BMC Genomics

RESEARCH ARTICLE Open Access

The prediction of the porcine pre-microRNAs in genome-wide based on support vector machine (SVM) and homology searching

Zhen Wang[1,2], Kan He[1,2,3], Qishan Wang[1,2], Yumei Yang[1,2] and Yuchun Pan[1,2*]

Abstract

Background: MicroRNAs (miRNAs) are a class of small non-coding RNAs that regulate gene expression by targeting mRNAs for translation repression or mRNA degradation. Although many miRNAs have been discovered and studied in human and mouse, few studies focused on porcine miRNAs, especially in genome wide.

Results: Here, we adopted computational approaches including support vector machine (SVM) and homology searching to make a global scanning on the pre-miRNAs of pigs. In our study, we built the SVM-based porcine pre-miRNAs classifier with a sensitivity of 100%, a specificity of 91.2% and a total prediction accuracy of 95.6%, respectively. Moreover, 2204 novel porcine pre-miRNA candidates were found by using SVM-based pre-miRNAs classifier. Besides, 116 porcine pre-miRNA candidates were detected by homology searching.

Conclusions: We identified the porcine pre-miRNA in genome-wide through computational approaches by utilizing the data sets of pigs and set up the porcine pre-miRNAs library which may provide us a global scanning on the pre-miRNAs of pigs in genome level and would benefit subsequent experimental research on porcine miRNA functional and expression analysis.

Keywords: Porcine, Pre-miRNA, SVM, Homology searching

Background

MicroRNAs (miRNAs) are a family of ~22nt endogenous non-coding RNAs [1,2]. Mature miRNAs are usually cleaved from ~90nt miRNA precursors (pre-miRNAs) which are derived from processing of a long primary miRNA (pri-miRNA) by a ribonuclease [3]. Increasing evidences have shown that miRNAs play fundamentally important roles in various biological processes, including cell proliferation [4-7], development timing [8,9], apoptosis [10,11], carcinogenesis [12-14], and response to different environmental stresses containing disease [15-17].

Since the first lin-4 miRNA of *C. elegans* was discovered in 1992 [18], more than 19000 miRNAs have been found in animals and plants. Currently, the miRNA Registry Database (Release 17, April 2011; http://mirbase.org), a comprehensive and searchable database of published miRNA sequences, contains 16772 entries representing hairpin pre-miRNAs, expressing 19724 mature miRNA products, in 153 species [19]. However, only 228 pre-miRNAs of pigs are included in this database, the number is far less than it really has.

Pre-miRNAs have similar hairpin-shaped stem loop structure, high minimal folding free energy index, and high evolutionary conservation. They become the important features which could be used in the computational identification of pre-miRNA [20-22]. To date, computational prediction has been broadly used to identify potential pre-miRNAs in animals and plants [23-25], because it is not limited by tissue specificity and time of miRNA expression. Especially, machine learning approaches such as random forest (RF) [26], naïve Bayes classifier [27], hidden Markov model [28,29] and SVM [30-32] have been adopted.

Although previous studies have identified a certain number of porcine pre-miRNAs, few researches in computational identification of pre-miRNAs based on the

* Correspondence: panyuchun1963@yahoo.com.cn
[1]School of Agriculture and Biology, Department of Animal Science, Shanghai Jiao Tong University, Shanghai 200240, PR China
[2]Shanghai Key Laboratory of Veterinary Biotechnology, Shanghai 200240, PR China
Full list of author information is available at the end of the article

whole genome sequences are being done. Furthermore, most of the machine learning approaches are based on the data sets of human, while the features of the pre-miRNAs also exhibit the species-specificity. Therefore, we are aimed to identify the porcine pre-miRNA in genome-wide through computational approaches by utilizing the data sets of pigs in our study, which may provide us a global scanning on the pre-miRNAs of pigs in genome level. In our study, we built the SVM-based porcine pre-miRNAs classifier with a sensitivity of 100%, a specificity of 91.2% and a total prediction accuracy of 95.6%, respectively. As a result, 2204 and 116 porcine pre-miRNA candidates were separately detected by using SVM-based pre-miRNAs classifier and homology searching.

Results and discussion

Performance of the SVM-based pre-miRNAs classifier

SVM-based porcine pre-miRNAs classifier was built by using the data sets of pigs. Interestingly, all of porcine pre-miRNAs of the test set were correctly detected by our classifier, which achieved a sensitivity (SE) of 100%, a specificity (SP) of 91.2% and a total prediction accuracy (ACC) of 95.6%, respectively. The power of the pre-miRNAs classifier was given in Table 1. Moreover, the performance of the classifier was also tested by a ROC curve. As shown in the Figure 1, the classifier achieved a five-fold cross-validation rate of 99.54%. In a word, it indicated that our classifier was available for the prediction of porcine pre-miRNAs. Additionally, it also demonstrated that the comprehensive use of the pre-miRNAs features of the secondary structure and sequence information was an effective strategy in pre-miRNAs prediction.

Xue et al. obtained an accuracy of 90% by using a set of features combining the local contiguous structures with sequence information to distinct the pre-miRNAs with that of pseudo pre-miRNAs [30], and those features have been used by several other pre-miRNA predicting methods [26,31,33]. Their studies demonstrated that those features were effective in pre-miRNA prediction. Thus, we also adopted those features in our study. Later, Jiang et al. found that the predicting performance significantly increased by combining the minimum of free energy (MFE) of the secondary structure or p-value feature with the local contiguous triplet structure composition feature. Their results indicated that a comprehensive feature vector was able to extract more information of a primary sequence and reach a better prediction performance [26]. Our classifier was capable of achieving a well prediction performance with an accuracy of 95.6% may be due to the using of a combined feature vector, because additional seven features used in our study have been proved to be one part of the optimized features subset in pre-miRNAs prediction by Wang et al. [3].

Table 1 Performance of the pre-miRNAs classifier on test sets.

Test set	Type	Size	Accuracy (%)
TE-S1	Real	40	100%
TE-S2	Pseudo	1000	91.20%

Test set represents positive and negative set used to test the power of the pre-miRNAs classifier. Type represents the classification of the test set. Size is the number of the real or pseudo pre-miRNAs contained in test set. Accuracy is the percentage of the real or pseudo correctly recognized by pre-miRNAs classifier.

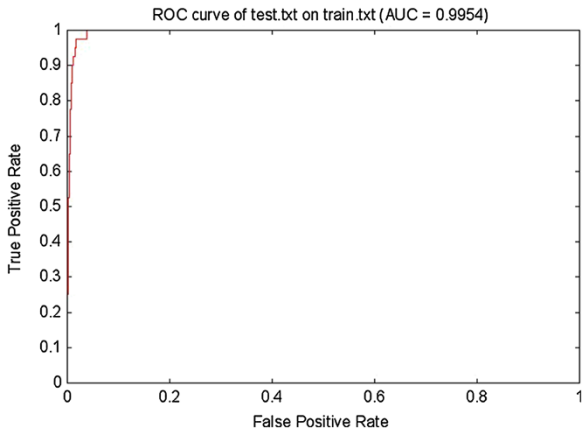

Figure 1 ROC curve for the pre-miRNAs classifier on the test set. The curve with more areas has better performance of the classifier. It showed the classifier reached a well performance.

Identification of pre-miRNAs candidates on pig genome using the SVM-based classifier

Since the genome sequences contain the full information of a species and the database of non-coding RNA of pigs is quite incompletely, thus we used whole genome sequences to construct the prediction set (PR-S). After splitting the pig genome, we obtained more than 222 million short sequences. The PR-S constructed by short sequences passed by pre-filter was further distinguished by our SVM-based pre-miRNAs classifier. As pre-filter parameters would be very useful in filtering the pseudo pre-miRNAs from huge number of similar pre-miRNA sequences, those pre-filters were incorporated into the SVM-based classifier to predict novel pre-miRNAs. Except for the redundancy and the known pre-miRNAs, we finally got 2204 pre-miRNA candidates with the probability more than 0.99995 in the pig genome. They were formed into 1849 clusters according to their locations in genome wide (inter-distance <=50kb [34]). Those pre-miRNA candidates were blasted with porcine CDS and other non-coding RNA (NONCODE v3.0, http://www.noncode.org/NONCODERv3/). The result shown that 6 novel pre-miRNAs (coverage >90%, identities =100% with CDS) overlap with coding region. Namely, 2198 out of 2204 new pre-miRNAs are in the

non-coding region. And none of pre-miRNAs (coverage >90%, identities >90% with non-coding RNAs) were found that overlap with other non-coding RNAs. The procedure for predicting porcine pre-miRNAs was given as Figure 2.

The large number of the novel pre-miRNA candidates indicated that there were still many unidentified pre-miRNAs in pigs. Previous studies estimated that the number of miRNAs have taken up to approximately 2-3% of the total number of genes in animal genomes [20]. According to our study, the number of the pre-miRNAs would be more than previous estimate. Expression profiling studies showed that most miRNAs were under the control of tissue-specific and development signaling, or both [35-37]. As a result, it may lead to a less number of miRNA identified by experimental methods and a low evaluate of pre-miRNAs' number. Indeed, in our studies, we regarded those pre-miRNA candidates as the real porcine pre-miRNAs in the view of bioinformatics. Meanwhile, those pre-miRNA candidates were set up to the porcine pre-miRNA library, the detail information of which was given in Additional file 1.

To explore the location distribution of all the pre-miRNA candidates, we calculated the number of pre-miRNA candidates in each chromosome. And the chromosome 1 covered the maximum number of pre-miRNAs candidates, while the chromosome 18 included the minimum. To a large extent, the number was consistent with the length of chromosome, namely the bigger of the chromosome the more number of pre-miRNA candidates it contained. The density analysis of pre-miRNA in chromosome showed that chromosome X, 8 and 16 maintained the highest density of pre-miRNA. The chromosome 8 was also found that it had a high density of quantitative trait locus (QTL) (http://www.animalgenome.org/cgi-bin/QTLdb/SS/index). Thus, the result suggested other researchers should pay more attention to study the chromosome 8 of pigs in the future. The result of density analysis of pre-miRNA and QTL in chromosome was given in Additional file 2.

At the same time, 215 unique pre-miRNAs were identified in pigs by Solexa sequencing in another published study [38]. Based on the comparison this data with ours, we found that 49 (coverage >90%, identities >90% with predicting pre-miRNAs) of above 215 unique pre-miRNAs were included in our study. In Chen et al.'s study, it mainly focused on identifying miRNAs in porcine backfat tissues. Tissues-specificity may lead to a bias on much more number of miRNAs identified in backfat tissues in their study, meanwhile some of their candidate miRNAs were unidentified by our method due to a limited length of 90-nt changed their features in our study. These may count for the low overlap rate. However, the result of Chen et al.'s study may still provide a piece of experimental evidence for our study. After the step of pre-filtering, a total of 160 known pre-miRNAs were retained in PR-S. 181 sequence fragments (coverage >90%, identities =100% with known pre-miRNAs) represented 115 known pre-miRNAs were detected by classifier. Namely, the sequence fragments of the known pre-miRNAs in the PR-S could be detected with the coverage of 72% (115 out of 160). The details those known pre-miRNAs sequence fragments were given in Additional file 3. There are several possible reasons accounting for that not all the reported porcine pre-miRNAs in miRNA Registry Database were covered in our studies. Firstly, not all the pre-miRNA sequences are expressed in the order of the genome sequence due to the RNA editing [39,40] , such as mir-381,mir-1271. According to our observation, 184 out of known 224 pre-miRNAs are completely identical to the sequence of the genome, thus 40 known pre-miRNA sequences unmapped to the genomic sequence data were filtered. Secondly, in order to reduce the pseudo pre-miRNAs as more as possible, the pre-filter parameters setting is up to some reported pre-miRNAs, such as the value of the minimal folding free energy index (MFEI). 160 out of 184 known pre-miRNAs were retained (20 known pre-miRNA were missed) after this step. Thirdly, the length of the short

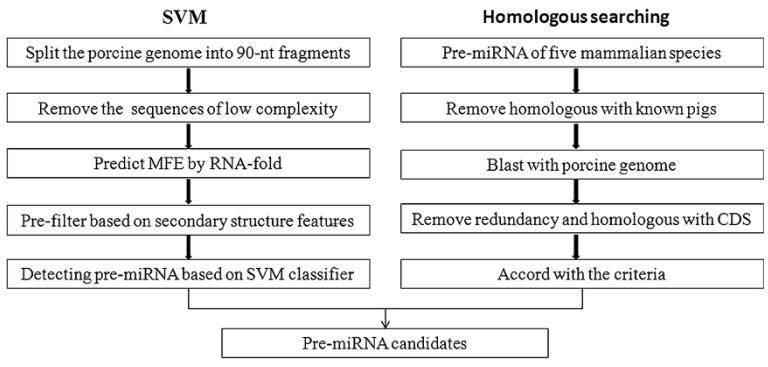

Figure 2 Flowchart of the porcine pre-miRNA prediction procedure.

sequence is limited to 90-nt, while some features of pre-miRNAs (such as adjusted minimal folding free energy (N(AMFE)) and the adjust number of paired nucleotides (N(ANNB)) have connection with the sequence length [32,41], which may influence the features of 45 reported pre-miRNAs and lead them to be undetected.

Although the classifier produced a specificity of 91.2%, the candidate hairpins could be lead to a certain number of false positives in genome-wide prediction. Thus, the next problem removing those pseudo pre-miRNAs in the library is needed to be considered deeply.

Identification of the pre-miRNAs candidates using the homologous searching

Since the pre-miRNA candidate sequences were split from genome with a specified length of 90-nt which may lead some of them undetected by our SVM classifier and the coverage of some model species with our SVM-based classifier result (coverage >85%, identities >85% with model species known pre-miRNAs) were 8% (human), 12% (mouse), 22% (rat), 16% (cow) and 31% (dog), which was not so high. The SVM-based classifier's training set was composed by the porcine known pre-miRNAs to predict the novel pre-miRNAs of pigs. The feature of pre-miRNAs exhibits the species-specificity. It may cause our SVM-based classifier have some biases to detect more pre-miRNA possessed only by pigs. The species-specificity and homologous porcine pre-miRNAs unidentified in model species may contribute to the low overlap rate. It was necessary to make it up by some other computational methods. At present, besides the SVM classifier the homologous searching is also a widely used method for identifying the pre-miRNAs, because the pre-miRNAs have a highly conservation among the different species [20]. What's more, in recent years, a large number of new pre-miRNAs were identified in some model species, such as Mouse, Human. Up to now, according to the records of miRNA Registry Database (Release 17, April 2011; http://mirbase.org), it contains human (1424), mouse (720), rat (408), cow (662) and dog (323). While, there are only 228 pre-miRNAs in pig. Therefore, it is quite necessary for us to do a homologous searching once again to find the new pre-miRNAs of porcine by using the identified pre-miRNAs in the other species.

According to the criteria mentioned in homologous searching method, we found 116 new pre-miRNAs candidates, and the detail information of which was given in Additional file 4. Interestingly, some pre-miRNAs candidates were mapped to more than one location of chromosomes. Guo et al. thought that cross-mapping events in pre-miRNAs revealed potential miRNA-mimics and evolutionary implications [42]. The newly identified porcine pre-miRNAs candidates belong to different miRNA families, such as miR-1282, miR-3059, miR-3120, miR-3618. Among them, miR-3120 initially identified from melanoma [43] and miR-3618 from human cervical cancer and normal cervices [44] have a highly conservation with pigs. We have also compared this result with the SVM-based and found no overlap between them. Actually, there were some of them passing SVM model before filtering in our study. However, when the prediction probability was set as more than 0.99995 to reduce false positive, they were filtered out with a result of no overlap between homology search and SVM-model candidates. There is no doubt that the high conservation of pre-miRNAs among the species also provides us a rapid way to identify the pig pre-miRNAs. This would be helpful to further enrich the resource of pre-miRNAs databases.

Conclusions

In conclusion, we built the SVM-based pre-miRNAs classifier using the known pre-miRNAs and CDS sets of the pigs. From the porcine genome, we discovered 2204 new pre-miRNAs candidates by our SVM-based classifier and 116 pre-miRNAs candidates by homology searching. Our study would provide guidance on further experimentally verifying swine pre-miRNA in the future and offer the opportunity to research gene function and the genetic mechanism of complex traits in genome level.

Methods

Sequence data collection

The porcine genomic sequences were available from UCSC database (Mar 2010, http://hgdownload.cse.ucsc.edu/goldenPath/susScr2/bigZips/). The precursor sequences of known miRNAs of *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Bos Taurus* (cow), *Canis familiaris* (dog) and *Sus scrofa* (pig) were obtained from miRNA Registry Database (Release 17, April 2011; http://mirbase.org) [19]. The porcine protein coding regions sequences (CDS) were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Sus_scrofa/RNA/), which were used as the pseudo pre-miRNA data.

The length of the pre-miRNAs sequences (LS)

The statistical length distribution of porcine pre-miRNA from miRNA Registry Database is that 86% of them within 75~105 nt. In our study, both the porcine genome sequences and CDS were divided into short sequences using a 90-nt sliding window with 9-nt increments at one time [3,33].

The complexity of the sequences

Low-complexity of the sequences, such as those with single nucleotide repeated > 8 times (for example, AAAAAAAA), dinucleotides repeated > 7 times (for

example, AGAGAGAGAGAGAG), trinucleotides repeated > 4 times (for example, ATGATGATGATG), were removed for further analysis, since we observed few known pre-miRNA possessed such sequences. Additionally, the sequences with the region of gap were removed.

MFE feature

MFE of the secondary structure was predicted by the Vienna RNA software package (RNAfold) (Version 1.8.5; http://www.tbi.univie. ac.at/~ivo/ RNA/) [45,46]. Previous studies indicated that pre-miRNAs have a high negative MFE and MFEI, which is a useful criterion to distinguish pre-miRNAs from all coding or non-coding RNAs [41]. The MFEI was calculated by the equation: MFEI = (- 100 × MFE/LS)/(G + C).

The three characteristics related to MFE were used as the feature vectors in SVM, and they were defined as follows:

$$N(\mathrm{MFE}) = (-\mathrm{MFE})/1000 \quad (1)$$

$$N(\mathrm{MFE}) = \mathrm{MFEI}/10 \quad (2)$$

$$N(\mathrm{AMFE}) = (-\mathrm{MFE})/(10 \times \mathrm{LS}) \quad (3)$$

Base-pairings and the secondary structure features

Because nucleic acid G can be paired with C or U, the base-pairings on the stem of the hairpin structure included the GU wobble pairs. And the threshold of the minimum base-parings of real pre-miRNA was 18. Indeed, the stem of the hairpin structure is highly conserved in pre-miRNAs, so we still only considered the stem regions of the pre-miRNA. The number of paired nucleotides (NNB), the adjust number of paired nucleotides (ANNB) and the number of nucleotides of the stem parts (NNS) were utilized as three feature vectors, defined as follows:

$$N(\mathrm{NNB}) = \mathrm{NNB}/1000 \quad (4)$$

$$N(\mathrm{ANNB}) = \mathrm{NNB}/\mathrm{LS} \quad (5)$$

$$N(\mathrm{NNS}) = \mathrm{NNB}/\mathrm{NNS} \quad (6)$$

Meanwhile, we denoted the contents of GC as follows:

$$N(\mathrm{GC}) = \mathrm{GC}/1000 \quad (7)$$

Besides, seven other features, including the structural diversity (N(Diversity)) (8), the frequency of the MFE structure (N(Freq/100)) (9) [46], adjusted base pair distance (N(dD)) (10) [47], average distance between internal loops (N(D_interlp/1000)) (11), the ratio of |A-U| to the length of sequence (N(|A-U|/LS)) (12), the length of the longest relaxed symmetry region (N(l_rsym_rgn/100)) (13) and the length of the longest symmetry region (N(l_sym_rgn/100)) (14), which were found as the optimized features for pre-miRNAs prediction according to the studies of Wang et.al [3], were also adopted.

The local adjacent sequence-structure features

Previous studies have shown that local sequence features play a crucial role in pre-miRNAs [48]. Additionally, Xue et al. found that the distributions of local contiguous sub-structures of pre-miRNAs are significantly distinguished with that of pseudo pre-miRNAs [30]. Therefore, in our study, we also characterized the secondary structure of pre-miRNAs by combining of the sequence information with the local contiguous structures.

There are only two conditions for each nucleotide in the predicted secondary structure by RNAfold [45], paired or unpaired, denoted by brackets "(" and dots ".", respectively. The left bracket "(" represents that paired nucleotide located near 5′-end which can be paired with another nucleotide at the 3′-end indicated by a right bracket ")". We used "(" for both situations without differentiating "(" or ")", because no evidence has indicated that mature miRNAs have a preference of the 3′ or 5′ arms of their hairpin precursors. Obviously, for any 3 adjacent nucleotides, there are eight possible structure units: "(((", "((.", "(.(", ".((", "(..", ".(. ", "..(", "...". Furthermore, by considering the left nucleotide among the three [31], there are 32 possible sequence-structure units, left-triplet coding ,denoted as "A(((", "U(((", "A((.", etc. as shown in Additional file 5 [31]. Here, we only considered the stem regions of a pre-miRNA by excluding the external single-stranded parts and the terminal loop. Similar features have been adopt by pioneer work, e.g. that Zhao et al. [31]. The frequency of each left-triplet coding of pre-miRNA was counted to create the 32 feature vectors. After normalizing, the frequency was used as input features for SVM. Combining with 14 features above, in all 46 feature vectors (summarized in Additional file 6) were taken as the input of SVM.

The pre-filter parameters of secondary structure features

Each sequence secondary structure, predicted by the Vienna RNAfold, was passed through a set of filter parameters. The filtering parameters [33,49,50]related to some terms of secondary structures were given as Additional file 7 [3], which were shown below.

(a) The number of hairpin loops = 1;
(b) The number of symmetrical loops < 6.
(c) The number of asymmetrical loops < 4.
(d) The number of bulges < 5.
(e) The total number of symmetrical and asymmetrical loops < 8.
(f) The total number of symmetrical, asymmetrical loops and bulges <10.
(g) The number of the base pairing >17.

(h) The value of ANNB is between 0.3~0.43.
(i) The length of symmetrical loops < 5.
(j) The length of asymmetrical loops <6.
(k) The length of bulges < 6.
(l) The MFE < −15kal/mol.
(m)The MFEI >0.7.
(n) The percentage of the GC contents is between 30-70%.

SVM data set

Among the 228 known porcine pre-miRNAs, whose secondary structures with no multiple loops were considered. 224 pre-miRNAs, covering more than 98% of all the reported porcine pre-miRNAs, were retained. We randomly extracted 184 pre-miRNAs from them as one part of training set (TR-S) and the remaining 40 pre-miRNAs formed into the test set 1 (TE-S1).

A pseudo pre-miRNAs set was collected from the porcine CDS and 5677 pseudo pre-miRNAs were selected due to their similar stem-loop structures to real pre-miRNAs. The criteria for extracting the pseudo pre-miNRAs from CDS segment was complied with the pre-filter parameters of the secondary structure features above. 184 pseudo pre-miRNAs selected randomly from the pseudo pre-miRNAs set composed another part of TR-S. Furthermore, we randomly took out 1000 pseudo pre-miRNA from the remaining pseudo pre-miRNAs set as test set 2 (TE-S2).

In addition, the porcine genome sequence fragments split from genome using a 90-nt sliding window with 9-nt increments at one time, passed the pre-filter parameters of secondary structure features (including (a),(g),(l) and (MFEI>0.6)), were collected for further identifying by SVM classifier and constructed the PR-S. The composition of each set was shown in Figure 3.

SVM

SVM, based on statistical theory [51], has a good generalization ability [52]. Therefore, in our study, SVM was adopted as a classifier to identify the real and pseudo pre-miRNAs. It was trained by the TR-S with the performance estimated by TE-S and applied to the PR-S. A 46-dimension feature vector referred to the above was taken as the input of SVM and the output was the number value “1”, which means the true, or “-1” indicating the false.

In our study, we downloaded a widely used software package Libsvm (Version 3.1, April 2011; http://www.csie.ntu.edu.tw/~cjlin/libsvm/) [53] to carry out our work. In order to acquire SVM classifier with optimal performance, we applied five cross-validation in model training, which could obtain the optimal penalty parameter C and the RBF kernel parameter g. Meanwhile, the performance of the SVM classifier was evaluated by following the assessment system used in RF [26].

Homologous searching

We chose pre-miRNAs of five other mammalian species (including human, mouse, rat, cow and dog), which have a highly homology with pigs. Firstly, we removed the

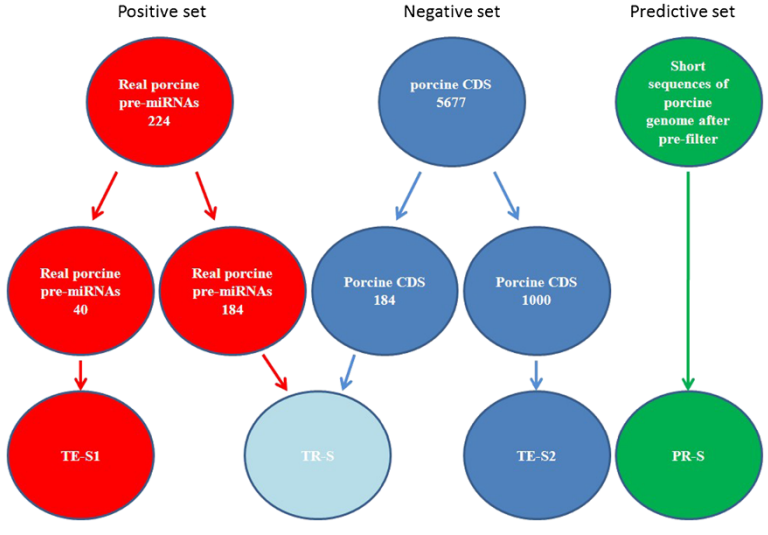

Figure 3 The composition of each set including the training set (TR-S), testing set (TE-S1 and TE-S2) and predictive set (PR-S). 184 real and pseudo porcine pre-miRNAs are randomly extracted from positive set (224 known real porcine pre-miRNAs) and negative set (5677 porcine CDS), respectively, and then they form into the training set. The remaining 40 real porcine pre-miRNAs compose the test set 1 (TE-S1). 1000 pseudo pre-miRNAs from the remaining negative set are randomly selected as test set 2 (TE-S2). Both TE-S1 and TE-S2 are used to test the performance of the SVM-based pre-miRNAs classifier. The predicting set (PR-S) is constructed by the porcine genome sequence fragments passed the pre-filter parameters of secondary structure features.

pre-miRNAs which have a highly homologous with 228 known porcine pre-miRNAs from the total pre-miRNAs of five species by utilizing the software of BLAST (ncbi-blast-2.2.25+; ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) [54]. Next, the remaining pre-miRNAs were blasted with the genome sequence of pigs and the sequence fragments (coverage >85%, identities >85% with pre-miRNAs) were retrieved from genome. Lastly, after discarding the redundant sequences, the sequences were regarded as pre-miRNA candidates if they accorded with the following criteria [55,56]:(i) an RNA sequence can fold into an stem-loop hairpin structure;(ii) predicted secondary structures had MFE less than -15kcal/mol;(iii) minimum base pairings on the stem of the hairpin structure is18;(iv) no multiple loops; (v) the GC contents is between 30~70%.

Additional files

Additional file 1: The list of porcine pre-miRNA candidates predicted by SVM-based classifier. The data provided represent the list of porcine pre-miRNA candidates predicted by SVM-based classifier in the whole genome of the pigs, and containing the information of their length, location in chromosome and genome location clusters.

Additional file 2: The result of density analysis of pre-miRNA and QTL in chromosome. The data provided the information of the number of pre-miRNA and QTL and their density in each chromosome.

Additional file 3: The list of porcine known pre-miRNA fragments of 90nt detected by SVM-based classifier. The data provided represents the list of porcine known pre-miRNA detected by SVM-based classifier in the whole genome of the pigs, and containing the information of their length, location in chromosome and the name of the represented known pre-miRNA.

Additional file 4: The list of porcine pre-miRNA candidates predicted by homology searching. The data provided represent the list of porcine pre-miRNA candidates predicted by homology searching, and containing the information of their length and location in chromosome.

Additional file 5: Local sequence-structure features of a hairpin were denoted by the left-triplet coding. Left-triplet elements are used to represent the local structure sequence features of a hairpin. The nucleotide type at the left and three local continuous substructures compose the left-triplet element. The appearances of all 32 possible triplet elements are counted along a hairpin segment to form a 32-dimensional vector, which is normalized to be the input vector for SVM.

Additional file 6: The 46 features used by SVM-based porcine pre-miRNAs classifier.

Additional file 7: The primary sequence of the has-let-7e precursor and the locations of some terms in the secondary structure. The upper part gives the primary structure of has-let-7e and the lower one shows the secondary structure and the correlative terms with varied colors.

Competing interests
The authors declare that they had no competing interests.

Authors' contributions
YP,QW and ZW designed the study. ZW collected the datasets from databases and analyzed the data, then prepared the original draft the manuscript. KH and YY guided the SVM analysis and the interpretation of the results. YP and QW reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was funded by National Natural Science Foundation of China (grant No. 31272414, 31072003 and 31000992), 2012 Animal Germplasm Resources Protection Project and Agriculture Development through Science and Technology Key Project of Shanghai (grant No. 2010 (1–3)).

Author details
[1]School of Agriculture and Biology, Department of Animal Science, Shanghai Jiao Tong University, Shanghai 200240, PR China. [2]Shanghai Key Laboratory of Veterinary Biotechnology, Shanghai 200240, PR China. [3]Department of Biology, Faculty of Science, Hong Kong Baptist University, Hong Kong, China.

Received: 26 December 2011 Accepted: 22 December 2012
Published: 27 December 2012

References
1. Ambros V: **The functions of animal microRNAs.** *Nature* 2004, **431**(7006):350–355.
2. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**(2):281–297.
3. Wang Y, Chen X, Jiang W, Li L, Li W, Yang L, Liao M, Lian B, Lv Y, Wang S, *et al*: **Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM.** *Genomics* 2011, **98**(2):73–78.
4. Sekiya Y, Ogawa T, Iizuka M, Yoshizato K, Ikeda K, Kawada N: **Down-regulation of cyclin E1 expression by microrna-195 accounts for interferon-beta-induced inhibition of hepatic stellate cell proliferation.** *J Cell Physiol* 2011, **226**(10):2535–2542.
5. Zhang Y, Wang Y, Wang X, Eisner GM, Asico LD, Jose PA, Zeng C: **Insulin promotes vascular smooth muscle cell proliferation via microRNA-208-mediated downregulation of p21.** *J Hypertens* 2011, **29**(8):1560–1568.
6. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM: **bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila.** *Cell* 2003, **113**(1):25–36.
7. Shah YM, Morimura K, Yang Q, Tanabe T, Takagi M, Gonzalez FJ: **Peroxisome proliferator-activated receptor alpha regulates a microRNA-mediated signaling cascade responsible for hepatocellular proliferation.** *Mol Cell Biol* 2007, **27**(12):4238–4247.
8. Wu G, Park MY, Conway SR, Wang JW, Weigel D, Poethig RS: **The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis.** *Cell* 2009, **138**(4):750–759.
9. Ambros V: **MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing.** *Cell* 2003, **113**(6):673–676.
10. Glass C, Singla DK: **ES cells overexpressing microRNA-1 attenuate apoptosis in the injured myocardium.** *Mol Cell Biochem* 2011, **357**(1-2):135–141.
11. Cheng AM, Byrom MW, Shelton J, Ford LP: **Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis.** *Nucleic Acids Res* 2005, **33**(4):1290–1297.
12. Jevnaker AM, Khuu C, Kjole E, Bryne M, Osmundsen H: **Expression of members of the miRNA17-92 cluster during development and in carcinogenesis.** *J Cell Physiol* 2011, **226**(9):2257–2266.
13. Osada H, Takahashi T: **MicroRNAs in biological processes and carcinogenesis.** *Carcinogenesis* 2007, **28**(1):2–12.
14. Hagan JP, Croce CM: **MicroRNAs in carcinogenesis.** *Cytogenet Genome Res* 2007, **118**(2–4):252–259.
15. Alvarez-Garcia I, Miska EA: **MicroRNA functions in animal development and human disease.** *Development* 2005, **132**(21):4653–4662.
16. Sayed D, Abdellatif M: **MicroRNAs in Development and Disease.** *Physiol Rev* 2011, **91**(3):827–887.
17. Garofalo M, Condorelli G, Croce CM: **MicroRNAs in diseases and drug response.** *Curr Opin Pharmacol* 2008, **8**(5):661–667.
18. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75**(5):843–854.
19. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32**(Database issue):D109–111.
20. Kim VN, Nam JW: **Genomics of microRNA.** *Trends Genet* 2006, **22**(3):165–173.

21. Li L, Xu J, Yang D, Tan X, Wang H: **Computational approaches for microRNA studies: a review.** *Mamm Genome* 2010, **21**(1–2):1–12.
22. Sheng Y, Engstrom PG, Lenhard B: **Mammalian microRNA prediction through a support vector machine model of sequence and structure.** *PLoS One* 2007, **2**(9):e946.
23. Zhang Y, Yu M, Yu H, Han J, Song C, Ma R, Fang J: **Computational identification of microRNAs in peach expressed sequence tags and validation of their precise sequences by miR-RACE.** *Mol Biol Rep* 2011, **39**(2):1975–1987.
24. Bhardwaj J, Mohammad H, Yadav SK: **Computational identification of microRNAs and their targets from the expressed sequence tags of horsegram (Macrotyloma uniflorum (Lam.) Verdc.).** *J Struct Funct Genomics* 2010, **11**(4):233–240.
25. Yue J, Sheng Y, Orwig KE: **Identification of novel homologous microRNA genes in the rhesus macaque genome.** *BMC Genomics* 2008, **9**:8.
26. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z: **MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.** *Nucleic Acids Res* 2007, **35**(Web Server issue): W339–344.
27. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK: **Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier.** *Bioinformatics* 2006, **22**(11):1325–1334.
28. Kadri S, Hinman V, Benos PV: **HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models.** *BMC Bioinforma* 2009, **10 Suppl 1**:S35.
29. Hsieh CH, Chang DT, Hsueh CH, Wu CY, Oyang YJ: **Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm.** *BMC Bioinforma* 2010, **11 Suppl 1**:S52.
30. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinforma* 2005, **6**:310.
31. Zhao D, Wang Y, Luo D, Shi X, Wang L, Xu D, Yu J, Liang Y: **PMirP: a pre-microRNA prediction method based on structure-sequence hybrid features.** *Artif Intell Med* 2010, **49**(2):127–132.
32. Batuwita R, Palade V: **microPred: effective classification of pre-miRNAs for human miRNA gene prediction.** *Bioinformatics* 2009, **25**(8):989–995.
33. Xu Y, Zhou X, Zhang W: **MicroRNA prediction with a novel ranking algorithm based on random walks.** *Bioinformatics* 2008, **24**(13):i50–58.
34. Li M, Liu Y, Wang T, Guan J, Luo Z, Chen H, Wang X, Chen L, Ma J, Mu Z, *et al*: **Repertoire of porcine microRNAs in adult ovary and testis by deep sequencing.** *Int J Biol Sci* 2011, **7**(7):1045–1055.
35. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse.** *Curr Biol* 2002, **12**(9):735–739.
36. Babak T, Zhang W, Morris Q, Blencowe BJ, Hughes TR: **Probing microRNAs with microarrays: tissue specificity and functional inference.** *RNA* 2004, **10**(11):1813–1819.
37. Breakfield NW, Corcoran DL, Petricka JJ, Shen J, Sae-Seaw J, Rubio-Somoza I, Weigel D, Ohler U, Benfey PN: **High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis.** *Genome Res* 2011, **22**(1):163–176.
38. Chen C, Deng B, Qiao M, Zheng R, Chai J, Ding Y, Peng J, Jiang S: **Solexa sequencing identification of conserved and novel microRNAs in backfat of Large White and Chinese Meishan pigs.** *PLoS One* 2012, **7**(2): e31426.
39. Rueter SM, Dawson TR, Emeson RB: **Regulation of alternative splicing by RNA editing.** *Nature* 1999, **399**(6731):75–80.
40. Luciano DJ, Mirsky H, Vendetti NJ, Maas S: **RNA editing of a miRNA precursor.** *RNA* 2004, **10**(8):1174–1177.
41. Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA: **Evidence that miRNAs are different from other RNAs.** *Cell Mol Life Sci* 2006, **63**(2):246–254.
42. Guo L, Liang T, Gu W, Xu Y, Bai Y, Lu Z: **Cross-mapping events in miRNAs reveal potential miRNA-mimics and evolutionary implications.** *PLoS One* 2011, **6**(5):e20517.
43. Stark MS, Tyagi S, Nancarrow DJ, Boyle GM, Cook AL, Whiteman DC, Parsons PG, Schmidt C, Sturm RA, Hayward NK: **Characterization of the Melanoma miRNAome by Deep Sequencing.** *PLoS One* 2010, **5**(3): e9685.
44. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO: **Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls.** *BMC Biol* 2010, **8**:58.
45. Hofacker IL: **RNA secondary structure analysis using the Vienna RNA package.** *Curr Protoc Bioinformatics* 2009, **Chapter 12**:Unit12 12.
46. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**(13):3429–3431.
47. Freyhult E, Gardner PP, Moulton V: **A comparison of RNA folding measures.** *BMC Bioinforma* 2005, **6**:241.
48. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics* 2004, **20**(17):2911–2917.
49. Oulas A, Poirazi P: **Utilization of SSCprofiler to predict a new miRNA gene.** *Methods Mol Biol* 2011, **676**:243–252.
50. Zhou B, Liu HL: **Computational identification of new porcine microRNAs and their targets.** *Anim Sci J* 2010, **81**(3):290–296.
51. Vapnik V: *Statistical Learning Theory*. Wiley-Interscience; 1998.
52. Davide A, Andrea B, Sandro R: **Evaluating the Generalization Ability of Support Vector Machines through the Bootstrap.** *Neural Processing Letters* 2000, **11**:51–58.
53. Chang C-C, Lin C-J: **LIBSVM: a library for support vector machines.** *ACM Transactions on Intelligent Systems and Technology* 2011, **2**(27):21–27. 27.
54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinforma* 2009, **10**:421.
55. Zhang B, Pan X, Anderson TA: **Identification of 188 conserved maize microRNAs and their targets.** *FEBS Lett* 2006, **580**(15):3753–3762.
56. Long JE, Chen HX: **Identification and characteristics of cattle microRNAs by homology searching and small RNA cloning.** *Biochem Genet* 2009, **47**(5–6):329–343.

doi:10.1186/1471-2164-13-729
Cite this article as: Wang *et al.*: The prediction of the porcine pre-microRNAs in genome-wide based on support vector machine (SVM) and homology searching. *BMC Genomics* 2012 **13**:729.