

RESEARCH ARTICLE

Open Access



# Identification and characterization of long intergenic noncoding RNAs in bovine mammary glands

Chao Tong<sup>1</sup>, Qiaoling Chen<sup>1</sup>, Lili Zhao<sup>1</sup>, Junfei Ma<sup>1</sup>, Eveline M. Ibeagha-Awemu<sup>2</sup> and Xin Zhao<sup>1,3\*</sup>

## Abstract

**Background:** Mammary glands of dairy cattle produce milk for the newborn offspring and for human consumption. Long intergenic noncoding RNAs (lincRNAs) play various functions in eukaryotic cells. However, types and roles of lincRNAs in bovine mammary glands are still poorly understood.

**Results:** Using computational methods, 886 unknown intergenic transcripts (UITs) were identified from five RNA-seq datasets from bovine mammary glands. Their non-coding potentials were predicted by using the combination of four software programs (CPAT, CNCI, CPC and hmmscan), with 184 lincRNAs identified. By comparison to the NONCODE2016 database and a domestic-animal long noncoding RNA database (ALDB), 112 novel lincRNAs were revealed in bovine mammary glands. Many lincRNAs were found to be located in quantitative trait loci (QTL). In particular, 36 lincRNAs were found in 172 milk related QTLs, whereas one lincRNA was within clinical mastitis QTL region. In addition, targeted genes for 10 lincRNAs with the highest fragments per kilobase of transcript per million fragments mapped (FPKM) were predicted by LncTar for forecasting potential biological functions of these lincRNAs. Further analyses indicate involvement of lincRNAs in several biological functions and different pathways.

**Conclusion:** Our study has provided a panoramic view of lincRNAs in bovine mammary glands and suggested their involvement in many biological functions including susceptibility to clinical mastitis as well as milk quality and production. This integrative annotation of mammary gland lincRNAs broadens and deepens our understanding of bovine mammary gland biology.

**Keywords:** Bovine mammary gland, LincRNAs, Quantitative trait loci

## Background

With the rapid development of sequencing technologies, especially the next-generation sequencing, a large number of transcripts have been accumulated through the RNA sequencing technology (RNA-seq). Some of these transcripts are noncoding RNAs (ncRNAs) and they may have diverse functions in biological processes. MicroRNAs, tRNA halves (tiRNAs) and Piwi-interacting RNA (piRNAs) are short ncRNAs with lengths of less than 200 bp, while ncRNAs longer than 200 bp are called long noncoding RNAs (lncRNAs). Based on

locations of lncRNAs, they can be classified as antisense lncRNAs, intronic lncRNAs, bidirectional lncRNAs, intergenic lncRNAs (lincRNAs) and sense-overlapping lncRNAs.

In recent years, a growing number of lncRNAs have been discovered in eukaryotic organisms. In particular, lincRNAs and lincRNA candidates have been catalogued for human, mouse, zebrafish, frog, fly, nematode, arabis, maize, yeast, pig, chicken and plasmodium [1, 2]. These lincRNAs are an important component of the regulatory architecture and they are involved in chromatin modification, epigenetic regulation, genomic imprinting, transcriptional control as well as pre- and posttranslational mRNA processing [1, 3].

Mammary glands have cycles of development and regression including pregnancy, lactation and involution throughout an adult female life. The role of lncRNAs in

\* Correspondence: xin.zhao@mcgill.ca

<sup>1</sup>College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, People's Republic of China

<sup>3</sup>Department of Animal Science, McGill University, 2111 Lakeshore Road, Ste-Anne-De-Bellevue, QC H9X 3V9, Canada

Full list of author information is available at the end of the article



development and differentiation of mammary glands have been studied in a few studies. The pregnancy induced noncoding RNA (*PINC*) is one of first reported lncRNAs having the function in regulating development of mouse mammary epithelial cells. Mouse *PINCL0* and mouse *PINCL6* are two splice forms in mouse mammary epithelial HC11 cells. Knockdown of *PINCL0* led to apoptosis, while knockdown of *PINCL6* induced cell proliferation in HC11 cells, respectively [4]. In addition, *Neat1* (nuclear-enriched abundant transcript 1) is an abundant lncRNA and conserved in the mammalian lineage [5]. It also plays a key role in mammary gland development. Loss of *Neat1* reduced numbers of luminal alveolar epithelial cells and influenced normal mammary gland development in mice [6]. Another lncRNA *Zfas1* was highly expressed in primary mammary epithelial cells from pregnant mice and knockdown of *Zfas1* increased the proliferation rate of cells and induced beta-casein mRNA expression [7].

A few studies have reported existence of bovine lncRNAs, mainly in non-mammary gland tissues. Qu and Adelson identified 12,614 intergenic ncRNAs and 9337 intronic ncRNAs from public bovine Expressed Sequence Tags (ESTs) data [8]. Huang et al. predicted 449 putative lncRNAs which were located in 405 intergenic regions from bovine ESTs [9]. Weikard et al. detected lncRNAs in bovine skin samples (pigmented and non-pigmented) and identified 4848 potential lncRNAs with most of them being classified as lincRNAs [10]. Billerey et al. explored the lincRNA in Limousin bull muscle samples and found 584 different lincRNAs [11]. Finally, Koufariotis et al. catalogued a comprehensive list of putative bovine lncRNA located within intergenic and pseudogene regions which were expressed in 18 tissues including mammary glands [12].

To identify noncoding RNAs and their corresponding genes and to simplify the analysis to avoid the complications arising from overlap with other types of genes, recent focuses have been on lincRNA, which do not overlap exons of either protein-coding or other non-lincRNA types of genes [1]. Up to now, very few studies have specifically profiled lincRNAs in bovine mammary glands. Thus, the focus of our current study was to profile lincRNA transcripts of bovine mammary glands. We assembled transcripts from bovine mammary glands by an integrative approach. Four programs including the Coding-Non-Coding Index (CNCL) [13], the Coding Potential Calculator (CPC) [14], the Coding Potential Assessment Tool (CPAT) [15] and the hmmscan [16] were used to identify lincRNA candidates in mammary gland RNA-seq datasets. Using stringent criteria, 184 lincRNAs were found and expression of six randomly selected lincRNAs was confirmed by real-time PCR. One hundred and twelve novel lincRNAs were identified,

contributing to the lincRNA repertoire. The potential relationships between lincRNAs and QTLs and between lincRNAs and signaling pathways in bovine mammary glands improve our understanding of lincRNAs in milk production and milk quantity.

## Methods

### Databases

Five sets of RNA-seq data were downloaded from the NCBI Sequence Read Archive (SRA) database (Additional file 1). These datasets include two single-read samples and three paired-end samples that were 36 to 100 bp long sequenced on Illumina platforms (~100 million reads for the total). The *Bos taurus* UMD3.1 reference genome FASTA file and the Gene Transfer Format (GTF) file were downloaded from the ensembl website (<http://asia.ensembl.org>). The UniRef90 (UniProt Reference Clusters) database was downloaded from the UniProt website (<http://www.ebi.ac.uk/uniprot/database/download.html>).

### Alignment of RNA-seq reads and assembly of transcripts

The quality control of downloaded RNA sequences was performed by the FastQC software (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>, version 0.11.2). Adaptors were filtered using the Trimmomatic program (<http://www.usadellab.org/cms/?page=trimmomatic>, version 0.33). RNA-seq reads from bovine mammary glands were aligned to the *Bos Taurus* UMD3.1 reference genome with the TopHat2 (version 2.0.12) [17]. Mapped reads were assembled with the Cufflinks (version 2.2.1) [18]. All assembled transcripts were then merged using the cuffcompare (version 2.2.1) [19].

### Identification of putative lincRNAs

There is no generally accepted or standard methodology that allows for easy discovery of lincRNAs. Thus, in order to identify true intergenic lincRNAs and avoid false-positive ones, stringent conditions were applied in this study to filter the mapped reads with the following criteria: (1) Only unknown intergenic transcripts (UITs) were used to identify putative lincRNAs; (2) If UITs had only one exon or the length of UITs was less than 200 bp, they were discarded [20]; (3) The UITs with low expression levels (fragments per kilobase of transcript per million fragments mapped, FPKM <1) and the minimal read coverage threshold of these transcript below three were discarded; (4) The UITs closest to the coding gene less than 1 kb were discarded; (5) To predict coding potentials of the remaining UITs which are not annotated in the bovine genome, four programs including the CNCL, the CPC, the CPAT and the hmmscan were used concurrently. The CPC (version 0.9-r2), a SVM based algorithm (<http://cpc.cbi.pku.edu.cn/>), uses UniRef90 to identify protein-coding UITs or noncoding UITs. We

selected the coding potential score  $> 0$  as the coding UITs and the coding potential score  $< 0$  as the noncoding UITs. The CPAT (version 1.2.2) uses a logistic regression model to assess the coding or noncoding transcripts in our UITs. We downloaded 10,000 bovine known protein sequences from the ensembl website and 10,000 bovine noncoding RNAs from the NONCODE database (version 4) (<http://www.noncode.org/index.php>). These sequences were used for training by CPAT. The software calculated the hexamer tables and a bovine specific logistic regression model was built. Because the coding probability score from the CPAT was different in different species, the cut-off value of 0.348 was chosen for reliability and sensitivity according a previous similar study of cows [11]. The UITs with scores  $< 0.348$  were retained as putative noncoding RNAs. The CNCI (version 2) program profiles the adjoining nucleotide triplets (ANT) to differentiate coding and noncoding sequences. The CNCI software was downloaded from the web (<https://github.com/www-bioinfo-org/CNCI>) and used with the default setting. Finally, all 886 UITs were translated into six possible open reading (ORF) frames. Six possible ORF frames contain three frames for the sense strand and three frames for antisense strand. The six ORF frames were compared against the Gene3D, Pfam, TIGRFAM and Superfamily databases using the hmmscan algorithm. If one or more motifs were found in any of six possible ORF frames, the UIT was considered as a coding UIT and was discarded. Only noncoding UITs identified by all four software programs were considered as our putative lincRNAs.

#### Localization of lincRNAs in quantitative trait loci

The positions of the 184 putative lincRNAs were compared with positions of known quantitative trait loci (QTL) on the *Bos taurus* UMD3.1 reference genome according to the AnimalQTLdb. The AnimalQTLdb is a public QTL database on animal species including cattle, chicken, horse, pig, rainbow trout and sheep (<http://www.animalgenome.org/QTLdb/>) [21].

#### Prediction of lincRNA–RNA interactions and pathway analyses

In order to predict the targets of lincRNAs in mammary glands and thus understand potential functions of lincRNAs, the LncTar tool with the first type of file format for predicting the lincRNA–RNA interactions was used. The first type of file format contains two files. One file was lincRNA sequence file and the other file was mRNA sequence file. During the analysis, only lincRNAs with the lowest normalized free energy  $< -0.14$  were selected as possible lincRNA target genes. Predicted gene targets were used for further analyses of gene ontology (GO) functional annotations and KEGG pathway analysis using the R package clusterProfiler [22].

#### Validation by RT-PCR

The PCR primers for six randomly selected lincRNAs were designed by the Primer Premier 5 (PREMIER Biosoft international, Palo Alto, CA, USA). Primer sequences are in Additional file 2. Expected lengths of PCR products were from 206 to 336 bp. Total RNA was extracted from bovine mammary epithelial cells (MAC-T) by Trizol according to the manufacturer's instructions (Invitrogen, Carlsbad, CA). The first strand of cDNA was synthesized using the PrimeScript™ RT reagent kit (Takara, Dalian, China) according to the manufacturer's instructions. PCR was performed using the  $2 \times$  EasyTaq PCR SuperMix (TransGen Biotech, Beijing, China). The following PCR cycling condition was used:  $94^\circ\text{C}$  5 min, followed by 35 cycles of  $94^\circ\text{C}$  for 30s, annealing for 30s (annealing temperatures for the lincRNAs are in Additional file 2),  $72^\circ\text{C}$  for 30s and a final extension step was  $72^\circ\text{C}$  for 10 min. Five  $\mu\text{l}$  of each PCR product was analyzed by 1% agarose gel.

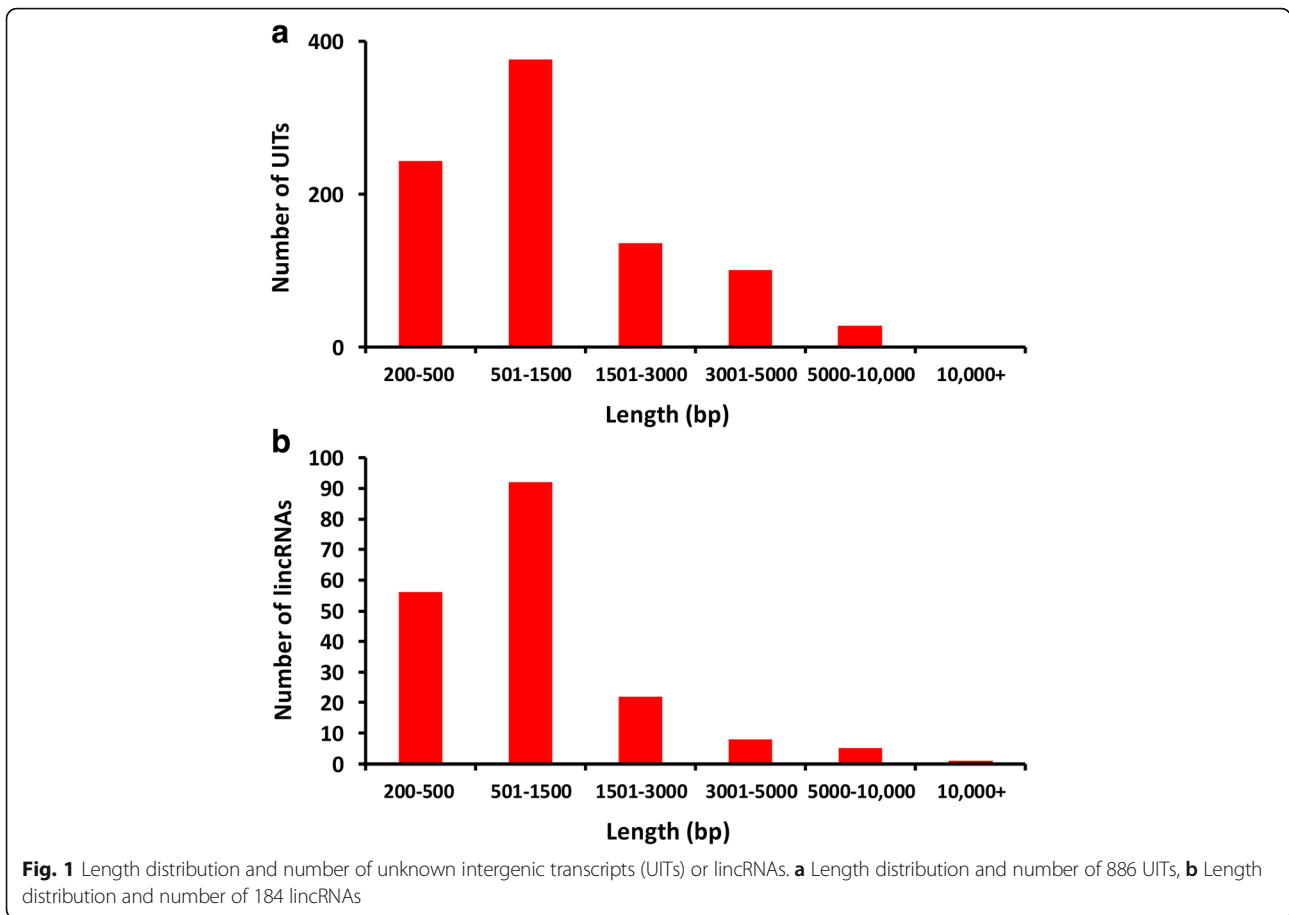
#### Result

##### Alignment of RNA-seq reads and transcriptome assembly

To comprehensively and accurately identify lincRNAs in bovine mammary glands, five un-stranded RNA-seq datasets from bovine mammary glands were used in this study. A total of 104,893,960 reads were obtained after trimming. The reads were aligned to the *Bos taurus* UMD3.1 reference genome using the TopHat2. Transcripts were reconstructed using the Cufflinks. To get a unique set of isoforms for each transcript locus, all reconstructed transcripts were assembled using the Cuffcompare against the *Bos taurus* UMD3.1 reference genome. Consequently, 129,889 transcripts were assembled.

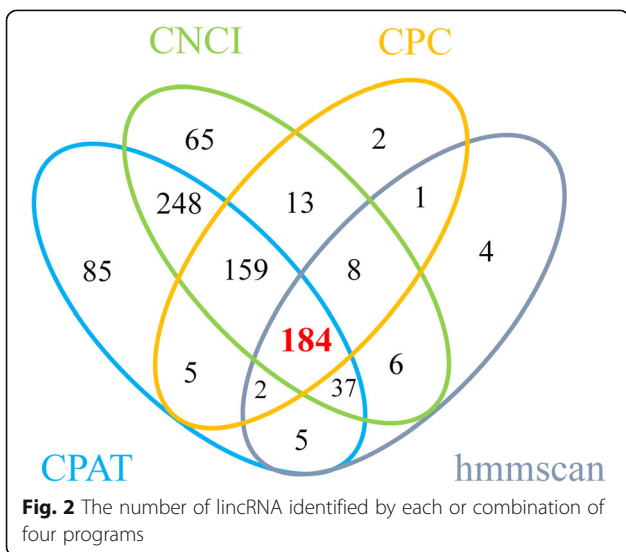
Since reads were not assembled from strand-specific RNA-seq libraries, only UITs were used to identify putative lincRNAs in the bovine mammary glands. Very stringent conditions were used to filter assembled transcripts, as described in the materials and methods. After the filtering, only 886 UITs remained with sizes ranging from 204 to 13,156 bp. Among them, approximately 42% had a length ranging from 501 to 1500 bp (Fig. 1a).

To discriminate noncoding UITs from coding ones, four software programs were applied to predict the protein-coding potential of 886 UITs. The CPC scores resulted in 374 putative noncoding UITs, while calculation by CPAT yielded 725 putative noncoding UITs. The CNCI analysis led to 720 putative noncoding UITs. Finally, 886 UITs were translated into six possible ORFs by the hmmscan algorithm against the Gene3D, Pfam, TIGRFAM and Superfamily databases. Six ORFs led to 247 putative noncoding UITs existing in all six ORFs. Based on the results from all four programs, 184 putative lincRNAs were identified (Fig. 2).



**Sizes and locations of identified lincRNAs**

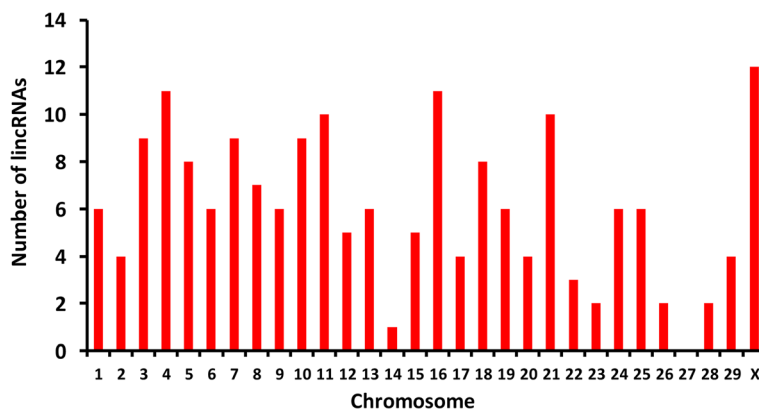
The size of 184 putative lincRNAs ranged from 205 to 13,156 bp (Fig. 1b). The length variation of lincRNAs was similar to that of the UITs. Only one lincRNA (TCONS\_00130959) was more than 10,000 bp long.



The distribution of 184 lincRNAs on each chromosome was plotted in Fig. 3. LincRNAs were present, but not equally, on every chromosome except chromosome 27. The highest numbers of lincRNAs were found on chromosome X (12 lincRNAs), chromosome 4 (11 lincRNAs) and chromosome 16 (11 lincRNAs). Interestingly, two lincRNAs were located in *Bos taurus* unplaced genomic scaffolds. The BLASTN (V2.6.0+) was used to compare our predicted lincRNAs with NON-CODE 2016 and ALDB databases [23, 24]. The standard for sequence similarity was adopted from a previous similar study [10] and was defined with a mapping identity of  $\geq 70\%$  and a total sequence identity of  $\geq 75\%$  in a covered region  $\geq 100$  nt. After the comparison of the two databases, 112 novel lincRNAs were detected (Additional file 3).

**Localizing lincRNAs in QTL-associated regions**

The AnimalQTLdb contains 17,908 cattle QTLs and 514 different cattle traits. By comparing locations of our 184 lincRNAs with QTL positions, 113 different lincRNAs were located in 399 cattle QTLs (Fig. 4a, Additional file 4). For example, one lincRNA was located in clinical mastitis QTLs and 36 lincRNAs in

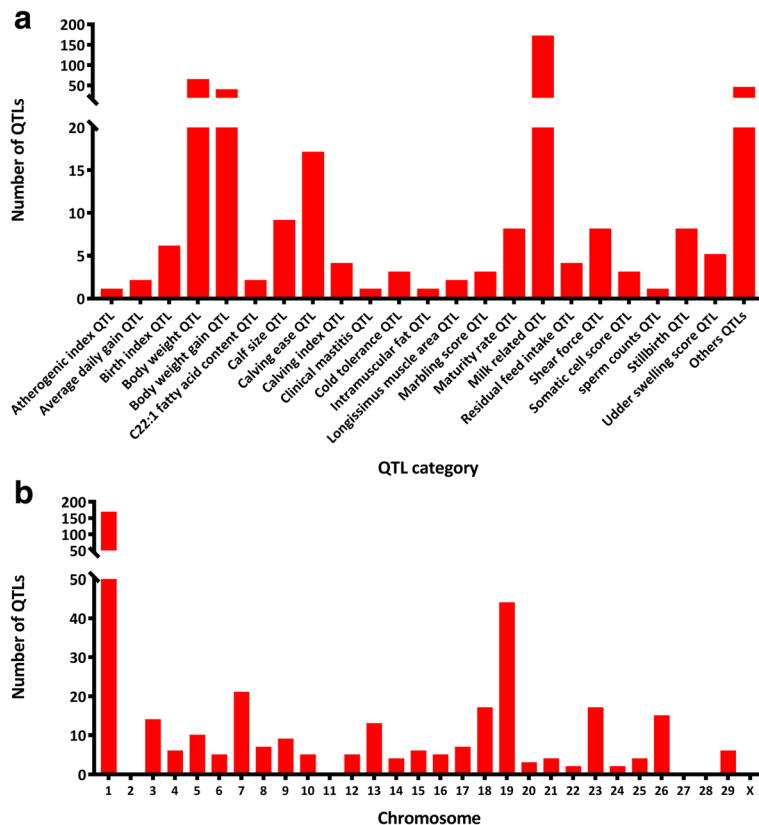


**Fig. 3** The number of 184 lincRNAs detected on each chromosome

169 unique milk related QTLs, suggesting that lincRNAs may be involved in mastitis susceptibility and regulation of milk yield and content (Additional file 4). The highest numbers of QTLs containing lincRNAs were found on chromosome 1 (Fig. 4b).

**Prediction of lincRNA functions**

In addition to the fact that location of a lincRNA within a QTL may suggest the function of the lincRNA in the QTL, we also used the LncTar to predict lincRNA-RNA interactions. To explore the functions of lincRNAs, 10



**Fig. 4** The category for all QTLs or QTLs detected on each chromosome. **a** The category for all QTLs, **b** QTLs detected on each chromosome. Milk related QTLs included Milk butyric acid percentage QTL, Milk capric acid percentage, Milk caprylic acid percentage, Milk caprylic acid percentage, Milk conjugated linoleic acid percentage, Milk decenoic acid percentage QTL, Milk iron content QTL, Milk kappa-casein content QTL, Milk lauric acid percentage QTL, Milk lauroleic acid percentage QTL, Milk myristic acid percentage, Milk oleic acid percentage QTL, Milk palmitic acid percentage QTL, Milk phosphorus content QTL, Milk protein content QTL and Milk yield QTL

lincRNAs with the highest FPKM values were used to identify lincRNA-RNA interactions (Additional file 5). For each lincRNA, predicted target genes with the normalized free energy  $< -0.14$  were identified and annotated by the GO and KEGG. GO functional annotation showed that five lincRNA target genes were enriched in different GO categories, such as biological processes of defense response, RNA phosphodiester bond hydrolysis, endonucleolytic and cAMP-mediated signaling, molecular function of oxidoreductase activity, cytokine receptor binding, protein transporter activity, cellular component of lysosome, membrane region, clathrin vesicle coat and intrinsic component of mitochondrial outer membrane (Additional file 6). For example, the target genes for TCONS\_00162862 were involved in lipid transporter activity and ligase activity and fatty acid transport. On the other hand, target genes for TCONS\_00162906, TCONS\_00174720, TCONS\_00078053, TCONS\_00078723 and TCONS\_00132942 were not enriched in any of the GO categories. As for the pathways, target genes for lincRNAs were enriched in several KEGG pathways such as the MAPK signaling pathway, calcium signaling pathway, glycosaminoglycan biosynthesis, PI3K-Akt signaling pathway, Insulin signaling pathway and Rap1 signaling pathway (Additional file 7). For example, target genes for TCONS\_00010113 were enriched in the PI3K-Akt signaling pathway and Chemokine signaling pathway. Target genes for TCONS\_0007853 were enriched in the Ras signaling pathway, Rap1 signaling pathway, MAPK signaling pathway and AMPK signaling pathway (Fig. 5, Additional file 7). However, target genes for TCONS\_00026311, TCONS\_00078723, TCONS\_

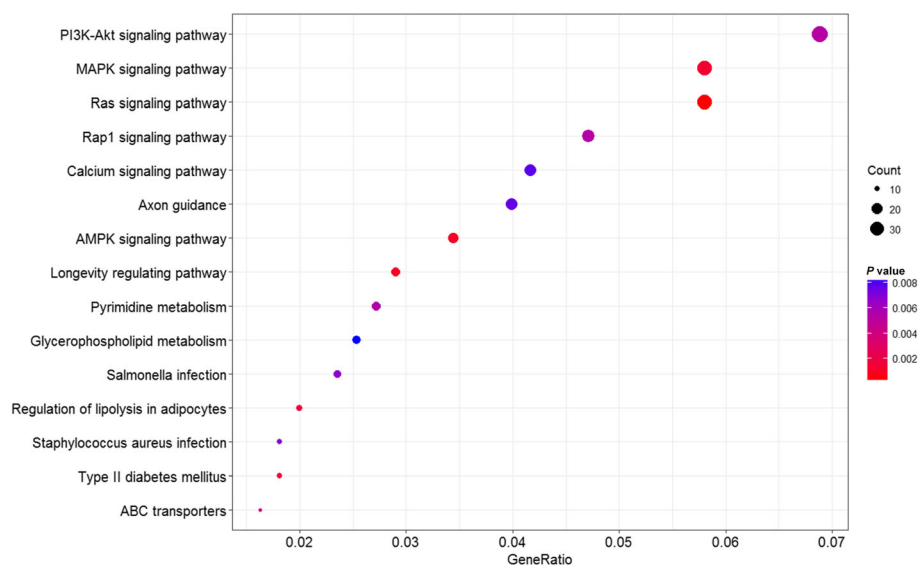
00132942, TCONS\_00174720 and TCONS\_00184509 were not annotated in any of the KEGG pathways.

### Selective validation of novel lincRNAs

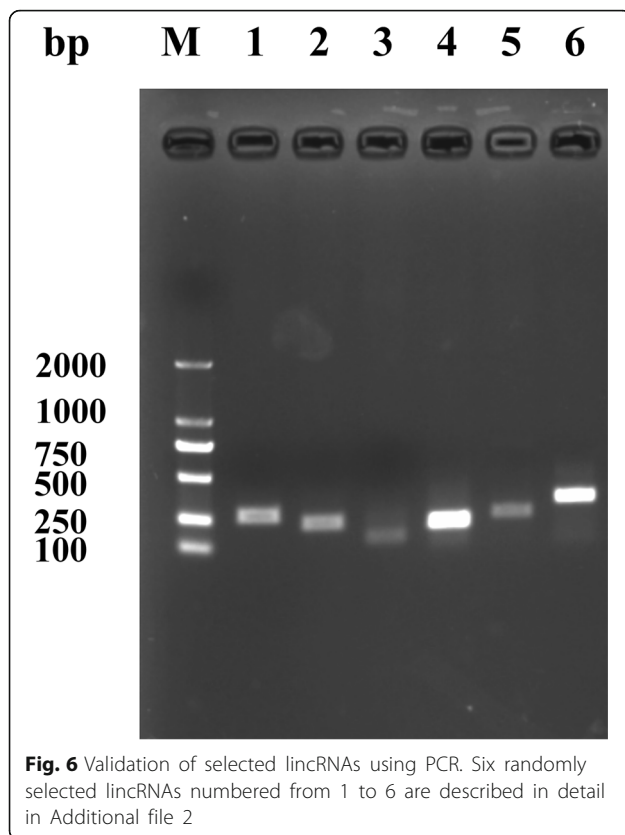
To confirm that our identified bovine lincRNAs were indeed expressed, six lincRNAs were randomly selected for PCR validation in vitro (Fig. 6). All PCR products had the expected size and our selected lincRNAs were amplified, suggesting that most, if not all, of our identified lincRNAs would be expressed in mammary gland tissues.

### Discussion

Long intergenic non-coding RNAs (lincRNAs) are a large and diverse class of transcribed RNA molecules with a length of more than 200 nucleotides that do not encode proteins. In the past few years, an increasing number of lincRNAs have been discovered in eukaryotic organisms, thanks to the advent of powerful and flexible RNA-seq methods. Despite the fact that only a few lincRNAs have been characterized experimentally in detail now, it is already known that they play regulatory and structural roles in almost every important biological process. Previous studies also used EST sequences to identify lincRNAs in cows [8]. However, such studies are incomplete, since ESTs sequences could not represent entire transcriptomes and reflect expression levels. On the other hand, RNA-seq can reveal dynamic transcriptome during different treatments or conditions and evaluate quantitative variations between different samples. Several studies have reported existence of bovine lincRNAs, mostly from non-mammary tissues.



**Fig. 5** The KEGG functional annotation of target genes for TCONS\_0007853. The sizes of the dots represent the counts of genes. The gene ratio indicates the ratio between the number of target genes associated with a KEGG term and the total number of genes in the KEGG term



In this study, we identified 184 bovine lincRNAs based on transcriptomic data from five RNA-seq datasets in bovine mammary gland tissues. The median expressed levels of lincRNA are low and only have about a tenth of the average expressed levels of mRNA [1]. Low expression levels of lincRNAs in transcripts present a challenge to distinguish them from thousands of one exon, low expression fragments and expression noise assembled from RNA-seq results [25]. In order to solve the problem, we have removed RNAs with only a single exon, the minimal read coverage threshold less than 3 or FPKM less than 1. While this process may remove some of lincRNAs with very low expression levels, thousands of unreliable fragments can be removed and only more reliable lincRNAs are kept. Similar strategies have been adopted in other studies. For example, Tang et al. annotated 300 human embryonic stem cell (hES) lincRNAs with FPKM >1 [26]. Similarly, Hangauer et al. selected FPKM > 1 as a criterion when they found tens of thousands lincRNAs in 127 human RNA-seq sequence files [27]. Using the stringent filter method, we identified 886 UITs as potential candidates for lincRNA.

To further identify true lincRNA, four programs including CPAT, CNCI, CPC and hmmscan were used collectively to calculate coding potentials of the 886 UITs. Accurate and quantitative assessment of coding potential is the most critical step in identifying lincRNAs. The

CPC uses a support vector machine (SVM) learning classifier to identify six features extracted from input transcript sequences, while the CPAT is an alignment-free program and has used a logistic regression model to distinguish lincRNAs from noncoding sequences. A bovine special logistic regression model has been used [11]. CNCI is powerful software identifying coding or non-coding transcripts without known annotations. It extracts five features from the transcripts sequences and profiles adjoining nucleotide triplets (ANT) to discriminate coding transcripts from noncoding ones. The Hmmscan can search the lincRNA sequences against the hidden Markov model (HMM) database. We assume that the intersection of four software programs would predict lincRNA results more reliably and accurately than each single program. Consequently, 184 non-protein coding transcripts identified by all four programs were considered as lincRNAs. Approximately 42% of our 184 putative lincRNAs had a length ranging from 501 to 1500 bp, in agreement with those previously reported for other bovine tissues [10]. On the other hand, the number of our identified lincRNA was smaller than previous studies, possibly due to our stringent criteria. We feel that the stringent standards can filter out more false positive lincRNAs and lead to more reliable results.

Identification of lincRNAs in a specific tissue is the first step to understand biological functions of lincRNAs. Up to now, few studies have worked on the functions of lincRNAs in bovine. As a first step, we hypothesized that the location of lincRNAs on the chromosome could be suggestive for potential biological functions. Compared to the position of QTL in AnimalQTLdb, 36 different lincRNAs were found in 172 unique QTL regions associated with milk butyric acid percentage, milk capric acid percentage, milk iron content and milk protein content in milk quality and content. Another TCONS\_00071212 is found in clinical mastitis QTL region (Additional file 4). These lincRNAs deserve particular attention in future studies. In addition, GO and KEGG analyses were performed for target genes of 10 lincRNAs with the highest FPKM values. We choose genes with the normalized free energy less than  $-0.14$  for each of the 10 lincRNAs with the highest FPKM values. The number of genes was chosen to illustrate the concept that GO and KEGG analyses can predict presumptive functions for lincRNAs.

Target genes of TCONS\_00010113 were significantly enriched in cellular component of clathrin coat and clathrin vesicle coat. Clathrin-coated vesicle were presented in the rabbit lactating mammary gland and involved in the biogenesis of casein-containing secretory vesicles and transcytotic pathway [28]. Intriguingly, KEGG pathway annotation showed that target genes for TCONS\_00010113 were enriched in the PI3K-Akt signaling pathway. By blocking the PI3K-Akt signaling

pathway using TK1258, Dey et al. induced apoptosis in breast cancer model cell lines 4T1 [29]. Target genes for TCONS\_0007853 were enriched in Rap1 signaling pathway and MAPK signaling pathway. Itoh et al. found that the Rap1 was a pivotal element in organizing acinar structure and inducing lumen formation in HMT-3522 human mammary epithelial cells. Increased activation of Rap1 induced tumor formation and progression to malignancy [30]. Four classes of MAPK signaling pathway, including the c-Jun N-terminal kinase (JNK) pathway, the ERK5 pathway, extracellular regulated kinase (ERK)1/2 pathway and the p38 pathway, functioned in mammary epithelial cells and involved in breast disease [31]. Based on these published results and the KEGG pathway, it is clear that target genes for TCONS\_0007853 could be involved in development of bovine mammary gland in bovine mammary gland and their roles need to be confirmed.

Analyzing the molecular basis for QTLs is a major challenge. Precisely identifying and verifying causative genes for production traits are still very complicated and difficult, considering that most traits are polygenic. Grisart et al. mapped a QTL region and identified DGAT1 on chromosome 14 as a major gene for milk fat content [32]. However, such a causative relationship is still not available for almost all other traits. Our results reveal that some of the 184 lincRNAs are located within QTL regions. Assuming that some lincRNAs can act on their neighboring gene expression and we can predict functions for lincRNAs accurately, we may be able to use lincRNAs to locate candidate genes for production traits. The *cis* relationship between lincRNAs and their target genes have been reported previously. For example, Faghihi et al. identified an lincRNA called b-secretase-1 antisense transcript (*BACE1-AS*), which regulates *BACE1* mRNA and protein expression *in vitro*. *BACE1* is an important enzyme in Alzheimer's disease [33]. Their results indicate the involvement of *BACE1-AS* in Alzheimer's disease. Similarly, Modarresi et al. found brain-derived neurotrophic factor (BDNF) BDNF, a factor involved in the Huntington's disease, was repressed by an lincRNA BDNF antisense RNA transcript (*BDNF-AS*) [34].

Our current results can provide interesting insight into the interaction between lincRNAs and QTL related marker-assisted selection genes in cows. Of course, finding the relationship between lincRNAs and QTL genes need other techniques. RNA fluorescent *in situ* hybridization (FISH) techniques can enable the absolute measurement of lincRNA transcript abundance and detect precise subcellular localization [35, 36]. Understanding the cellular and molecular mechanism of lincRNAs also relies on identification of RNA-protein and RNA-DNA interactions. RNA immunoprecipitation (RIP) and

ultraviolet cross-linking and immunoprecipitation (CLIP) method are able to identify RNA-protein interactions [37, 38]. Chromatin isolation by RNA purification (ChIRP) and capture hybridization analysis of RNA targets (CHART) can be used to map the DNA binding sites for lincRNAs [39, 40]. These methods as well as more research on lincRNAs will certainly lead to better understanding the mechanism of lincRNAs and guide our future experimental designs for finding causative genes in QTLs for milk yield and milk quality as well as for mastitis resistance.

## Conclusion

In this study, 184 lincRNAs including 112 novel ones were identified in bovine mammary glands. In addition, 113 different lincRNAs were located within 399 unique different cattle QTL regions. In particular, several lincRNAs were placed in QTLs affecting clinical mastitis, milk quality or production and their involvement in these parameters is worthy of further investigation. In addition, the identification of novel lincRNAs significantly expanded the repertoire of lincRNAs for dairy cattle and consequently will help understand functions of the bovine lincRNAs.

## Additional files

**Additional file 1:** RNA-Seq datasets used in this study. (XLSX 10 kb)

**Additional file 2:** Primer sequences used for RT-PCR. (DOCX 14 kb)

**Additional file 3:** List of lincRNA information identified in bovine mammary glands of this study; Additional file 3a. List of 112 novel lincRNA sequences; Additional file 3b. List of 72 known lincRNA sequences, Additional file 3c. GTF file of 184 lincRNAs. (XLSX 114 kb)

**Additional file 4:** List of identified lincRNAs located within known QTL regions. (XLSX 38 kb)

**Additional file 5:** Predicted target genes for 10 lincRNAs with the highest FPKM values; Additional file 5a. Predicted target genes for 10 lincRNAs with the highest FPKM values by LncTar software; Additional file 5b: The neighbouring coding genes (30 K upstream and downstream) of 10 lincRNAs. (XLSX 143 kb)

**Additional file 6:** Gene ontology (GO) functional annotations for target genes of 10 lincRNAs with the highest FPKM values. The *P* value cut-off <0.05 and *q* value cut-off <0.2 were adopted to include the target genes. (XLSX 23 kb)

**Additional file 7:** KEGG pathway annotation results for target genes of 10 lincRNAs with the highest FPKM values. The *P* value cut-off <0.05 and *q* value cut-off <0.2 were adopted to include the target genes. (XLSX 14 kb)

## Abbreviations

FPKM: Fragments per kilobase of transcript per million fragments mapped; GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes; lincRNAs: Intergenic lincRNAs; QTL: Quantitative Trait Loci; UITs: Unknown intergenic transcripts

## Acknowledgements

Not applicable.



**Funding**

This study was financially supported by a grant from National Natural Science Foundation of China (31372282). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Availability of data and materials**

The datasets used during the current study are available in the NCBI the Sequence Read Archive (SRA) database (<https://www.ncbi.nlm.nih.gov/sra/>) under the accession number SRR1449277, SRR1449278, SRR1449280, SRR1639249 and SRR1640194.

**Authors' contributions**

CT and XZ designed the experiments, did the data analysis and wrote the paper. CT and JFM did the PCR. QLC, LLZ and EMI helped the data analysis and contributed to manuscript writing. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**

<sup>1</sup>College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, People's Republic of China. <sup>2</sup>Agriculture and Agri-Food Canada, Research and Development Centre, Sherbrooke, QC J1M 0C8, Canada. <sup>3</sup>Department of Animal Science, McGill University, 21111 Lakeshore Road, Ste-Anne-De-Bellevue, QC H9X 3V9, Canada.

Received: 19 March 2017 Accepted: 12 June 2017

Published online: 19 June 2017

**References**

- Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013;154(1):26–46.
- Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet*. 2016;17(10):601–14.
- Geisler S, Collier J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol*. 2013;14(11):699–712.
- Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, Gonzalez-Rimbau MF, et al. A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc Natl Acad Sci U S A*. 2006;103(15):5781–6.
- Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics*. 2007;8:39.
- Standaert L, Adriaens C, Radaelli E, Van Keymeulen A, Blanpain C, Hirose T, et al. The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA*. 2014;20(12):1844–9.
- Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, Clark MB, et al. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*. 2011;17(5):878–91.
- Qu Z, Adelson DL. Bovine ncRNAs are abundant, primarily intergenic, conserved and associated with regulatory genes. *PLoS One*. 2012;7(8):e42638.
- Huang W, Long N, Khatib H. Genome-wide identification and initial characterization of bovine long non-coding RNAs from EST data. *Anim Genet*. 2012;43(6):674–82.
- Weikard R, Hadlich F, Kuehn C. Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. *BMC Genomics*. 2013;14:789.
- Billerey C, Boussaha M, Esquerre D, Rebours E, Djari A, Meersseman C, et al. Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics*. 2014;15:499.
- Koufariotis LT, Chen YP, Chamberlain A, Vander Jagt C, Hayes BJ. A catalogue of novel bovine long noncoding RNA across 18 tissues. *PLoS One*. 2015;10(10):e0141225.
- Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*. 2013;41(17):e166.
- Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 2007;35:W345–9.
- Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41(6):e74.
- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39:W29–37.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc*. 2012;7(3):562–78.
- Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25(18):1915–27.
- Hu ZL, Park CA, Reecy JM. Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res*. 2016;44(D1):D827–33.
- Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–7.
- Xie CY, Yuan J, Li H, Li M, Zhao GG, Bu DC, et al. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res*. 2014;42(D1):D98–D103.
- Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: a domestic-animal long noncoding RNA database. *PLoS One*. 2015;10(4):e0124003.
- Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22(9):1775–89.
- Tang X, Hou M, Ding Y, Li Z, Ren L, Gao G. Systematically profiling and annotating long intergenic non-coding RNAs in human embryonic stem cell. *BMC Genomics*. 2013;14(Suppl 5):S3.
- Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*. 2013;9(6):e1003569.
- Pauloin A, Tooze SA, Michelutti I, Delpal S, Ollivier-Bousquet M. The majority of clathrin coated vesicles from lactating rabbit mammary gland arises from the secretory pathway. *J Cell Sci*. 1999;112:4089–100.
- Dey JH, Bianchi F, Voshol J, Bonenfant D, Oakeley EJ, Hynes NE. Targeting fibroblast growth factor receptors blocks PI3K/AKT signaling, induces apoptosis, and impairs mammary tumor outgrowth and metastasis. *Cancer Res*. 2010;70(10):4151–62.
- Itoh M, Nelson CM, Myers CA, Bissell MJ. Rap1 integrates tissue polarity, lumen formation, and tumorigenic potential in human breast epithelial cells. *Cancer Res*. 2007;67(10):4759–66.
- Whyte J, Bergin O, Bianchi A, McNally S, Martin F. Key signalling nodes in mammary gland development and cancer. Mitogen-activated protein kinase signalling in experimental models of breast cancer progression and in mammary gland development. *Breast Cancer Res*. 2009;11(5):209.
- Grisart B, Coppeters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res*. 2002;12(2):222–31.
- Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med*. 2008;14(7):723–30.
- Modarresi F, Faghihi MA, Lopez-Toledano MA, Fatemi RP, Magistri M, Brothers SP, et al. Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat Biotechnol*. 2012;30(5):453–9.
- Itzkovitz S, van Oudenaarden A. Validating transcripts with probes and imaging technology. *Nat Methods*. 2011;8:512–9.

36. Chakraborty D, Kappei D, Theis M, Nitzsche A, Ding L, Paszkowski-Rogacz M, et al. Combined RNAi and localization for functionally dissecting long noncoding RNAs. *Nat Methods*. 2012;9(4):360–2.
37. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science*. 2003;302(5648):1212–5.
38. Selth LA, Gilbert C, Svejstrup JQ. RNA immunoprecipitation to determine RNA-protein associations in vivo. *Cold Spring Harb Protoc*. 2009; doi:10.1101/pdb.prot5234.
39. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell*. 2011;44(4):667–78.
40. Simon MD, Wang CI, Kharchenko PV, West JA, Chapman BA, Alekseyenko AA, et al. The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A*. 2011;108(51):20497–502.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

