both *longissimus dorsi* skeletal muscle and liver tissues (Figure S2A and B) retrieved from Göttingen minipigs. Both ssc-miR-483 and ssc-miR-484 were also detected as consistently expressed in the skeletal muscle of Duroc gilts from our small RNA-seq experiment. The ssc-miR-200a was also detected in our generated data set but at very low expression levels. Nevertheless, its expression was further confirmed independently by RT-qPCR analyses. Amplification profiles and melting curves for the three novel miRNA candidates detected by RT-qPCR are shown in File S4.

Discussion

In the discovery of novel miRNA genes, one essential issue is the generation of pre-miRNA sequence candidates, given that the majority of miRNA prediction tools are based on feature extraction from the well-defined pre-miRNA hairpin structure [62]. At the cellular level, the most abundant and stable miRNA transcripts are the mature miRNA forms. Indeed, precursor stages, such as pri or pre-miRNAs, are much less abundant and have shorter half-lives than mature miRNAs [63,64]. Therefore, the accurate definition of pre-miRNA boundaries reconstructed from mature miRNAs is a crucial issue in order to predict folding structure and minimum free energy (MFE) estimates in a robust manner.

Noteworthy, the majority of state-of-the-art methods for miRNA prediction are solely focused on the miRNA classification of predefined candidate sequences. Moreover, many of them do not contemplate the generation of such candidates for the identification of unannotated miRNAs. On the contrary, they rely on well-known hairpins or on sets of manually curated candidate sequences that are embedded in their prediction pipelines [30,31,65-72].

Several other algorithms take advantage of the automated generation of hairpin candidates, adopting fixed defined elongation patterns in order to reconstruct pre-miRNA candidates from mature miRNA sequences [9,11,73,74]. However, fixed assumptions about elongation patterns do not take into consideration the expected variable length of pre-miRNA loci, and tend to generate candidate sequences that, despite harboring mature miRNAs, might have unreliable boundaries. This may lead to inaccuracies in the folding prediction and thus to an augmentation of the false negative rate. Even worse, non-miRNA hairpin-like sequences strongly resembling pre-miRNAs may be generated through the blind elongation of short sequences, which could result in the emergence of false positive candidates. This situation is particularly critical when analyzing the reliability of miRNA annotation in public databases [27,75,76]. Other approaches have also adopted a multiple hairpin candidate search for each query sequence to further select those showing a higher structural stability [77-79]. By using this strategy, we explored the influence of flanking processing motifs on the accurate determination of the length and boundaries of pre-miRNA candidates. By doing so, we have demonstrated that the inclusion of processing motif search criteria for the estimation of pre-miRNA boundaries resulted in an improved ability to better assess the optimal candidate sequences to be used for miRNA prediction.

Compared with miRDeep2 [54], the eMIRNA pipeline showed an improved ability to better assess the already annotated miRNA loci boundaries after pre-miRNA sequence reconstruction. However, the presence of embedded processing motifs within the boundaries of miRNA genes is not a universal feature, with a non-negligible amount of miRNA loci lacking the well-known CNNC and UG motifs [44], as well as the CHC and GHG mismatches [42] in their proximal surroundings. Additional work is needed to better characterize other processing motifs or structural determinants that may also contribute to miRNA maturation.

In contrast with pre-existing supervised methods for miRNA discovery, few semi-supervised methods have been developed for such purpose [31,80]. From a biological perspective, the scarce miRNA annotation typically found in non-model species poses a great challenge when attempting to predict novel miRNA loci uniquely based on labeled data. This happens because the amount of unknown non-miRNA sequences with hairpin-like secondary structures is expected to be hundreds of times larger than the number of confidently annotated miRNAs to be used for training supervised algorithms. Despite the fact that good performance statistics may be obtained after classifier training, supervised algorithms heavily depend on the existence of an extensive miRNA annotation. Indeed, the ability of such classifiers to detect unannotated miRNA sequences is mainly driven by the amount and diversity of positive and negative instances used for learning training.

On the contrary, semi-supervised transductive approaches [31] are able to overcome such limitation by incorporating unlabeled cases to the training process, with the aim of increasing the variability of the data used for target sequences classification. In fact, allowing the classifier to check hundreds or thousands of unknown unlabeled sequences has proven to increase the validity of microRNA prediction over other methods solely based on labeled data [31], a result that was also verified when comparing the semi-supervised approach used in this study with other broadly reported supervised methods (Table 2). This strategy is particularly reliable when few positive data are available and the annotated negative data set only represent a small proportion of the whole non-miRNA class. Besides, in classification problems where the negative class is expected to be dozens or hundreds of times larger than the positive class, the accurate identification of false positives is crucial. Indeed, such scenario is completely applicable to miRNAs, where thousands of non-miRNA sequences exist compared with the few hundreds of reliably annotated miRNA genes, and the annotation of negative hairpin-like sequences only represents a small proportion of the whole non-miRNA class.

After miRNA prediction, the detection of multiple isoforms for each single predicted miRNA loci evidenced the existence of a broad array of isomiR sequences expressed at significant levels in our *gluteus medius* muscle transcriptomic data set (File S3). Previous studies have highlighted the importance of isomiRs in expanding the biological diversity of miRNA function [81-84]. Like canonical miRNAs, isomiRs are also evolutionary conserved [81]. Both 5′ and 3′ miRNA isoforms can be generated either from alternative processing sites of DROSHA and Dicer [43,85] or from post-transcriptional modifications, influencing miRNA half-lives as well as their interactions with RNA-binding proteins (RBPs) [86,87].

More recently, other integrative approaches have addressed the detection of isomiRs and the potential functional influence that subtle modifications in the 3′ and 5′ boundaries of mature miRNA sequences might have on target recognition [88-91]. Other studies have also reported 5′ alternative processing events in a large number of miRNAs, contributing to the expansion of their target repertoire at a higher rate than previously thought [92]. Despite these promising results, the biological implications of miRNA alternative processing events leading to the generation of isomiRs are still poorly understood and further research is needed in order to exclude potential biases in isomiR quantification and functional validation, as variations in 3′ or 5′ ends of mature miRNAs can strongly affect the reliability of stem-loop qPCR amplification protocols [93].

One potential limitation of our study is that 17 of the novel miRNAs predicted with eMIRNA and based on muscle transcriptomic data have not been further investigated in order to confirm their existence by RT-qPCR, so their experimental validation is still pending. Indeed, we only investigated 3 out of 20 predicted novel porcine miRNAs. Noteworthy, the three selected miRNAs were successfully confirmed as bona fide miRNAs by RT-qPCR thus suggesting that eMIRNA predictions are accurate.

Among the three validated miRNAs, it is worth mentioning miR-483, which has been functionally associated with cell growth regulation [94] as well as with insulin resistance and metabolic syndrome susceptibility likely due to its strong implication in the regulation of glucose metabolism [95,96]. Additionally, the expression of miR-483, whose coding sequence maps to the second intron of the insulin growth factor 2 (*IGF2*) gene, has been tightly associated with an enhancement of *IGF2* gene expression. This is achieved through the binding of miR-483 to transcription factors in a positive feed-back loop [97], although other authors have questioned such dependence [98]. Other relevant successfully profiled miRNAs were ssc-miR-200a and ssc-miR-484. The miR-200a gene has been mainly reported as a regulator of cell growth and differentiation through targeting several protein-encoding transcripts like the growth factor receptor-bound 2 (*GRB2*), α-smooth muscle actin (*α-SMA*) or the fibroblast-specific protein-1 (*FSP-1*), thus hampering the endothelial-mesenchymal transition [99]. Furthermore, miR-484 has been associated with the inhibition of Fis1-mediated mitochondrial fission and apoptosis signaling [100].

Conclusions

In this study we have implemented an end-to-end pipeline that may facilitate the identification of novel miRNAs in the porcine genome. We have tested the eMIRNA pipeline by following a homology-based approach making use of the well annotated human microRNA transcriptome. Besides, we have analyzed the presence of non-annotated miRNAs in the porcine genome using data from a small RNA-seq experiment comprising muscle samples from 48 Duroc gilts. We have also taken into consideration several issues that are critical to robustly predict miRNA genes, such as the accurate reconstruction of candidate pre-miRNAs, the correct definition of negative training data sets and the evaluation of the high class-imbalance phenomenon, which is not fully addressed in many miRNA-prediction studies. In parallel, we have established hard-threshold filtering steps to keep false positive predictions at a minimum. We have also demonstrated the usefulness of positional refinement through flanking motif search to better determine the boundaries of pre-miRNA hairpin-like candidate sequences. The expression of several of the novel miRNAs described in this work was further confirmed by RT-qPCR analyses. In the light of these results, we believe that the eMIRNA pipeline will facilitate the discovery and annotation of novel miRNAs, thus broadening the miRNA catalogue of non-model species with yet poorly annotated genome assemblies.

Support

The research presented in this publication was funded by grants AGL2013-48742-C2-1-R and AGL2013-48742-C2-2-R awarded by the Spanish Ministry of Economy and Competitivity. E. Mármol-Sánchez was funded with a Ph.D. fellowship FPU15/01733 awarded by the Spanish Ministry of Education and Culture (MECD).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank the Department of Veterinary Animal Sciences in the Faculty of Health and Medical Sciences of the University of Copenhagen for providing their facilities and materials for RT-qPCR experiments. We would also like to express our gratitude to Dr. Caroline M. Junker Mentzel for kindly providing RNA samples for qPCR analyses. We also acknowledge Selección Batallé S.A. for providing animal material and the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016-2019 (SEV-2015-0533) grant awarded to the Center for

Research in Agricultural Genomics (CRAG). Thanks also to the CERCA Programme of the Generalitat the Catalunya for their support.

References

[1] A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function, Nucleic Acids Res. 47 (2019) D155–D162.

[2] C. Backes, T. Fehlmann, F. Kern, T. Kehl, H.-P. Lenhof, E. Meese, A. Keller, miRCarta: a central repository for collecting miRNA candidates, Nucleic Acids Res. 46 (2018) D160–D167.

[3] B. From, D. Domanska, L. Høye, V. Ovchinnikov, W. Kang, E. Aparicio-Puerta, M. Johansen, K. Flatmark, A. Mathelier, E. Hovig, M. Hackenberg, M.R. Friedländer, K.J. Peterson, MirGeneDB2.0: the metazoan microRNA complement, Nucleic Acids Res. (2019) gkz885.

[4] J. Meunier, F. Lemoine, M. Soumillon, A. Liechti, M. Weier, K. Guschanski, H. Hu, P. Khaitovich, H. Kaessmann, Birth and expression evolution of mammalian microRNA genes, Genome Res. 23 (2013) 34–45.

[5] M. Warnefors, A. Liechti, J. Halbert, D. Valloton, H. Kaessmann, Conserved microRNA editing in mammalian evolution, development and disease, Genome Biol. 15 (2014) R83.

[6] L.P. Lim, N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, D.P. Bartel, The microRNAs of Caenorhabditis elegans, Genes Dev. 17 (2003) 991–1008.

[7] E.C. Lai, P. Tomancak, R.W. Williams, G.M. Rubin, Computational identification of Drosophila microRNA genes, Genome Biol. 4 (2003) R42.

[8] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang, Y. Li, MicroRNA identification based on sequence and structure alignment, Bioinformatics. 21 (2005) 3610–3614.

[9] A. Mathelier, A. Carbone, MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data, Bioinformatics. 26 (2010) 2226–2234.

[10] K. Qian, E. Auvinen, D. Greco, P. Auvinen, miRSeqNovel: An R based workflow for analyzing miRNA sequencing data, Mol. Cell. Probes 26 (2012) 208–211.

[11] J. An, J. Lai, M.L. Lehman, C.C. Nelson, MiRDeep*: An integrated application tool for miRNA identification from RNA sequencing data, Nucleic Acids Res. 41 (2013) 727–737.

[12] T.B. Hansen, M.T. Venø, J. Kjems, C.K. Damgaard, miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs, Nucleic Acids Res. 42 (2014) e124.

[13] D. Kleftogiannis, A. Korfiati, K. Theofilatos, S. Likothanassis, A. Tsakalidis, S. Mavroudi, Where we stand, where we are moving: surveying computational techniques for identifying miRNA genes and uncovering their regulatory role, J. Biomed. Inform. 46 (2013) 563–573.

[14] M. Bortolomeazzi, E. Gaffo, S. Bortoluzzi, A survey of software tools for microRNA discovery and characterization using RNA-seq, Brief. Bioinform. 20 (2017) 918–930.

[15] G. Stegmayer, L.E. Di Persia, M. Rubiolo, M. Gerard, M. Pividori, C. Yones, L.A. Bugnon, T. Rodriguez, J. Raad, D.H. Milone, Predicting novel microRNA: a

comprehensive comparison of machine learning approaches, Brief. Bioinform. (2018) bby037.

[16] A. Rajendiran, A. Chatterjee, A. Pan, Computational approaches and related tools to identify microRNAs in a species: a bird's eye view, Interdiscip. Sci. Comput. Life Sci. 10 (2018) 616–635.

[17] J.-E. Long, H.-X. Chen, Identification and characteristics of cattle microRNAs by homology searching and small RNA cloning, Biochem. Genet. 47 (2009) 329–343.

[18] Z. Wang, K. He, Q. Wang, Y. Yang, Y. Pan, The prediction of the porcine premicroRNAs in genome-wide based on support vector machine (SVM) and homology searching, BMC Genomics 13 (2012) 729.

[19] X. Hou, Z. Tang, H. Liu, N. Wang, H. Ju, K. Li, Discovery of microRNAs associated with myogenesis by deep sequencing of serial developmental skeletal muscles in pigs, PLoS One 7 (2012) e52123.

[20] C. Yuan, X. Wang, R. Geng, X. He, L. Qu, Y. Chen, Discovery of cashmere goat (Capra hircus) microRNAs in skin and hair follicles by Solexa sequencing, BMC Genomics 14 (2013) 511.

[21] J. Sun, M. Li, Z. Li, J. Xue, X. Lan, C. Zhang, C. Lei, H. Chen, Identification and profiling of conserved and novel microRNAs from Chinese Qinchuan bovine longissimus thoracis, BMC Genomics 14 (2013) 42.

[22] T. Buza, M. Arick, H. Wang, D.G. Peterson, Computational prediction of disease microRNAs in domestic animals, BMC Res. Notes. 7 (2014) 403.

[23] B. Sadeghi, H. Ahmadi, S. Azimzadeh-Jamalkandi, M.R. Nassiri, A. Masoudi-Nejad, BosFinder: a novel pre-microRNA gene prediction algorithm in Bos taurus, Anim. Genet. 45 (2014) 479–484.

[24] J. Wu, H. Zhu, W. Song, M. Li, C. Liu, N. Li, F. Tang, H. Mu, M. Liao, X. Li, W. Guan, X. Li, J. Hua, Identification of conservative microRNAs in Saanen dairy goat testis through deep sequencing, Reprod. Domest. Anim. 49 (2014) 32–40.

[25] Z. Li, H. Wang, L. Chen, L. Wang, X. Liu, C. Ru, A. Song, Identification and characterization of novel and differentially expressed microRNAs in peripheral blood from healthy and mastitis Holstein cattle by deep sequencing, Anim. Genet. 45 (2014) 20–27.

[26] D.M.D. Saçar, H. Hamzeiy, J. Allmer, Can miRBase provide positive data for machine learning for the detection of miRNA hairpins? J. Integr. Bioinform. 10 (2013) 1–11.

[27] N. Ludwig, M. Becker, T. Schumann, T. Speer, T. Fehlmann, A. Keller, E. Meese, Bias in recent miRBase annotations potentially associated with RNA quality issues, Sci. Rep. 7 (2017) 5162.

[28] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, Q. Zou, Improved and promising identification of human microRNAs by incorporating a high-quality negative set, IEEE/ACM Trans. Comput. Biol. Bioinforma. 11 (2014) 192–201.

[29] M. Yousef, J. Allmer, W. Khalifa, Accurate plant microRNA prediction can be achieved using sequence motif features, J. Intell. Learn. Syst. Appl. 8 (2016) 9–22.

[30] G. Stegmayer, C. Yones, L. Kamenetzky, D.H. Milone, High class-imbalance in premiRNA prediction: a novel approach based on deepSOM, IEEE/ACM Trans. Comput. Biol. Bioinforma. 14 (2017) 1316–1326.

[31] C. Yones, G. Stegmayer, D.H. Milone, C. Sahinalp, Genome-wide pre-miRNA discovery from few labeled examples, Bioinformatics. 34 (2018) 541–549.

[32] Y. Wang, X. Li, B. Tao, Improving classification of mature microRNA by solving class imbalance problem, Sci. Rep. 6 (2016) 25941.

[33] T.F. Cardoso, R. Quintanilla, J. Tibau, M. Gil, E. Mármol-Sánchez, O. GonzálezRodríguez, R. González-Prendes, M. Amills, Nutrient supply affects the mRNA expression profile of the porcine skeletal muscle, BMC Genomics 18 (2017) 603.

[34] M. Ballester, M. Amills, O. González-Rodríguez, T.F. Cardoso, M. Pascual, R. González-Prendes, N. Panella-Riera, I. Díaz, J. Tibau, R. Quintanilla, Role of AMPK signalling pathway during compensatory growth in pigs, BMC Genomics 19 (2018) 682.

[35] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics. 26 (2010) 841–842.

[36] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT suite: a web server for clustering and comparing biological sequences, Bioinformatics. 26 (2010) 680–682.

[37] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA Package 2.0, Algorithms Mol. Biol. 6 (2011) 26.

[38] C. Yones, HextractoR: Integrated tool for hairpin extraction of RNA sequences, R Package Version 1.3, 2018 https://cran.r-project.org/package=HextractoR.

[39] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal. 17 (2011) 10.

[40] Y. Lu, A.S. Baras, M.K. Halushka, miRge 2.0 for comprehensive analysis of microRNA sequencing data, BMC Bioinforma. 19 (2018) 275.

[41] B. Langmead, C. Trapnell, M. Pop, S. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.

[42] W. Fang, D.P. Bartel, The menu of features that define primary microRNAs and enable de novo design of microRNA genes, Mol. Cell 60 (2015) 131–145.

[43] C. Roden, J. Gaillard, S. Kanoria, W. Rennie, S. Barish, J. Cheng, W. Pan, J. Liu, C. Cotsapas, Y. Ding, J. Lu, Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation, Genome Res. 27 (2017) 374–384.

[44] V.C. Auyeung, I. Ulitsky, S.E. McGeary, D.P. Bartel, Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing, Cell. 152 (2013) 844–858.

[45] E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, Bioinformatics. 20 (2004) 2911–2917.

[46] M. Jiang, J. Anderson, J. Gillespie, M. Mayne, uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts, BMC Bioinforma. 9 (2008) 192.

[47] I. Lopes, A. Schliep, A.C.L.F. de Carvalho, The discriminant power of RNA features for pre-miRNA recognition, BMC Bioinforma. 15 (2014) 124.

[48] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Appl. Intell. 7 (1997) 39–55.

[49] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23–69, https://doi.org/10.1023/ A:1025667309714.

[50] R. Batuwita, V. Palade, Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics data sets learning, J. Bioinforma. Comput. Biol. 10 (2012) 1250003.

[51] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, ACM Int. Conf. Proceeding Ser. (2006) 233–240.

[52] G.S. França, M.D. Vibranovski, P.A.F. Galante, Host gene constraints and genomic context impact the expression and evolution of human microRNAs, Nat. Commun. 7 (2016) 11438.

[53] M.R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, N. Rajewsky, Discovering microRNAs from deep sequencing data using miRDeep, Nat. Biotechnol. 26 (2008) 407–415.

[54] M.R. Friedländer, S.D. MacKowiak, N. Li, W. Chen, N. Rajewsky, MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, Nucleic Acids Res. 40 (2012) 37–52.

[55] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.

[56] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[57] C.M.J. Mentzel, C. Anthon, M.J. Jacobsen, P. Karlskov-Mortensen, C.S. Bruun, C.B. Jørgensen, J. Gorodkin, S. Cirera, M. Fredholm, Gender and obesity specific microRNA expression in adipose tissue from lean and obese pigs, PLoS One 10 (2015) e0131650.

[58] I. Balcells, S. Cirera, P.K. Busk, Specific and sensitive quantitative RT-PCR of miRNAs with DNA primers, BMC Biotechnol. 11 (2011) 70.

[59] P.K. Busk, A tool for design of primers for microRNA-specific quantitative RT-qPCR, BMC Bioinforma. 15 (2014) 29.

[60] S. Cirera, P.K. Busk, Quantification of miRNAs by a simple and specific qPCR method, Methods Mol. Biol. (2014) 73–81.

[61] K. Kim, T. Duc Nguyen, S. Li, T. Anh Nguyen, SRSF3 recruits DROSHA to the basal junction of primary microRNAs, RNA. 24 (2018) 892–898.

[62] D.P. Bartel, Metazoan microRNAs, Cell. 173 (2018) 20–51.

[63] L. Gan, B. Denecke, Profiling pre-microRNA and mature microRNA expressions using a single microarray and avoiding separate sample preparation, Microarrays. 2 (2013) 24–33.

[64] Y. Guo, J. Liu, S.J. Elfenbein, Y. Ma, M. Zhong, C. Qiu, Y. Ding, J. Lu, Characterization of the mammalian miRNA turnover landscape, Nucleic Acids Res. 43 (2015) 2326–2341.

[65] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinforma. 6 (2005) 310, https://doi.org/10.1186/1471-2105-6- 310.

[66] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, Z. Lu, MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features, Nucleic Acids Res. 35 (2007) W339–W344.

[67] R. Batuwita, V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction, Bioinformatics. 25 (2009) 989–995.

[68] Y. Wu, B. Wei, H. Liu, T. Li, S. Rayner, MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences, BMC Bioinforma. 12 (2011) 107.

[69] A. Gudyś, M.W. Szcześniak, M. Sikora, I. Makałowska, HuntMi: an efficient and taxon-specific approach in pre-miRNA identification, BMC Bioinforma. 14 (2013) 83.

[70] Q. Zou, Y. Mao, L. Hu, Y. Wu, Z. Ji, miRClassify: an advanced web server for miRNA family classification and annotation, Comput. Biol. Med. 45 (2014) 157–160.

[71] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, S. Mavroudi, YamiPred: a novel evolutionary method for predicting pre-miRNAs and selecting relevant features, IEEE/ACM Trans. Comput. Biol. Bioinforma. 12 (2015) 1183–1192.

[72] D.M.D. Saçar, J. Baumbach, J. Allmer, On the performance of pre-microRNA detection algorithms, Nat. Commun. 8 (2017) 330.

[73] D.M. Vitsios, E. Kentepozidou, L. Quintais, E. Benito-Gutiérrez, S. van Dongen, M.P. Davis, A.J. Enright, Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests, Nucleic Acids Res. 45 (2017) e177.

[74] R.J. Peace, M. Sheikh Hassani, J.R. Green, miPIE: NGS-based prediction of miRNA using integrated evidence, Sci. Rep. 9 (2019) 1548.

[75] M.J. Axtell, B.C. Meyers, Revisiting criteria for plant microRNA annotation in the era of big data, Plant Cell 30 (2018) 272–284.

[76] J. Alles, T. Fehlmann, U. Fischer, C. Backes, V. Galata, M. Minet, M. Hart, M. AbuHalima, F.A. Grässer, H.-P. Lenhof, A. Keller, E. Meese, An estimate of the total number of true human miRNAs, Nucleic Acids Res. 47 (2019) 3353–3364.

[77] J. Lei, Y. Sun, miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-seq data, Bioinformatics. 30 (2014) 2837–2839.

[78] M. Evers, M. Huttner, A. Dueck, G. Meister, J.C. Engelmann, miRA: adaptable novel miRNA identification in plants using small RNA sequencing data, BMC Bioinforma. 16 (2015) 370, https://doi.org/10.1186/s12859-015-0798-3.

[79] C. Paicu, I. Mohorianu, M. Stocks, P. Xu, A. Coince, M. Billmeier, T. Dalmay, V. Moulton, S. Moxon, miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing data sets, Bioinformatics. 33 (2017) 2446–2454.

[80] M. Sheikh Hassani, J.R. Green, Multi-view co-training for microRNA prediction, Sci. Rep. 9 (2019) 10931.

[81] G.C. Tan, E. Chan, A. Molnar, R. Sarkar, D. Alexieva, I.M. Isa, S. Robinson, S. Zhang, P. Ellis, C.F. Langford, P.V. Guillot, A. Chandrashekran, N.M. Fisk, L. Castellano, G. Meister, R.M. Winston, W. Cui, D. Baulcombe, N.J. Dibb, 5′ isomiR variation is of functional and evolutionary importance, Nucleic Acids Res. 42 (2014) 9424–9435.

[82] A.G. Telonis, P. Loher, Y. Jing, E. Londin, I. Rigoutsos, Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity, Nucleic Acids Res. 43 (2015) 9158–9175.

[83] F. Yu, K.A. Pillman, C.T. Neilsen, J. Toubia, D.M. Lawrence, A. Tsykin, M.P. Gantier, D.F. Callen, G.J. Goodall, C.P. Bracken, Naturally existing isoforms of miR-222 have distinct functions, Nucleic Acids Res. 45 (2017) 11371–11385.

[84] P. Sheng, C. Fields, K. Aadland, T. Wei, O. Kolaczkowski, T. Gu, B. Kolaczkowski, M. Xie, Dicer cleaves 5′-extended microRNA precursors originating from RNA polymerase II transcription start sites, Nucleic Acids Res. 46 (2018) 5737–5752.

[85] B. Kim, K. Jeong, V.N. Kim, Genome-wide mapping of DROSHA cleavage sites on primary microRNAs and noncanonical substrates, Mol. Cell 66 (2017) 258–269.e5.

[86] C.T. Neilsen, G.J. Goodall, C.P. Bracken, IsomiRs – the overlooked repertoire in the dynamic microRNAome, Trends Genet. 28 (2012) 544–549.

[87] X. Bofill-De Ros, A. Yang, S. Gu, IsomiRs: expanding the miRNA repression toolbox beyond the seed, Biochim. Biophys. Acta - Gene Regul. Mech. (2019) 194373.

[88] G. Urgese, G. Paciello, A. Acquaviva, E. Ficarra, isomiR-SEA: an RNA-seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation, BMC Bioinforma. 17 (2016) 148.

[89] Y. Zhang, Q. Zang, B. Xu, W. Zheng, R. Ban, H. Zhang, Y. Yang, Q. Hao, F. Iqbal, A. Li, Q. Shi, IsomiR Bank: a research resource for tracking IsomiRs, Bioinformatics. 32 (2016) 2069–2071.

[90] X. Bofill-De Ros, K. Chen, S. Chen, N. Tesic, D. Randjelovic, N. Skundric, S. Nesic, V. Varjacic, E.H. Williams, R. Malhotra, M. Jiang, S. Gu, QuagmiR: a cloud-based application for isomiR big data analytics, Bioinformatics. 35 (2019) 1576–1578.

[91] X. Bofill-De Ros, W.K. Kasprzak, Y. Bhandari, L. Fan, Q. Cavanaugh, M. Jiang, L. Dai, A. Yang, T.-J. Shao, B.A. Shapiro, Y.-X. Wang, S. Gu, Structural differences between pri-miRNA paralogs promote alternative Drosha cleavage and expand target repertoires, Cell Rep. 26 (2019) 447–459.e4.

[92] H. Kim, J. Kim, K. Kim, H. Chang, K. You, V.N. Kim, Bias-minimized quantification of microRNA reveals widespread alternative processing and 3′ end modification, Nucleic Acids Res. 47 (2019) 2630–2640.

[93] A. Schamberger, T.I. Orbán, 3’ IsomiR species and DNA contamination influence reliable quantification of microRNAs by stem-loop quantitative PCR, PLoS One 9 (2014) e106315.

[94] T.H. Vu, N.V. Chuyen, T. Li, A.R. Hoffman, M. Blick, F. Fornari, N. Zanesi, H. Alder, G. D'Elia, L. Gramantieri, L. Bolondi, G. Lanza, P. Querzoli, A. Angioni, C.M. Croce, M. Negrini, Loss of imprinting of IGF2 sense and antisense transcripts in Wilms' tumor, Cancer Res. 63 (2003) 1900–1905.

[95] D. Ferland-McCollough, D.S. Fernandez-Twinn, I.G. Cannell, H. David, M. Warner, A.A. Vaag, J. Bork-Jensen, C. Brøns, T.W. Gant, A.E. Willis, K. Siddle, M. Bushell, S.E. Ozanne, Programming of adipose tissue miR-483-3p and GDF-3 expression by maternal diet in type 2 diabetes, Cell Death Differ. 19 (2012) 1003–1012.

[96] F. Pepe, S. Pagotto, S. Soliman, C. Rossi, P. Lanuti, C. Braconi, R. MarianiCostantini, R. Visone, A. Veronese, Regulation of miR-483-3p by the O-linked N-acetylglucosamine transferase links chemosensitivity to glucose metabolism in liver cancer cells, Oncogenesis. 6 (2017) e328.

[97] M. Liu, A. Roth, M. Yu, R. Morris, F. Bersani, M.N. Rivera, J. Lu, T. Shioda, S. Vasudevan, S. Ramaswamy, S. Maheswaran, S. Diederichs, D.A. Haber, The IGF2 intronic miR-483 selectively enhances transcription from IGF2 fetal promoters and enhances tumorigenesis, Genes Dev. 27 (2013) 2543–2548.

[98] A. Veronese, L. Lupini, J. Consiglio, R. Visone, M. Ferracin, F. Fornari, N. Zanesi, H. Alder, G. D'Elia, L. Gramantieri, L. Bolondi, G. Lanza, P. Querzoli, A. Angioni, C.M. Croce, M. Negrini, Oncogenic role of miR-483-3p at the IGF2/483 locus, Cancer Res. 70 (2010) 3140–3149.

[99] H. Zhang, J. Hu, L. Liu, MiR-200a modulates TGF- β 1-induced endothelial-to-mesenchymal shift via suppression of GRB2 in HAECs, Biomed. Pharmacother. 95 (2017) 215–222.

[100] K. Wang, B. Long, J.-Q. Jiao, J.-X. Wang, J.-P. Liu, Q. Li, P.-F. Li, miR-484 regulates mitochondrial network through targeting Fis1, Nat. Commun. 3 (2012) 781.

Figures

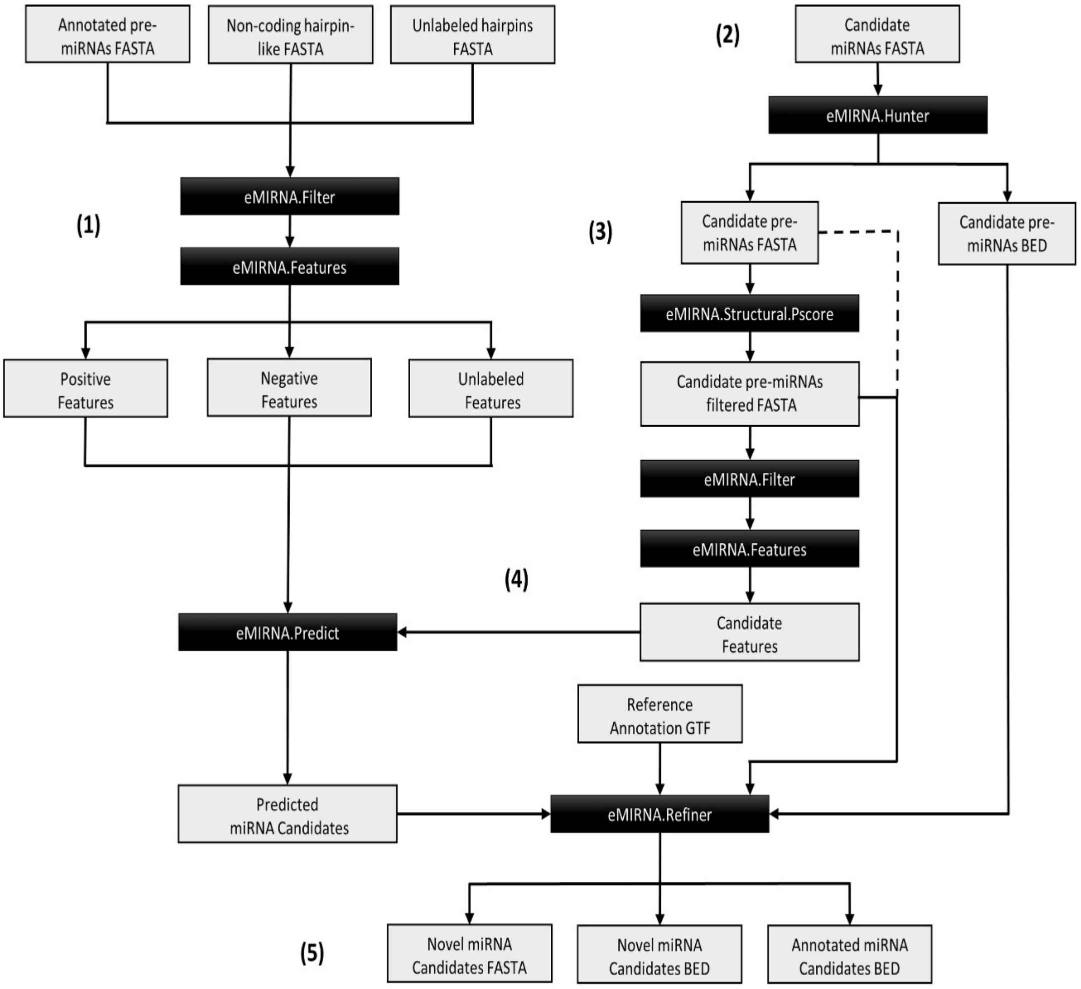

Figure 1: eMIRNA pipeline scheme for homology-based miRNA prediction using data from closely related species and *de novo* miRNA prediction from small RNA-seq data. (**1**) Positive, negative and unlabeled data are filtered based on size and secondary folding structure and a set of features is extracted for each sequence. (**2**) Mature miRNA sequences from small RNA-seq data or related model species are mapped against the selected genome assembly and elongated to reconstruct putative pre-miRNA candidates. (**3**) Candidate precursors are filtered based on size and secondary folding structure and a set of features is extracted for each candidate sequence. Optionally, sequences showing unstable secondary structure are removed. (**4**) Candidate sequences are embedded in the semi-supervised transductive classifier and a list of putative miRNAs is predicted. (**5**)

Predicted miRNAs are either assigned to already annotated miRNA loci in the provided reference assembly or classified as putative novel miRNAs genes.

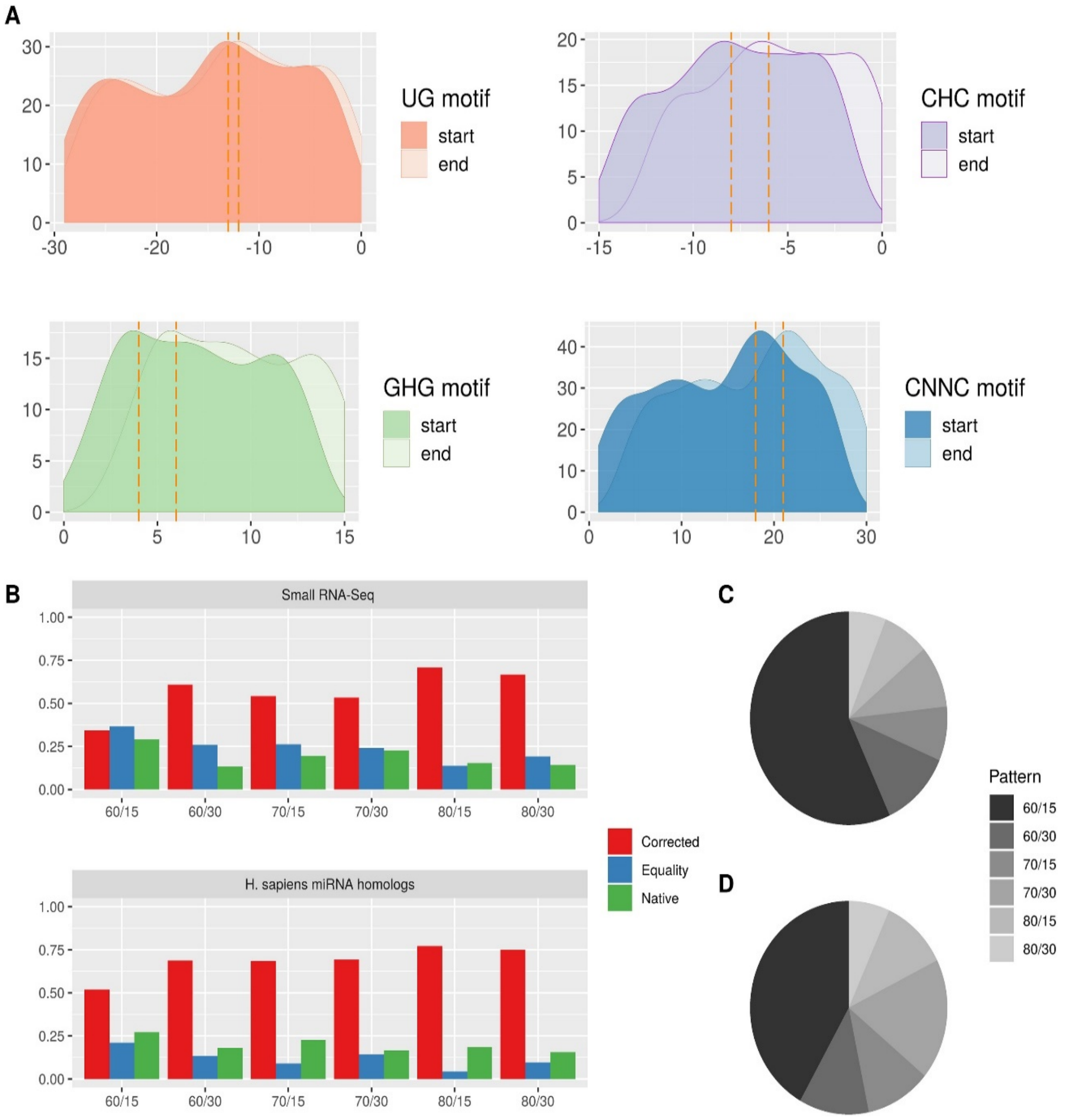

Figure 2: Processing motifs distribution and structural stability metrics. (**A**) Positional distribution of upstream and downstream motifs across annotated pre-miRNA boundaries in the porcine genome. (**B**) Proportion of candidate sequences for each elongation pattern showing the most stable folding structure according to randfold p score. The proportion of sequences for which the structural stability was higher in motif corrected candidates or, conversely, in non-corrected (native) candidates are shown as red and green bars, respectively. The proportion of sequences for which the structural stability was equivalent between motif corrected and native candidates were labeled as equally stable (blue). (**C**) Proportion of selected pre-miRNA candidates detected in the porcine *gluteus medius* muscle small RNA-seq data and (**D**) Proportion of selected pre-miRNA candidates detected through a *H. sapiens* homology-based miRNA search strategy, according to the most structurally stable elongation pattern tested. If two or more pre-miRNA sequences showed equivalent stability, the shortest motif-corrected candidate was selected.

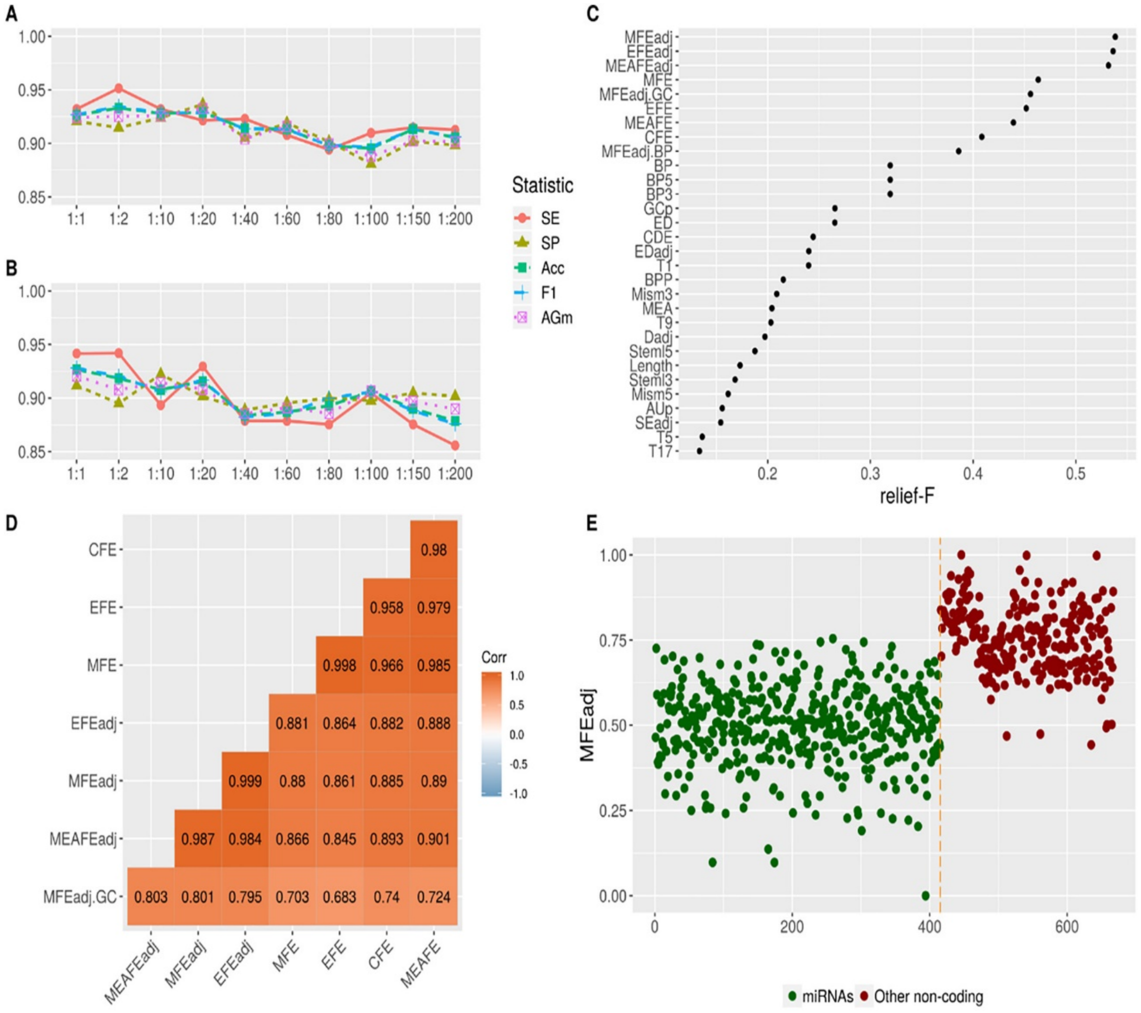

Figure 3: Classification performance and feature importance statistics. Performance metrics for Sensitivity (SE), Specificity (SP), Accuracy (Acc), F1-score (F1) and Adjusted Geometric-mean (Agm) across incremental imbalance-ratios by using positive miRNAs from (**A**) Ensembl and (**B**) miRCarta databases. (**C**) Thirty most discriminant features according to the relief-F algorithm. (**D**) Pearson's correlation coefficient among the seven most discriminant features associated with secondary structure stability metrics. (**E**) Comparison of the folding structure stability between annotated miRNAs and other hairpin-like non-coding RNA sequences present in the porcine genome. Stability is expressed as the scaled Minimum Free Energy of the folded hairpins adjusted by sequence length (MFEadj).

Tables

Table 1: List of calculated features extracted from candidate hairpins.

Sequence Features	**Symbol**	**Number of variables**
Triplet Elements by SVM-Triplet	T1 … T32	32
Sequence Length	Length	1
G+C/Length	GC	1
A+U/G+C	AU.GCr	1
A, U, G, C/Length	Ar, Ur, Gr, Cr	4
Dinucleotide/Length	Aar, GGr, CCr …	16
Secondary Structure metrics	**Symbol**	**Number of variables**
Hairpin loop Length	Hl	1
5' and 3' Stems Length	Steml5, Steml3	2
Basepairs in Secondary Structure	BP	1
Matches in 5' and 3' Stems	BP5, BP3	2
Mismatches in 5' and 3' Stems	Mism5, Mism3	2
Bulges in 5' and 3' Stems	B5, B3	2
Bulges in 5' and 3' Stems of types 1 to 7 mismatch	BN1.5, BN1.3 …	14
A-U, G-C and G-U basepairs	Aup, GCp, Gup	3
Structural Statistics	**Symbol**	**Number of variables**
Minimum Free Energy	MFE	1
Ensemble and Centroid Free Energy	EFE, CFE	2
Centroid Distance to Ensemble	CDE	1
Maximum Expected Accuracy	MEA, MEAFE	2
BP/Length	BPP	1
MFE Ensemble Frequency	Efreq	1

Ensemble Diversity	ED	1
MFE/Length, EFE/Length and CDE/Length	MFEadj, EFEadj, Dadj	3
Shannon Entropy/Length	Seadj	1
MFE-EFE/Length	DiffMFE.EFE	1
MFEadj/GC and MFEadj/BP	MFEadj.GC, MFEadj.BP	2
MEAFE/Length and ED/Length	MEAFEadj, Edadj	2

Table 2: Comparative benchmarking between the semi-supervised transductive learning approach employed by the *miRNAss* algorithm and other state-of-the-art supervised algorithms (i.e. SVM: Support Vector Machine, RF: Random Forest, KNN: k-Nearest Neighbors, NB: Naïve Bayes, XGB: Extreme Gradient Boosting and lGBM: light Gradient Boosting Tree) for miRNA classification. Only labeled positive and negative data sets were used for training.

SE: Sensitivity; SP: Specificity; F-1: F-score measure of the harmonic mean of the precision and recall; AUROC: Area under the Receiver Operating Characteristics (ROC) curve; AUPR: Area under the Precision-Recall curve.

Statistic	SVM	RF	KNN	NB	XGB	lGBM	miRNAss
SE	0.932	0.932	0.9223	0.9126	0.9515	0.9223	0.8835
SP	0.8413	0.9524	0.9524	0.9683	0.9365	0.9048	0.9683
F-1	0.9187	0.9505	0.9453	0.9447	0.9561	0.9314	0.9226
AUROC	0.6428	0.7246	0.5757	0.4291	0.7063	0.9781	0.9783
AUPR	0.7222	0.8489	0.6751	0.5818	0.8509	0.9873	0.987

Table 3: Novel porcine miRNA genes predicted through a homology-based comparison with human miRNA annotation and on the basis of data generated by sequencing small RNAs expressed in the *gluteus medius* muscle of Duroc pigs.

Chr: Chromosome; N: Neighborhood score.

Chr	Start	End	Strand	ID	N
1	191218572	191218651	+	miR-3529	0.33
1	268816970	268817050	+	miR-219b	0.92
2	32718	32792	+	miR-6743	0.82
2	1473428	1473495	-	miR-483	0.84
2	1474436	1474513	-	3229-4643	-
2	40104336	40104403	-	1325-14520	-
2	134660802	134660897	-	1323-14559	-
3	7180536	7180603	-	miR-484	0.1
3	40421320	40421409	+	427-63874	-
3	40772345	40772445	+	176-178526	-
4	22195784	22195880	+	2340-6855	-
5	3397056	3397130	-	1111-18619	-
5	17410008	17410122	+	1794-9841	-
5	95548384	95548458	+	miR-3059	1
6	56426941	564267012	-	miR-520e	0.3
6	63490755	63490822	+	miR-200a	0.6
8	1205684	1205760	-	miR-4800	0.85
9	52087075	52087155	+	1864-9314	-
9	114528009	114528076	+	miR-3120	0.7
10	27079413	27079489	-	miR-24-1	0.79
11	1824995	1825062	+	504-51258	-
11	49808356	49808431	-	miR-3665	0.86
12	1538011	1538119	+	337-84973	-

12	1601453	1601506	-	miR-3065	0.82
12	18989584	18989651	+	399-69074	-
12	45088806	45088863	+	miR-451b	0.78
12	45597382	45597459	+	miR-4523	0.81
12	46211527	46211594	-	miR-3184	0.61
12	48162620	48162704	-	miR-132	0.84
12	56201226	56201300	-	518-49963	-
13	30242047	30242114	+	772-29980	-
13	33152284	33152383	+	miR-4787	0.83
13	197168804	197168901	+	miR-6501	0.97
14	87673881	87673954	+	3552-4147	-
14	109233945	109234032	-	miR-3085	0.95
14	122706280	122706361	+	miR-6715a	0.96
14	122706285	122706353	-	miR-6715b	0.96
14	127016706	127016794	-	miR-9851	0.83
14	140979533	140979627	+	3525-4198	-
15	128165751	128165827	-	miR-5702	0.86
17	61915309	61915376	+	1544-12001	-
X	41793240	41793315	+	451-58980	-
X	43716471	43716538	+	miR-502	0.73
X	59551153	59551220	+	miR-374c	0.8
X	94122543	94122610	+	miR-1264	0.83
X	96979691	96979765	+	miR-1277	0.68
X	124724889	124724956	-	miR-718	0.89

Supplementary Materials

Figure S1: (A) Receiver Operating Characteristics (ROC) and **(B)** Precision-Recall (PR) curves computed for each pre-defined imbalance scenario using porcine Ensembl annotation for positive (miRNAs) and negative (other hairpin-like non-coding RNAs) data sets.

Figure S2: RT-qPCR results of selected novel miRNAs. Successfully profiled novel miRNAs in **(A)** the *longissimus dorsi* skeletal muscle and **(B)** liver tissues from 7 Göttingen minipigs.

File S1: FASTA file of collapsed expressed sequences (RPM > 10) used in the *de novo* discovery of miRNAs expressed in the porcine *gluteus medius* skeletal muscle.

File S2: Non-redundant annotated mature miRNA sequences obtained from the *H. sapiens* GRCh38.p12 genome assembly used as a reference in the homology-based search of novel miRNAs in the current release of the porcine genome (Sscrofa11.1).

File S3: List of already annotated miRNAs and all isomiRs detected as expressed (RPM > 10) in the porcine *gluteus medius* skeletal muscle.

File S4: Amplification profiles and melting curves for the three novel miRNA candidates subjected to confirmation by RT-qPCR analyses.

Table S1: Area under the curve (AUC) computed for each pre-defined imbalance scenario using Ensembl annotation for positive and negative data sets.

Table S2: True positive ratio of porcine miRNA loci annotated in the Ensembl and miRCarta databases and identified by the eMIRNA pipeline in all considered imbalance scenarios.

Table S3: Mature miRNAs and primers used for RT-qPCR confirmation of selected novel miRNA candidates.

Table S4: Feature importance according to the relief-F algorithm.

Table S5: Previously annotated miRNAs genes that are correctly classified as miRNAs by eMIRNA and miRDeep2.

Table S6: miRDeep2 algorithm results for miRNA prediction using the *gluteus medius* muscle small RNA-seq data generated in the present study.

Table S7: Deviation rates (dr) and Differential deviation (ΔD) estimates for miRNA genomic positional prediction with eMIRNA and miRDeep2.