RESEARCH NOTE

An integrative bioinformatics pipeline for the genomewide identification of novel porcine microRNA genes

WEI FANG[1], NA ZHOU[1], DENGYUN LI[2], ZHIGANG CHEN[1], PENGFEI JIANG[2,3] and DELI ZHANG[2,3]*

[1]*Center for Bioinformation, College of Life Science,* [2]*Department of Preventive Veterinary Medicine, College of Veterinary Medicine, and* [3]*Institute of Veterinary Immunology, Northwest A and F University, Yangling 712100, Shaanxi, People's Republic of China*

[Fang W., Zhou N., Li D., Chen Z., Jiang P. and Zhang D. 2013 An integrative bioinformatics pipeline for the genomewide identification of novel porcine microRNA genes. *J. Genet.* **92**, 587–593]

Introduction

MicroRNA (miRNA) is a pivotal type of noncoding RNA gene in posttranscriptional gene regulation (David 2004). The majority of miRNAs in pig (*Sus scrofa*), an important domestic animal, remain unknown. From this perspective, we attempted the genomewide identification of novel porcine miRNAs. Here, we propose a novel integrative bioinformatics pipeline to identify conservative and non-conservative novel miRNAs. We used methods such as homology searching against known metazoan miRNAs, conservation filtering, and *ab initio* approaches. As a result, 222 new porcine miRNAs were identified and 14 more porcine miRNA gene families were discovered. miRNA:target pairs, 78,060, between 224 miRNAs and 4384 mRNAs have been predicted, and these miRNA targets were involved in a wide spectrum of regulatory functions and metabolic biological processes. An overall analysis of genome-scale gene locations and sequence characteristics was also conducted, and a detailed user manual of the pipeline is provided. Through this study, the number of identified porcine miRNAs increased, and an effective alternative strategy for genomewide miRNA identification was developed.

Materials and methods

Starting sequence dataset

All metazoan miRNA precursors and mature sequences were retrieved from the miRBase Release 19. To avoid redundancy or overlapping of miRNAs, we removed the repeated items in the miRNA precursors and mature sequences. At the end, 14,179 pre-miRNA sequences and 12,004 mature miRNA sequences remained. Unique pre-miRNA and mature miRNA sequences were used as query sequences to identify the conserved miRNA genes in the pig genome.

The repeat-masked Nov. 2009 (SGSC Sscrofa9.2/susScr2) pig genome was downloaded from the UCSC Genome Browser database (Dreszer *et al.* 2012). The genome comprised 18 autosomes and the sex chromosome X. The assembly sequences were saved in FASTA format, one chromosome in each file, for further analysis.

Identification of novel porcine miRNA genes

The identification of novel miRNA genes from the pig genome involved two steps. Homology-based approaches were adopted in the first stage, and *ab initio* approaches were used in the second stage. Figure 1 shows an overview of this prediction process.

Homology-based approaches to extract candidate miRNA genes: Conserved miRNA genes were extracted from the pig genome using a previously reported method (Baev *et al.* 2009). Sequence conservation and structure conservation of miRNA genes were considered in this method. Each chromosome sequence of the pig genome was processed using a local copy of the BLAT analysis tool (Kent 2002) and searched against the unique pre-miRNA sequence set with default parameters. A Perl script was used to convert the BLAT result into a BED format file, which included coordinates and identity scores for all homology regions. The BED

*For correspondence. E-mail: zhangdeli@tsinghua.org.cn.

Keywords. integrated approach; homology search; *ab initio* identification; *Sus scrofa*.

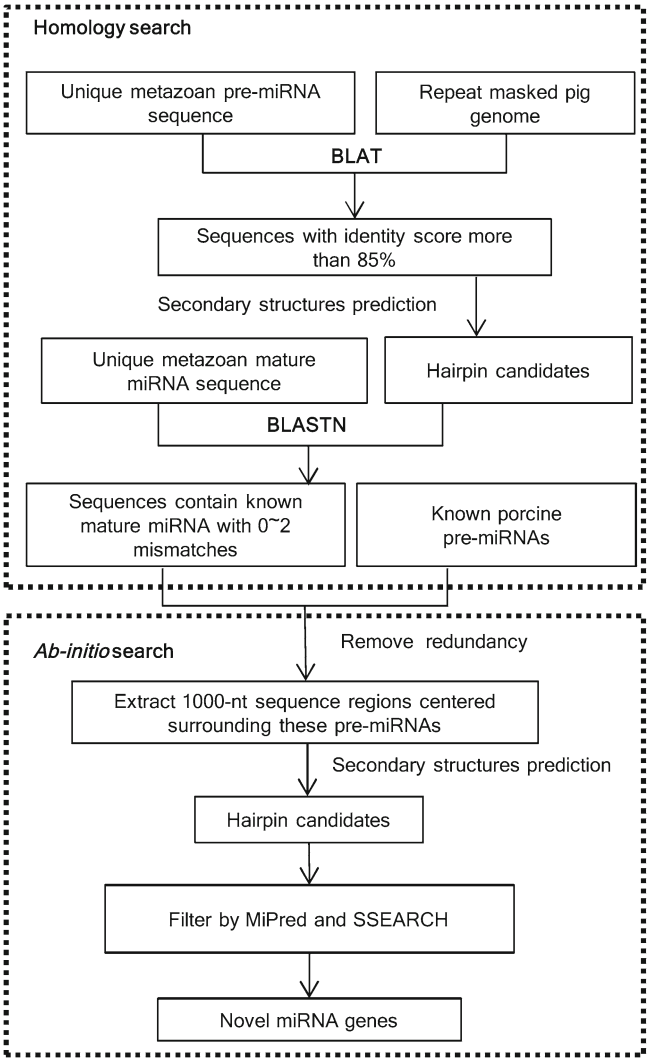

Figure 1. Flowchart of the porcine miRNA gene prediction procedure.

file was then uploaded into the Galaxy tool (Goecks *et al.* 2010) and linked to genome susScr2. The sequences with an identity score higher than 85% were fetched in FASTA format according to the coordinates. The secondary structure of these fetched sequences (see figures 1 and 2 in electronic supplementary material at http://www.ias.ac.in/jgenet/) was predicted by the RNA-fold procedure in Vienna RNA package ver. 2.0.7 (Lorenz *et al.* 2011). The threshold free energy for the secondary structures was −25 kcal/mol. Pre-miRNAs with multiple loops in the stem–loop structure were removed. Next, BLAST analysis was conducted to check the presence of mature miRNAs in the candidate precursor sequences. The final dataset contained cases of 100% identical mature miRNAs, as well as cases where the precursor had a proper stem–loop structure and the mature miRNA region was slightly altered (2 nt changes at most). Redundant sequences of the selected miRNAs and known porcine miRNAs were removed according to genomic location, and the miRNA gene set 'S1' was obtained.

Discovery of candidate miRNA genes by ab initio approaches: Previous studies have shown that miRNA genes often form clusters in animal genomes within several kilobase distances (Altuvia *et al.* 2005; Lagos-Quintana *et al.* 2003). Therefore, additional miRNAs may be likely discovered in genomic regions within the vicinity of known miRNAs (Jiang *et al.* 2007). To predict the hairpin structure, 1000-nt sequence regions surrounding the miRNA gene set 'S1' were extracted with a 100-nt sliding window using the script implemented in CRAVELA (Mendes *et al.* 2010). A locally installed MiPred (Jiang *et al.* 2007) was then used to screen novel pre-miRNA candidates. All screened pre-miRNAs were piped to SSEARCH analysis (http://mirbase.org/) to detect the mature miRNAs present in these precursors. The miRNA gene sequences containing mature miRNAs with at most 4-nt mismatches against known metazoan mature miRNAs were stored to form gene set 'S2'.

Prediction of miRNA targets in pig

The potential targets of porcine miRNAs were predicted using the miRanda (http://www.microrna.org/microrna/home.do), PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html) and RNA hybird (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid) programmes with default parameters. To search for potential target mRNAs, mature miRNA sequences were searched against the 3′ untranslated regions of porcine mRNAs downloaded from UTRdb. For the purpose of reducing false positives, we took the intersection of the prediction results achieved by the above-mentioned three programs as final results.

To better understand the roles of the predicted miRNA target genes in pig, the biological process categories of the miRNA targets were searched by gene ontology. We subjected potential miRNA targets to functional enrichment analysis against gene ontology (GO) biological process terms database by using Biological Networks Gene Ontology (BiNGO). For enrichment *P* value calculation (at a significance level of 0.0001 or better), a hypergeometric distribution statistical testing method was selected to ensure that target genes are not hitting their corresponding biological function/process classes purely by random chance. For multiple hypotheses testing, the Benjamini and Hochberg false discovery rate correction was applied to reduce false negatives at the cost of a few more false positives (Benjamini and Hochberg 1995).

Characteristics analysis of miRNA genes

All miRNA sequence features, including length, G+C content, minimal folding free energies (MFEs), and minimal folding free energy indexes (MFEIs), were explored using

a Python script. Genome-scale location analysis was performed to further understand the distribution characteristics of porcine miRNAs. Further, Thatcher's (2008) method was adopted to analyse the miRNA gene family and miRNA gene cluster. Family members were strictly determined using identical seed sequences (2nd to 7th nts of the 5′ end of the mature miRNA), and genes within 3 kb of a known or predicted miRNA were classified as polycistronic family members (Thatcher *et al.* 2008).

Results and discussion

Novel porcine miRNA genes

Several studies have attempted to identify miRNAs in pig genome (Li *et al.* 2011; Lian *et al.* 2012; Liu *et al.* 2013). However, compared with the computational approach, these identification methods were deficient in that they were tissue specific and time point specific, and may fail in identifying new miRNAs at a genomewide scale. To date, several computational approaches have been developed to screen novel miRNAs from whole genome sequences, but most are based on sequence alignment analysis and are unable to detect the distant homologs that diverge in sequence but maintain a conservative structure (Legendre *et al.* 2005). More sensitive methods can be developed by considering structure conservation (Mendes *et al.* 2009). Accordingly, we report an improved integrative computational method that considers sequence conservation and structure conservation of miRNA genes to enhance prediction efficiency and accuracy.

We used the improved integrative computational approach to identify novel miRNAs from the pig genome. In total, 222 new miRNA genes were recognized across all these chromosomes, including 239 mature miRNAs located in the 3′ or 5′ arm of these miRNA genes (table 1 in electronic supplementary material). The mature miRNAs could be transcribed and processed from sense and antisense transcripts derived from the same genomic loci (Chen *et al.* 2004). We found a total of 17 pairs of -5p/-3p mature miRNAs (opposite arms of the same pre-miRNA), which accounted for 7.11% of all newly identified mature miRNAs.

We searched the predicted novel porcine miRNAs against known miRNAs and found that seven were annotated in the genomes of other species. For example, mir-208a, identified in an intron of MYH7 in human genome, has been proved to be associated with RNA-induced silencing complex, which affects RNA interference.

With the method described in the Materials and Methods section, we identified a total of 78,060 miRNA:target pairs between 224 identified unduplicated mature miRNAs and 4384 porcine miRNA 3′ untranslated region sequences. The 4384 potential miRNA target hits belonged to thousands of gene families and had different biological functions. There were 1286 potential target hits for miR-762, which ranked the most targets in the 224 mature miRNAs. However, only 34 potential target hits were identified for miR-981-3p, which had the fewest targets. There was an average of 348 potential target hits identified for each mature miRNA.

To better understand the roles of the predicted miRNA target genes in pig, the biological process categories of the miRNA targets were searched for using GO. The targets were annotated using the GO annotations available from the UniProt-GOA database. BiNGO was applied to study target enrichment and to construct a hierarchical ontology tree in Cytoscape. As shown in figure 3 in electronic supplementary material, the results showed that the miRNA families in pig preferentially target genes are involved in a wide spectrum of regulatory functions and metabolic biological processes.

Characteristics of porcine pre-miRNAs and mature miRNAs

Details on the identified porcine miRNA precursors, including length, G+C content, minimal folding free energies (MFEs) and minimal folding free energy indexes (MFEIs) are summarized in table 1. The length of miRNA precursors varied from 52 to 117 nt, with an average length of 87.36 ± 28.49 nt. The majority of the newly identified miRNAs (77%) contained 60 to 99 nucleotides (figure 2a). The G+C content of the novel miRNA precursors ranged from 28.71% to 85.54% (table 1). The G+C contents were consistent with those of known miRNAs in pig, human and mouse genomes. Previous prediction methods have shown that MFE is an important feature of miRNAs. These newly identified pig miRNA precursors have relatively higher negative MFE (20.5–59.2 kcal/mol). However, MFEs are strongly and positively correlated with their sequence length (Zhang *et al.* 2008). The longer the RNA sequence, the lower the MFE. Thus, a comparison based only on the original MFE is not appropriate. Zhang *et al.* (2008) developed two new terms called the adjusted MFE (AMFE) and MFEI to make the MFEs comparable. Their findings demonstrated that MFEI is an important criterion to distinguish miRNA from other types of RNA (Zhang *et al.* 2008). The MFEIs of newly found porcine pre-miRNAs ranged from 0.35 to 1.29, with a comparatively small average of 0.66 ± 0.29, and were within the range of known human and mouse pre-miRNAs (table 1). Figure 2a illustrates the length distribution of all metazoan and porcine miRNAs (miRNA gene set 'S1' and 'S2'), which show that the length distribution in porcine pre-miRNA sequences and in all metazoan pre-miRNA sequences had a similar pattern.

We further studied the length distribution of mature miRNAs. As shown in figure 2b, the length distribution of mature miRNAs in pig was similar to that in metazoans in general. The length of mature miRNAs in pig ranged from 17 to 27 nt, and the majority of the miRNAs had a length of 22 nt.

Genomic locations of porcine miRNAs

To determine the distribution characteristics of miRNAs in the pig genome, all of the 467 (26 previously documented

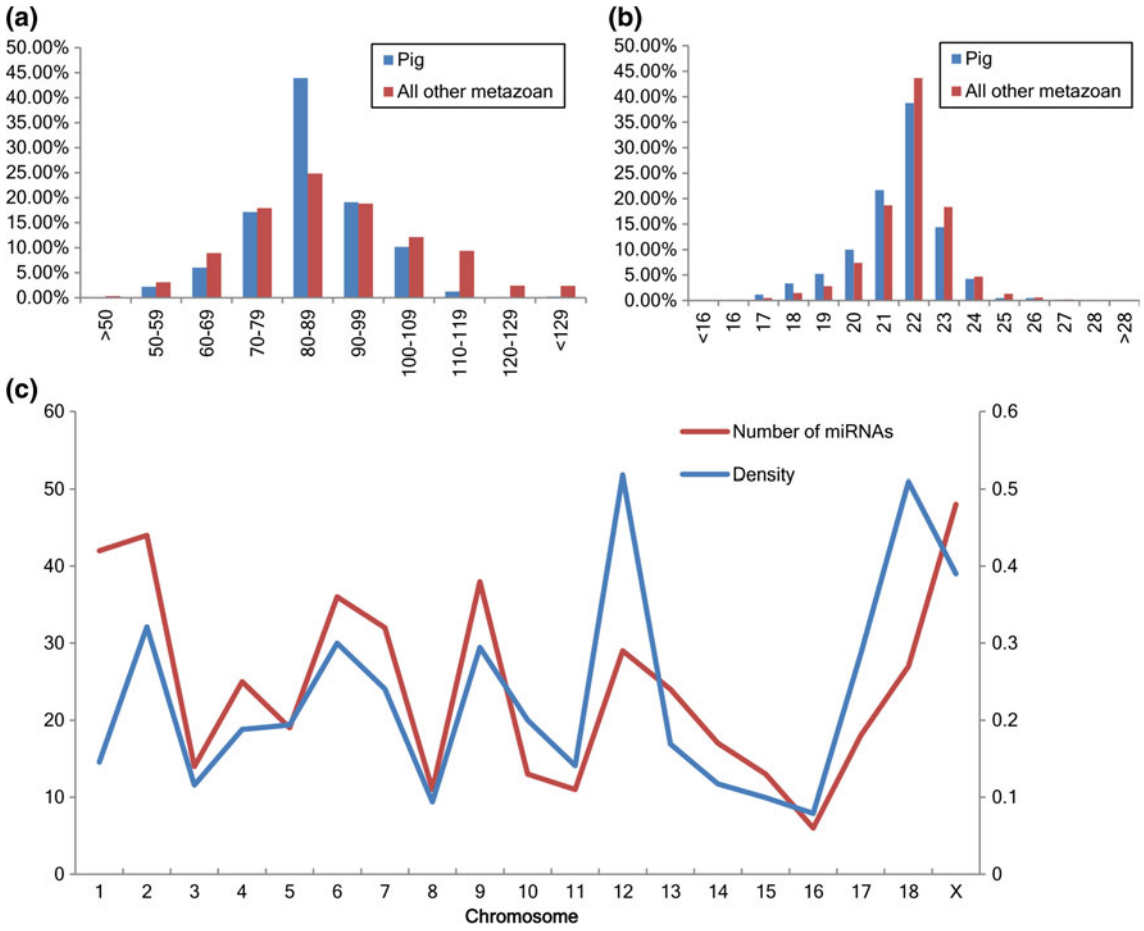

Figure 2. Characteristics of porcine miRNA genes. (a) Length distribution of all porcine and metazoan pre-miRNAs. (b) Length distribution of all porcine and metazoan mature miRNAs. (c) Number of miRNA genes and miRNA density across the pig genome. The miRNA density is defined as the number miRNA genes per Mb.

porcine miRNAs could not be located in the chromosome) porcine miRNA genes were mapped onto the pig genome in Ensembl (Flicek *et al.* 2012). Figure 3 shows the location of the miRNAs across 19 porcine chromosomes. Some chromosomes were relatively miRNA deserts (chromosomes 8, 11 and 16) while others encoded large numbers of miRNAs (chromosomes 1, 2 and X). The detailed porcine miRNA locations are presented in supplementary table 1. We also studied the relationship between the number of miRNA genes and the miRNA density across 19 porcine chromosomes (figure 2c). Here, miRNA density was defined as the number of miRNA genes per Mb. As shown in figure

Table 1. Characteristics of the newly identified miRNAs, and known pig, human and mouse miRNAs.

Characteristic	Newly identified pig miRNAs ($n = 222$)	Known miRNAs of pig ($n = 271$)	Known miRNAs of human ($n = 1600$)	Known miRNAs of mouse ($n = 855$)
Length of precursors	52–117	64–149	41–180	49–141
	87.36±28.49	82.4±18.04	83.47±33.34	87.89±33.92
Per cent G+C content	28.71–85.54	25–86.92	16.33–92.86	21.67–83.12
	58.87±25.64	50.09±19.93	50.16±25.44	48.9±19.71
MFEs	16.2–67.8	19.2–75.9	3.9–144.4	12.6–80.8
	33.16±20.01	36.33±16.76	39.41±25.57	35.77±18.51
MFEIs	0.35–1.29	0.44–1.71	0.28–2.31	0.38–1.94
	0.66±0.29	0.90±0.34	0.99±0.67	0.86±0.44

The range and mean with standard errors of all characteristics are presented. All the data of known miRNAs were retrieved from miRBase (Release 19.0). Per cent G+C content, G+C content over pre-miRNA sequence; MFEs, minimal folding free energies; and MFEIs, minimal folding free energy indexes.

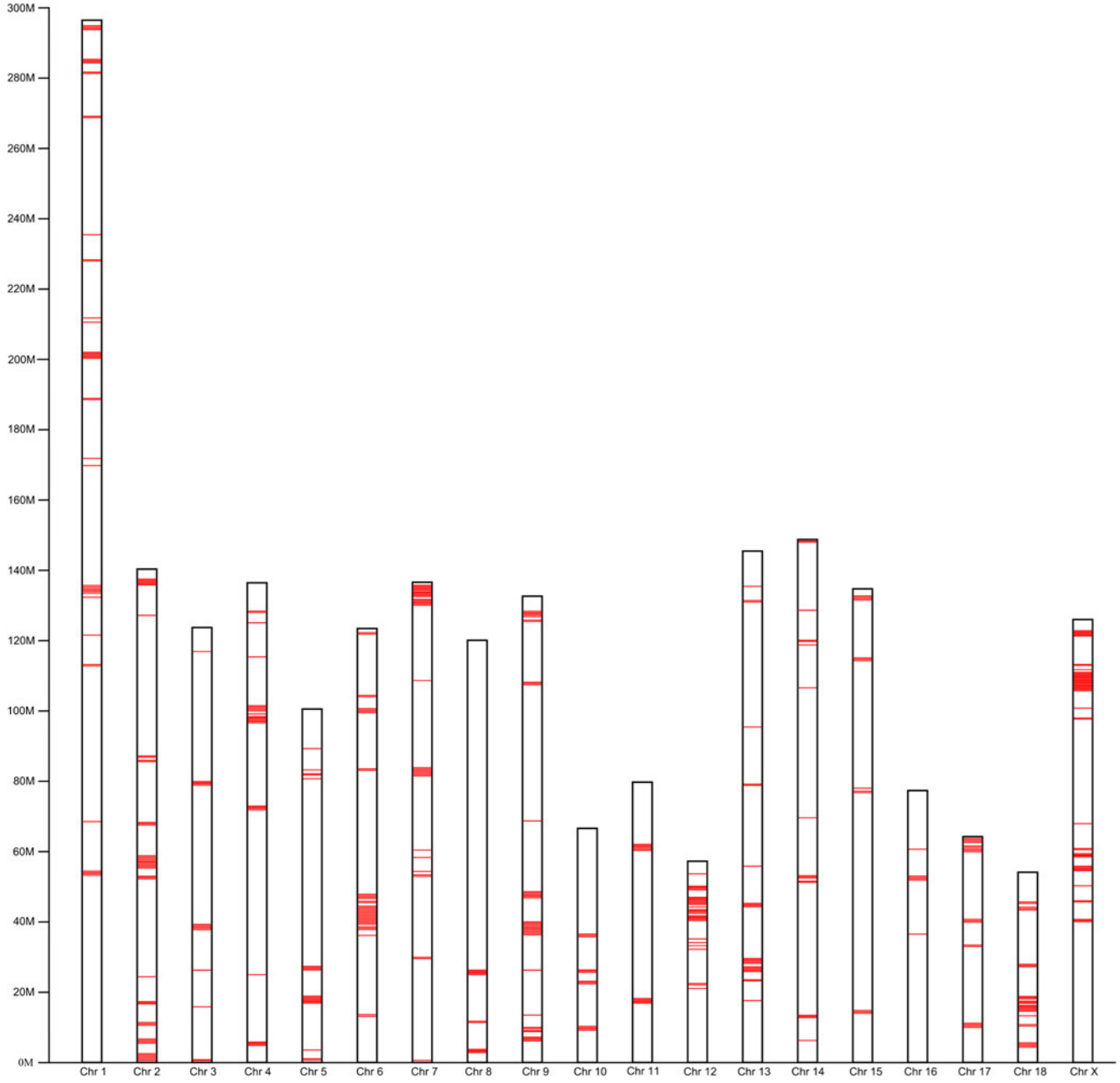

Figure 3. Chromosomal location of porcine miRNAs. The relative locations of 467 miRNA genes are shown across 19 chromosomes in the pig genome. The red line represents the miRNA gene, and the colour intensity represents the number of miRNA genes in this region.

2c, chromosomes 12, 18 and X had the highest miRNA density. We then determined whether a correlation existed between miRNA density and chromosome length. The correlation analysis result across the 19 chromosomes indicated that miRNA density was not significantly correlated with chromosome length (Spearman's $\rho = 0.36$).

Based on chromosomal location and additional genomic analysis, we distinguished miRNAs that were encoded as distinct transcripts versus those encoded within the transcription units of other genes. Approximately 50% of all mammalian miRNAs are located within other transcription units (Rodriguez *et al.* 2004; Griffiths-Jones *et al.* 2008). For pig, 240 (51%) miRNAs are hosted within intronic regions, and 227 (49%) miRNAs are found in intergenic regions (table 1 in electronic supplementary material). We defined 'intergenic miRNAs' as miRNAs that reside between protein-coding genes, and 'intronic miRNAs' as miRNAs that overlap protein-coding genes.

Novel miRNA genes clusters in pig genome

miRNA genes often form clusters in the genome (Lagos-Quintana *et al.* 2001). Clusters can be defined as miRNA genes present in the same orientation and transcribed in

one polycistronic unit (Baev *et al.* 2009). We used 3 kb as the distance threshold to determine whether miRNAs could be classified as part of polycistronic transcripts. A total of 111 miRNA gene clusters were identified and 314 miRNAs accounting for 62.5% of the total porcine miRNAs were part of polycistronic transcripts (table 2 in electronic supplementary material), which was larger than that found in zebrafish genome (50%) (Thatcher *et al.* 2008) and horse genome (36%) (Zhou *et al.* 2009). The 111 porcine miRNA clusters contained 57 pairs, 24 triplets, 20 groups of four, six groups of five, two groups of six and two groups of seven, and were differentially located onto the 19 chromosomes (figure 3; table 2 in electronic supplementary material). The length of these miRNA clusters varied from 163 to 2929 bp, with an average of 809 bp. Approximately 64% of the clusters had lengths ranging from 313 to 1132 bp. Given that porcine miRNAs are located in the protein-coding regions or the intergenic regions, the identified 111 miRNA gene clusters were classified into two groups: miRNA clusters from intergenic regions and intronic regions. Our analysis results show that 69 of these miRNA gene clusters were located in intergenic regions and 39 in intronic regions. However, the last three gene clusters were located in overlapping intergenic and intronic regions (table 2 in electronic supplementary material).

Numerous studies have shown that cluster miR-17-92, which encodes miR-17-5p, miR-17-3p, miR-18a, miR-19a, miR-20a, miR-19b and miR-92a, considerably contributes to the development of the heart, lung and immune system, and is highly conserved among human, mouse, rat and horse genomes (Olive *et al.* 2010). The cluster was also found in pig genome, specifically in chromosome 11 (table 2 in electronic supplementary material). Cluster miR-106-363, a paralogous cluster of miR-17-92, was found in human and mouse chromosome X. This cluster, which contains six genes, miR-18b, miR-19b, miR-20b, miR-92a, miR-106a and miR-363, was also identified in pig chromosome X (table 2 in electronic supplementary material). These results indicate that these two gene clusters may also be highly conserved in the pig genome.

Novel porcine miRNA gene families

Animal miRNAs are usually grouped as gene families. On the basis of the sequences identified within the seed region, miRNAs can be grouped with the prediction that specific mRNAs can be targeted by multiple miRNAs if these miRNAs contain identical seed sequences even if other downstream nucleotides vary (Thatcher *et al.* 2008). The seed region is defined as nucleotides 2–7 from 5′ end of mature miRNA sequences (Gitlin *et al.* 2005). We used a previously described method (Thatcher *et al.* 2008) and placed porcine miRNAs with identical seed regions in the same family. Consequently, 14 novel porcine gene families were found (table 3 in electronic supplementary material).

Pipeline implementation

A detailed user guide to implement the pipeline and Python scripts to facilitate data processing are presented in supplementary material.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant no. 31072115), the opening foundation of the State Key Laboratory of Biology Macromolecule of the Institute of Biophysics, Chinese Academy of Sciences (grant no. O5SY021107), and a preparatory project sponsored by the National Ministry of Science and Technology of China of the First Batch in the Basic Research Category of the National Program of Science and Technology in the Field of Countryside for 2011 to 2015 (preparatory project no. NC2010CD0178).

References

Altuvia Y., Landgraf P., Lithwick G., Elefant N., Pfeffer S., Aravin A. *et al.* 2005 Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res.* **33**, 2697–2706.

Baev V., Daskalova E. and Minkov I. 2009 Computational identification of novel microRNA homologs in the chimpanzee genome. *Comput. Biol. Chem.* **33**, 62–70.

Benjamini Y. and Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300.

Chen J., Sun M., Kent W. J., Huang X., Xie H., Wang W. *et al.* 2004 Over 20% of human transcripts might form sense–antisense pairs. *Nucleic Acids Res.* **32**, 4812–4820.

David P. B. 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297.

Dreszer T. R., Karolchik D., Zweig A. S., Hinrichs A. S., Raney B. J., Kuhn R. M. *et al.* 2012 The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40**, 918–923.

Flicek P., Amode M. R., Barrell D., Beal K., Brent S., Carvalho-Silva D. *et al.* 2012 Ensembl 2012. *Nucleic Acids Res.* **40**, 84–90.

Gitlin L., Stone J. K. and Andino R. 2005 Poliovirus escape from RNA interference: short interfering RNA-target recognition and implications for therapeutic approaches. *J. Virol.* **79**, 1027–1035.

Goecks J., Nekrutenko A., Taylor J. and Team G. 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86.

Griffiths-Jones S., Saini H. K., van Dongen S. and Enright A. J. 2008 miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, 154–158.

Jiang P., Wu H., Wang W., Ma W., Sun X. and Lu Z. 2007 MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, 339–344.

Kent W. J. 2002 BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

Lagos-Quintana M., Rauhut R., Lendeckel W. and Tuschl T. 2001 Identification of novel genes coding for small expressed RNAs. Science **294**, 853–858.

Lagos-Quintana M., Rauhut R., Meyer J., Borkhardt A. and Tuschl T. 2003 New microRNAs from mouse and human. *RNA* **9**, 175–179.

Legendre M., Lambert A. and Gautheret D. 2005 Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics* **21**, 841–845.

Li G., Li Y., Li X., Ning X., Li M. and Yang G. 2011 MicroRNA identity and abundance in developing swine adipose tissue as determined by solexa sequencing. *J. Cell Biochem.* **112**, 1318–1328.

Lian C., Sun B., Niu S., Yang R., Liu B., Lu C. *et al.* 2012 A comparative profile of the microRNA transcriptome in immature and mature porcine testes using Solexa deep sequencing. *FEBS J.* **279**, 964–975.

Liu Y., Li M., Ma J., Zhang J., Zhou C., Wang T. *et al.* 2013 Identification of differences in microRNA transcriptomes between porcine oxidative and glycolytic skeletal muscles. *BMC Mol. Biol.* **14**, 7.

Lorenz R., Bernhart S., Höner zu Siederdissen C., Tafer H., Flamm C., Stadler P. *et al.* 2011 ViennaRNA Package 2.0. *Algorithm. Mol. Biol.* **6**, 26.

Mendes N., Freitas A. and Sagot M. F. 2009 Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res.* **37**, 2419–2433.

Mendes N., Freitas A., Vasconcelos A. and Sagot M. F. 2010 Combination of measures distinguishes pre-miRNAs from other stem-loops in the genome of the newly sequenced Anopheles darlingi. *BMC Genomics* **11**, 529.

Olive V., Jiang I. and He L. 2010 mir-17-92, a cluster of miRNAs in the midst of the cancer network. *Int. J. Biochem. Cell Biol.* **42**, 1348.

Rodriguez A., Griffiths-Jones S., Ashurst J. L. and Bradley A. 2004 Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14**, 1902–1910.

Thatcher E. J., Bond J., Paydar I. and Patton J. G. 2008 Genomic organization of zebrafish microRNAs. *BMC Genomics* **9**, 253.

Zhang B., Pan X. and Stellwag E. J. 2008 Identification of soybean microRNAs and their targets. *Planta* **229**, 161–182.

Zhou M., Wang Q., Sun J., Li X., Xu L., Yang H. *et al.* 2009 In silico detection and characteristics of novel microRNA genes in the *Equus caballus* genome using an integrated *ab initio* and comparative genomic approach. *Genomics* **94**, 125–131.

Received 8 March 2013, in final revised form 30 June 2013; accepted 8 July 2013
Published on the Web: 6 December 2013