RESEARCH ARTICLE

ALDB: A Domestic-Animal Long Noncoding RNA Database

Aimin Li[1,2], Junying Zhang[1]*, Zhongyin Zhou[3,4], Lei Wang[2], Yujuan Liu[5], Yajun Liu[1,6]

1 School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, PR China, **2** School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi, PR China, **3** Department of Molecular and Cell Biology, School of Life Sciences, University of Science and Technology of China, Hefei, PR China, **4** State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, PR China, **5** Xi'an DongXing Branch, CMST Development Co. Ltd., Xi'an, Shaanxi, PR China, **6** Higher Technology College, Xi'an University of Technology, Xi'an, Shaanxi, PR China

* jyzhang@mail.xidian.edu.cn

OPEN ACCESS

Citation: Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y (2015) ALDB: A Domestic-Animal Long Noncoding RNA Database. PLoS ONE 10(4): e0124003. doi:10.1371/journal.pone.0124003

Academic Editor: Igor B. Rogozin, National Center for Biotechnology Information, UNITED STATES

Received: November 18, 2014

Accepted: February 25, 2015

Published: April 8, 2015

Data Availability Statement: Data have been deposited to Figshare: http://dx.doi.org/10.6084/m9.figshare.1320987.

Funding: This work was supported by the Natural Science Foundation of China (http://www.nsfc.gov.cn/) under Grants 61070137, 91130006, 61201312 and 61303122; and the Research Fund for the Doctoral Program of Higher Education of China (http://www.moe.edu.cn/, 20130203110017); and the Fundamental Research Funds for the Central Universities of China (http://www.moe.edu.cn/, BDY171416 and JB140306). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Background

Long noncoding RNAs (lncRNAs) have attracted significant attention in recent years due to their important roles in many biological processes. Domestic animals constitute a unique resource for understanding the genetic basis of phenotypic variation and are ideal models relevant to diverse areas of biomedical research. With improving sequencing technologies, numerous domestic-animal lncRNAs are now available. Thus, there is an immediate need for a database resource that can assist researchers to store, organize, analyze and visualize domestic-animal lncRNAs.

Results

The domestic-**a**nimal **l**ncRNA **d**ata**b**ase, named ALDB, is the first comprehensive database with a focus on the domestic-animal lncRNAs. It currently archives 12,103 pig intergenic lncRNAs (lincRNAs), 8,923 chicken lincRNAs and 8,250 cow lincRNAs. In addition to the annotations of lincRNAs, it offers related data that is not available yet in existing lncRNA databases (lncRNAdb and NONCODE), such as genome-wide expression profiles and animal quantitative trait loci (QTLs) of domestic animals. Moreover, a collection of interfaces and applications, such as the Basic Local Alignment Search Tool (BLAST), the Generic Genome Browser (GBrowse) and flexible search functionalities, are available to help users effectively explore, analyze and download data related to domestic-animal lncRNAs.

Conclusions

ALDB enables the exploration and comparative analysis of lncRNAs in domestic animals. A user-friendly web interface, integrated information and tools make it valuable to researchers in their studies. ALDB is freely available from http://res.xaut.edu.cn/aldb/index.jsp.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Noncoding RNAs (ncRNAs) are a large class of RNAs that do not encode proteins. Based on their sequence characteristics and biological functions, ncRNAs can be classified as (1) housekeeping RNAs; (2) small ncRNAs; (3) long noncoding RNAs (lncRNAs). As the advances of sequencing technologies, thousands of lncRNAs have been identified in animals and plants [1–3]. Functional research has shown that lncRNAs play important roles in fundamental biological processes such as dosage compensation, transcriptional regulation and epigenetic regulation [4–6]. To identify lncRNAs cleanly, and to simplify their analysis by avoiding the complications arising from overlap with other types of genes, we here focus on intergenic long noncoding RNAs (lincRNAs), a family of lncRNAs that are located within intergenic regions.

Domestic animals show considerable phenotype diversity due to thousands of years' domestication. Identification of domestic animal lincRNAs has been previously conducted in the gonad transcriptome of two extreme male pigs, chicken skeletal muscle and bovine skin using RNA-seq data [7–10], and also been performed in bovine using EST data [11] (Table A in S1 File). These reports indicate that some lincRNAs present homology with human, and suggest that some lincRNAs serve important roles in muscle development or in the modulation of pigmentation processes [7–10]. Recently, Zhou *et al*. [12] have also carried out the genome-wide identification of pig lincRNAs from RNA-seq and NCBI UniGene data using a computational pipeline with a stringent criterion. Among the lincRNAs identified by them, *linc-sscg2561* exhibits synteny and sequence conservation. Both *linc-sscg2561* and its neighbor *Dnmt3a*, associated with emotional behaviors, show differential expression in the frontal cortex between domesticated pigs and wild boars, suggesting *linc-sscg2561*'s possible role in pig domestication. In this study, we used RNA-seq transcriptome datasets in combination with NCBI UniGene data to identify the lincRNAs of chicken and cow employing a similar method as Zhou *et al*.'s [12]. We built a database, named ALDB, to store and explore lncRNAs.

Compared with existing lncRNA databases, such as NONCODE and lncRNAdb [13,14], ALDB has several distinctive features (Table 1). ALDB provides the expression levels of lncRNA and mRNA genes from various tissues for domestic animals. As lncRNAs are known to possibly interact with chromatin proteins to positively or negatively regulate the expression of neighboring genes [15–17], we also presented the expression of ten protein-coding genes adjacent to lincRNAs. QTLs are important to analyzing putative functions of genes. In ALDB, users can find the lincRNAs that are located in QTL regions. In summary, ALDB provides series of functions to facilitate research on domestic animal lncRNAs.

Materials and Methods

Ethics statement

This study involved analysis of existing, publicly available datasets, therefore no ethics statement is required for this work.

Data acquisition and database content

ALDB currently provides the following data for further domestic-animal lncRNA research: (1) Annotated pig, chicken and cow lincRNAs; (2) Gene (including lincRNA and mRNA, etc.) expression levels in diverse tissue samples; (3) Overlap between lincRNA genes and QTLs; (4) Analysis and statistics on gene expression, QTLs and homology. For the convenience of analysis, we also integrated the Ensembl gene set [25].

Datasets used for lincRNA identification. The annotations of pig lincRNAs with CPC (Coding Potential Calculator) scores <0 were provided by Zhou *et al*. [12]. To identify a more

Table 1. Comparison between ALDB and related databases.

Database	lncRNA*	Expression	QTL	Source	Species	Description	Ref
NONCODE	16188**	Human, mouse	×	Literature	10 species	An integrated knowledge database of non-coding RNAs; http://www.noncode.org/	[13]
lncRNAdb	16	Human, mouse	×	Literature	Around 60 species	A reference database for long noncoding RNAs; http://www.lncrnadb.org/	[14]
PLncDB	0	Arabidopsis	×	Tiling Arrays, RNA-seq	Arabidopsis	A plant long noncoding RNA database; http://chualab.rockefeller.edu/gbrowse2/homepage.html	[18]
NRED	0	Human, mouse	×	Design arrays	Human, mouse	A database of long noncoding RNA expression; http://nred.matticklab.com/cgi-bin/ncrnadb.pl	[19]
LncRNADisease	0	×	×	Literature	Human	A database for long-non-coding RNA-associated diseases;http://www.cuilab.cn/lncrnadisease	[20]
LNCipedia	0	×	×	Literature	Human	A comprehensive compendium of human long non-coding RNAs; http://www.lncipedia.org/	[21]
The Functional lncRNA Database	0	×	×	Literature	Human, mouse, rat	A repository of mammalian long non-protein-coding transcripts that have been experimentally shown to be both non-coding and functional; http://www.valadkhanlab.org/database	[22]
ChIPBase	0	Human	×	ChIP-Seq	6 species	A database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data; http://deepbase.sysu.edu.cn/chipbase/	[23]
starBase	0	×	×	CLIP-Seq	Human, mouse, Caenorhabditis elegans	Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data; http://starbase.sysu.edu.cn/	[24]
ALDB	29276	Pig, chicken, cow	Pig, chicken, cow	RNA-seq, EST	Pig, chicken, cow	A domestic-animal long noncoding RNA database; http://res.xaut.edu.cn/aldb/	

* The number of domestic animal lncRNA transcripts.

** 16,188 chicken lncRNA transcripts were gathered from an unpublished source.

doi:10.1371/journal.pone.0124003.t001

comprehensive set of lincRNAs for chicken and cow, we jointly used both RNA-seq and UniGene (EST) datasets. EST sequences have been proven to be extremely valuable for the identification of lincRNAs in plants and mammals [26,27]. To minimize transcriptional noise, high quality and non-redundancy sequences (UniGene) were used in our study. We collected RNA-seq transcriptome datasets across 4 different experiments from the Sequence Read Archive (SRA)[28–31], and obtained chicken and cow unigenes from the NCBI UniGene database [32] (Tables B, C and D in S1 File).

Ensembl gene set (release 73) and the following genomic sequences, used for alignment, assembly and filtering, were downloaded from the Ensembl Genome Browser [25]: GG4.0 (chicken), UMD3.1 (cow).

Computational pipeline for identifying lincRNAs. LncRNAs can be crudely classified based on their genomic proximity to protein-coding genes, including sense, antisense, bidirectional, intronic or intergenic [33]. To identify lncRNAs neatly by escaping the complications arising from overlap with other types of genes, we currently focus on intergenic lncRNAs (lincRNAs), which lie within the genomic interval between two genes.

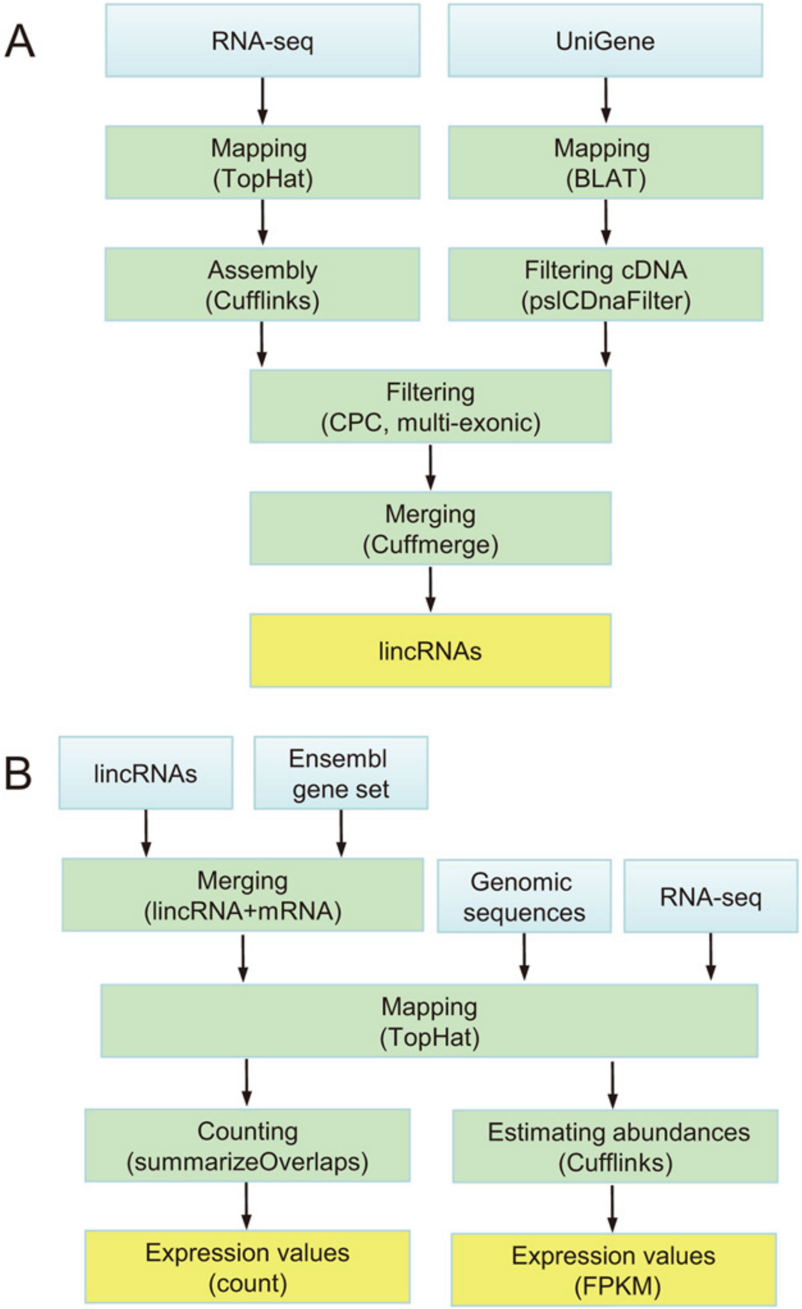


Fig 1. Identification of lincRNAs and calculation of expression levels. (**A**) Computational pipeline to identifying lincRNAs. (**B**) Pipeline of calculation of expression levels for protein-coding and lncRNA genes. Blue box: input data. Green box: processing. Orange box: outcomes.

doi:10.1371/journal.pone.0124003.g001

In this study, the bioinformatics pipeline of identification of lincRNAs was similar to Zhou *et al.*'s except that the cutoff of CPC scores was 0 (in this study), not -1 (in Zhou *et al.*'s study), in order to identify a more thorough lincRNA set [12,34].

Fig 1A highlights the main steps of the pipeline. The approach to extracting high-confidential and reliable transcripts from RNA-seq data differs from that from UniGene data.

For the RNA-seq data, all reads from each dataset were mapped to genomic sequences with an alignment tool TopHat (v2.0.8) [35] (where we set the option of searching novel splice variants be on and we provided gene annotations from Ensembl gene set to maximize spliced alignment accuracy). The resulting aligned data were then separately fed to an assembler Cufflinks (v2.0.2) [35] for their assembly to transcripts and the subsequent transcripts were compared with the Ensembl gene set (release 73) by using the Cuffcompare program in Cufflinks suite (v2.0.2) for the identification of novel transcripts (*class code = "u"*). To reduce transcriptional noise, only those assembled transcript isoforms that had at least 3 reads to support their junctions and whose coverage of exon nucleotides was >80% in at least one sample were retained for further analysis [12].

For the UniGene data, each unigene was first mapped to the genomic sequences using BLAT (the BLAST-like alignment tool) (v35) [36], and then low-quality contigs were removed using the pslCDnaFilter program (http://hgdownload.cse.ucsc.edu/admin/exe/) with the parameters:*-minId = 0.95*, *-minCover = 0.25*, *-globalNearBest = 0.0025*, *-minQSize = 20*, *-minNonRepSize = 16*, *-ignoreNs-bestOverlap*. Contigs that equally mapped to multiple locations of the genomes were removed from the further analysis.

The resulting transcripts supported by UniGene and/or RNA-seq were further filtered as follows:(1) Size selection. The length of a transcript is at least 200 nt. (2) Keeping intergenic lncRNAs. The location of a transcript is at least 500 bp away from the closest annotated genes (Ensembl genes). (3) Coding potential test. The CPC score of a transcript is less than 0 [34]. (4) Retaining multi-exonic transcripts. A transcript has at least 2 exons [12].

Finally, the retained confidential high-quality transcripts were combined by using the Cuffmerge (v2.0.2) program in Cufflinks suite with default parameters to form final non-redundant lincRNA catalog (We used Cuffmerge as follows: *cuffmerge-s genome_file-o directory_of_output_file input_list_file*).

Using the above pipeline, we identified 8,923 chicken lincRNA transcripts from 6,151 loci, and 8,250 cow lincRNA transcripts from 6,213 loci, respectively.

Characterization of expression profiles. In order to profile the expression patterns of the lincRNAs and other Ensembl genes, we derived count and FPKM (fragments per kilobase per million mapped reads) expression levels from RNA-seq data by using TopHat, Cufflinks and summarizeOverlaps [37]. Of the 13 RNA-seq datasets from the SRA and EBI databases (Table C in S1 File), we selected the following RNA-seq datasets which were produced from relatively higher-depth sequencing and contained more tissue samples: ERA178851 (pig, 20 samples, 10 tissues), SRA059960 (chicken, 26 samples, 9 tissues; cow, 27 samples, 9 tissues).

The following approach begins with lincRNAs that we identified, and concludes with gene expression levels (Fig 1B). First, lincRNAs and Ensembl genes were combined to form a novel annotated gene set. Then, RNA-seq reads for each sample were mapped to the genomic sequences with TopHat (with the option—*no-novel-juncs* to skips gene and transcript discovery), and the novel annotated gene set were also fed to TopHat (with the option-*G*). The resulting alignment files were then fed to Cufflinks to quantify gene abundance with the parameter-*G* and output expression values into the file *genes.fpkm_tracking*. For achieving count expression values, we used the summarizeOverlaps function from the GenomicRanges package (version 1.14.4) [37] with the default parameter *Mode = Union*. Reads were counted from a list of Binary Alignment/Map (BAM) files derived from TopHat and count expression values were returned in a matrix file for possible use in further analysis such as those offered in DESeq and edgeR [38,39].

Integration of animal QTLs. QTLs provide useful information to link genome loci with phenotypes [40]. We downloaded 9,862 pig QTLs, 3,919 chicken QTLs and 8,305 cow QTLs from the animal QTL database (release 22) [41]. These QTLs and their overlap with lincRNAs are curated in ALDB and can be conveniently visualized in GBrowse.

Table 2. Description of MySQL database tables.

Table	Description
gene	The lncRNA genes and the Ensembl genes.
transcript	The lncRNA transcripts and the Ensembl transcripts.
expression	The expression levels of all genes.
neighborgenes	The neighboring genes of each gene.
neighborlnc	The neighboring lncRNA genes of each protein-coding gene.
qtl	The QTL data.
overlapqtl	The overlap between genes/transcripts and QTLs.
exon	The start and end locations of each exon.
sequence	The nucleotide sequence of each transcript.

doi:10.1371/journal.pone.0124003.t002

Database construction

Database design and implementation. The web pages of ALDB were built using a combination of Java Server Pages (JSP) and jQuery EasyUI 1.3.5 (http://www.jeasyui.com). They run on a Linux system (Centos 6.4) and are powered by a Tomcat (7.0.53) server. All data were stored in a MySQL (5.1.66) database (http://www.mysql.com), except that the lincRNA and expression data for downloads were stored as files in multiple format (GTF, GFF3, FASTA, etc.).

We created nine MySQL tables for ALDB, and the relationship of these tables is presented in S2 File. The description of these tables is listed in Table 2.

Generic genome browser. We used GBrowse 2.55 [42], a web-based application for displaying genomic annotations and other features, to visualize the genomes, lincRNAs, Ensembl gene set (release 73) and Animal QTLs (release 22). Gbrowse was built by Perl CGI scripts and was powered by an Apache server (2.2.15). In Gbrowse, users can scroll and zoom through arbitrary regions of a genome, enter a region of the genome by searching for a landmark or performing a full text search of all features, and enable and disable tracks. Users can also download GFF3/FASTA files of derived features from GBrowse.

Online BLAST service. To enable sequence homology searches against lincRNAs, we incorporated a standalone BLAST server using NCBI's wwwblast tool [43] (version 2.2.26, http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/wwwblast/), which can be easily accessed through web browsers.

Results

Web interface of ALDB

We created a freely accessible web interface for visiting ALDB. It comprises six core-page sections: “Home”, “GBrowse”, “Search”, “BLAST”, “Analysis” and “Download” (Fig 2).

Home page. Users can access ALDB at http://res.xaut.edu.cn/aldb/index.jsp. The home page contains an introduction to ALDB and provides a sitemap (Fig 2), which gives users an outline of ALDB and is convenient for them to view and retrieve the information that they are interested in.

GBrowse. GBrowse greatly facilitates users to browse any region of interest in pig, chicken and cow genomes. It allows for a comprehensive analysis of the genomic neighborhood of a single gene or transcript. Users can select data source, and search transcripts or genes by their ids or names, or by a genomic region. A variety of track features can be accessed depending on the region’s position on the genomes, including lncRNA genes, protein-coding genes and QTLs (Fig 3A). In viewing the features, users can zooms out or zooms in a region to view the

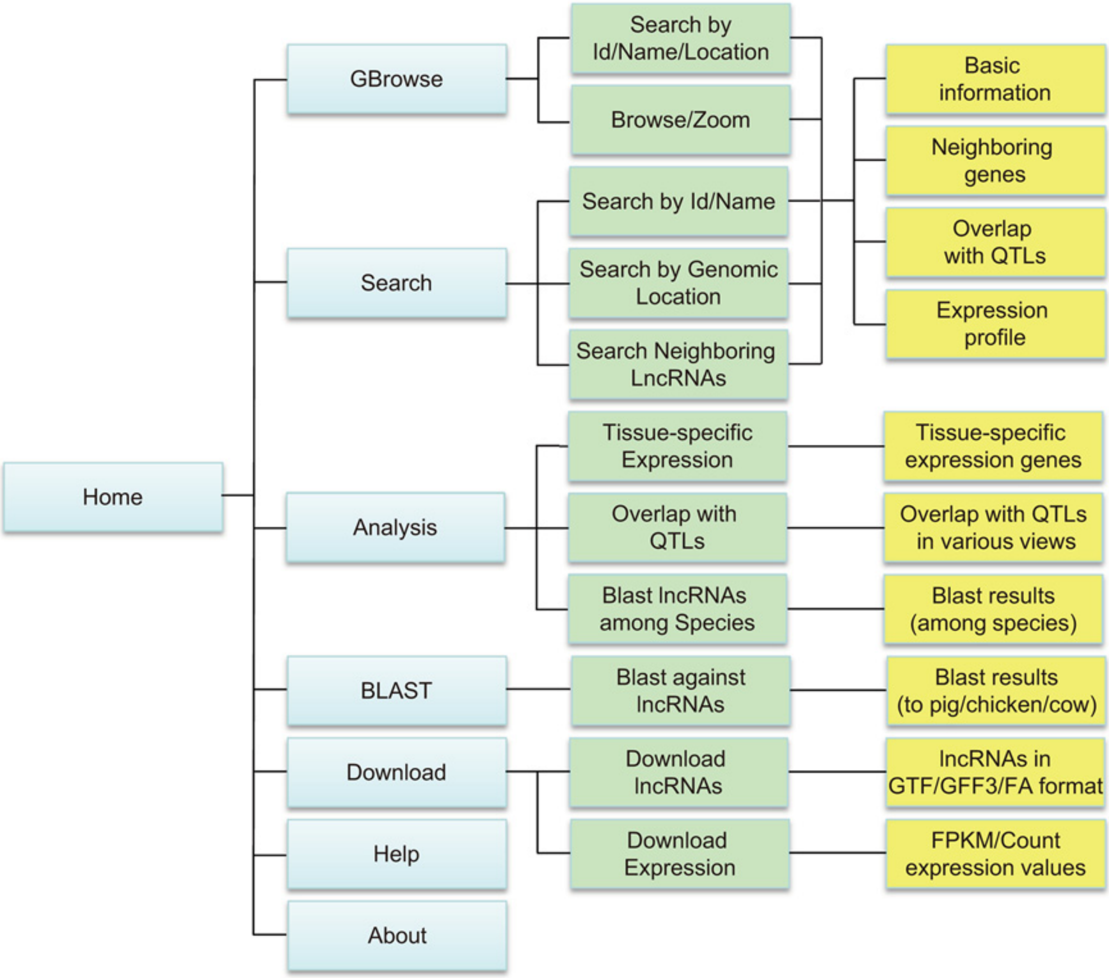
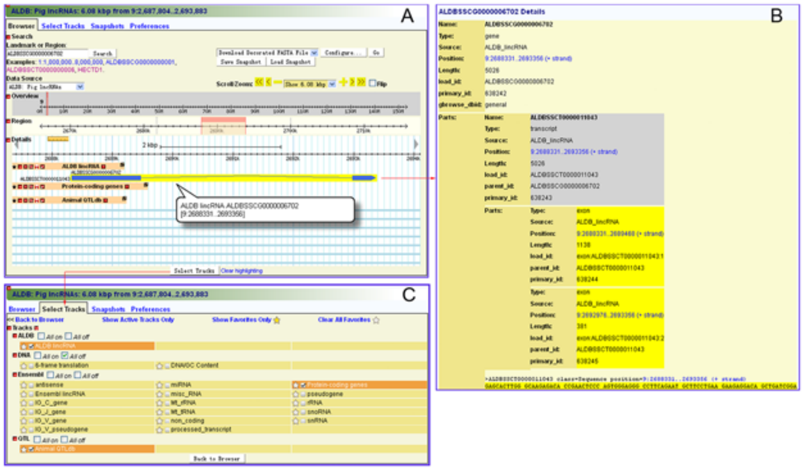

Home
GBrowse
Search
Analysis
BLAST
Download
Help
About
Search by Id/Name/Location
Browse/Zoom
Search by Id/Name
Search by Genomic Location
Search Neighboring LncRNAs
Tissue-specific Expression
Overlap with QTLs
Blast lncRNAs among Species
Blast against lncRNAs
Download lncRNAs
Download Expression
Basic information
Neighboring genes
Overlap with QTLs
Expression profile
Tissue-specific expression genes
Overlap with QTLs in various views
Blast results (among species)
Blast results (to pig/chicken/cow)
lncRNAs in GTF/GFF3/FA format
FPKM/Count expression values

Fig 2. Schematic illustration of ALDB sitemap. Blue box: navigation bar. Green box: processing. Orange box: outcomes.

doi:10.1371/journal.pone.0124003.g002



ALDB: Pig lncRNAs: 6.08 kbp from 9:2,687,804..2,693,883

Landmark or Region: ALDBSSCG0000006702

Examples: 1:1,000,000..8,000,000, ALDBSSCG0000000001, ALDBSSCT0000000006, HECTD1.

Data Source: ALDB: Pig lncRNAs

Download Decorated FASTA File

Show 6.08 kbp

ALDB lincRNA ALDBSSCG0000006702
[9:2688331..2693356]

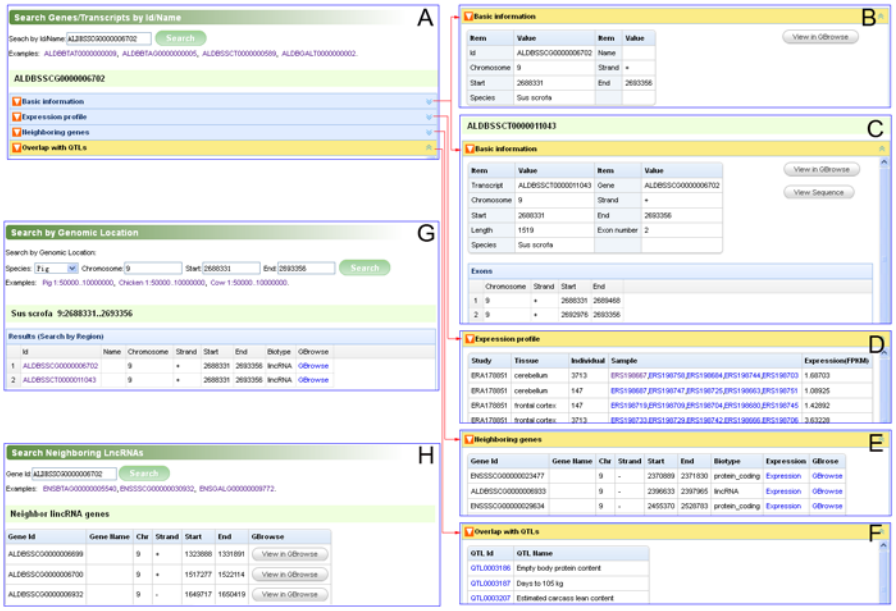
ALDBSSCG0000006702 Details

Name:	ALDBSSCG0000006702		
Type:	gene		
Source:	ALDB_lincRNA		
Position:	9:2688331..2693356 (+ strand)		
Length:	5026		
load_id:	ALDBSSCG0000006702		
primary_id:	638242		
gbrowse_dbid:	general		
Parts:	Name:	ALDBSSCT0000011043	
	Type:	transcript	
	Source:	ALDB_lincRNA	
	Position:	9:2688331..2693356 (+ strand)	
	Length:	5026	
	load_id:	ALDBSSCT0000011043	
	parent_id:	ALDBSSCG0000006702	
	primary_id:	638243	
	Parts:	Type:	exon
		Source:	ALDB_lincRNA
		Position:	9:2688331..2689468 (+ strand)
		Length:	1138
		load_id:	exon:ALDBSSCT0000011043:1
		parent_id:	ALDBSSCT0000011043
		primary_id:	638244
		Type:	exon
		Source:	ALDB_lincRNA
		Position:	9:2692976..2693356 (+ strand)
		Length:	381
		load_id:	exon:ALDBSSCT0000011043:2
		parent_id:	ALDBSSCT0000011043
		primary_id:	638245

>ALDBSSCT0000011043 class=Sequence position=9:2688331..2693356 (+ strand)
GAGCACTTGG GCAAGAGACA CCGAACTCCC AGTGGGAGGG CCTTCAGAAT GCTTCCTGAA GAAGAGGACA GCTGATCGGA

Fig 3. Visualization of genome annotation using GBrowse. (**A**) Users can search and browse features using GBrowse. (**B**) Detail page of a gene. (**C**) Tracks in GBrowse.

doi:10.1371/journal.pone.0124003.g003



Search Genes/Transcripts by Id/Name

Seach by Id/Name: ALDBSSCG0000006702

Examples: ALDBBTAT0000000009, ALDBBTAG0000000005, ALDBSSCT0000000589, ALDBGALT0000000002.

ALDBSSCG0000006702

Basic information

Expression profile

Neighboring genes

Overlap with QTLs

Basic information

Item	Value	Item	Value
Id	ALDBSSCG0000006702	Name	
Chromosome	9	Strand	+
Start	2688331	End	2693356
Species	Sus scrofa		

ALDBSSCT0000011043

Basic information

Item	Value	Item	Value
Transcript	ALDBSSCT0000011043	Gene	ALDBSSCG0000006702
Chromosome	9	Strand	+
Start	2688331	End	2693356
Length	1519	Exon number	2
Species	Sus scrofa		

Exons

	Chromosome	Strand	Start	End
1	9	+	2688331	2689468
2	9	+	2692976	2693356

Expression profile

Study	Tissue	Individual	Sample	Expression(FPKM)
ERA178851	cerebellum	3713	ERS198667,ERS198758,ERS198684,ERS198744,ERS198703	1.68703
ERA178851	cerebellum	147	ERS198687,ERS198747,ERS198725,ERS198663,ERS198751	1.08925
ERA178851	frontal cortex	147	ERS198719,ERS198709,ERS198704,ERS198680,ERS198745	1.42892
ERA178851	frontal cortex	3713	ERS198733,ERS198729,ERS198742,ERS198666,ERS198706	3.63228

Neighboring genes

Gene Id	Gene Name	Chr	Strand	Start	End	Biotype	Expression	GBrose
ENSSSCG00000023477		9	-	2370889	2371830	protein_coding	Expression	GBrowse
ALDBSSCG0000006933		9	-	2396633	2397965	lncRNA	Expression	GBrowse
ENSSSCG00000029634		9	-	2455370	2528783	protein_coding	Expression	GBrowse

Overlap with QTLs

QTL Id	QTL Name
QTL0003186	Empty body protein content
QTL0003187	Days to 105 kg
QTL0003207	Estimated carcass lean content

Search by Genomic Location

Search by Genomic Location:

Species: Pig Chromosome: 9 Start: 2688331 End: 2693356

Examples: Pig 1:50000..10000000, Chicken 1:50000..10000000, Cow 1:50000..10000000.

Sus scrofa 9:2688331..2693356

Results (Search by Region)

	Id	Name	Chromosome	Strand	Start	End	Biotype	GBrowse
1	ALDBSSCG0000006702		9	+	2688331	2693356	lncRNA	GBrowse
2	ALDBSSCT0000011043		9	+	2688331	2693356	lncRNA	GBrowse

Search Neighboring LncRNAs

Gene Id: ALDBSSCG0000006702

Examples: ENSBTAG00000005540, ENSSSCG00000030932, ENSGALG00000009772.

Neighbor lincRNA genes

Gene Id	Gene Name	Chr	Strand	Start	End	GBrowse
ALDBSSCG0000006699		9	+	1323888	1331891	View in GBrowse
ALDBSSCG0000006700		9	+	1517277	1522114	View in GBrowse
ALDBSSCG0000006932		9	-	1649717	1650419	View in GBrowse

Fig 4. Screenshot of the search pages. (**A**) Search for transcripts or genes by their ids or names. (**B**) Basic information for a gene. (**C**) Basic information for a transcript. (**D**) The expression levels of genes in various tissue samples. (**E**) About ten genes adjacent to the current gene are represented. (**F**) The QTLs which overlap with the current gene or transcript. (**G**) Search for transcripts or genes by genomic region. (**H**) Search for neighboring lncRNAs of protein-coding genes.

doi:10.1371/journal.pone.0124003.g004

outline or details of a feature. If the user holds the cursor over a specified gene or transcript, a popup balloon will appear to show gene tracks, symbols and location information. Alternatively, by clicking on a specific gene, users can jump to a new corresponding gene information page to view its more considerable details (Fig 3B). Users can view different elements on different tracks by customizing tracks on the page "Select Tracks" (Fig 3C).

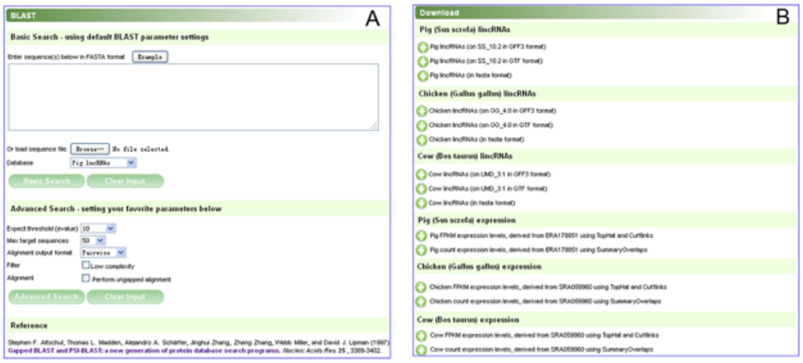
Search. Search functionalities are designed to facilitate users to retrieve useful information, such as the basic information of transcripts and genes, expression profiles, neighboring genes, overlap with QTLs and the lncRNA genes adjacent to a particular protein-coding gene. These functional units are contained in the "Search" section on the main navigation bar. This section mainly includes "Search by Id/Name", "Search by Genomic Location", and "Search Neighboring LncRNAs" pages.

In the "Search by Id/Name" page, users can find transcripts and genes by inputting their ids or names (Fig 4A). For a transcript or gene, the following relative information will be represented: "Basic information" (Fig 4B and 4C), "Expression profile" (Fig 4D), "Neighboring genes" (Fig 4E), "Overlap with QTLs" (Fig 4F). The content panels can be expanded or collapsed when their head titles (inverted triangle icons) are clicked.

The "Search by Genomic Location" page is designed to help users to retrieve all the genes and transcripts located in a specific genomic region (Fig 4G). If users are interested in a transcript or gene, they can click its id in the first column to view its details, or click "GBrowse" in the last column to visualize it in GBrowse.

The "Search Neighboring LncRNAs" page can be used to search the lncRNA genes adjacent to a protein-coding gene (Fig 4H).

BLAST. Since ALDB maintained nucleotide sequence data of lincRNAs, a sequence alignment tool, called wwwblast, based on the Basic Local Alignment Search Tool (BLAST) [43] was integrated into ALDB. Users can compare a query sequence with a library of lincRNA sequences



Basic Search - using default BLAST parameter settings

Enter sequence(s) below in FASTA format

Or load sequence file: No file selected.

Database: Pig lncRNAs

Advanced Search - setting your favorite parameters below

Expect threshold (evalue): 10

Max target sequences: 50

Alignment output format: Pairwise

Filter: □ Low complexity

Alignment: □ Perform ungapped alignment

Reference

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 , 3389-3402.

Pig (Sus scrofa) lincRNAs

- Pig lincRNAs (on SS_10.2 in GFF3 format)
- Pig lincRNAs (on SS_10.2 in GTF format)
- Pig lincRNAs (in fasta format)

Chicken (Gallus gallus) lincRNAs

- Chicken lincRNAs (on GG_4.0 in GFF3 format)
- Chicken lincRNAs (on GG_4.0 in GTF format)
- Chicken lincRNAs (in fasta format)

Cow (Bos taurus) lincRNAs

- Cow lincRNAs (on UMD_3.1 in GFF3 format)
- Cow lincRNAs (on UMD_3.1 in GTF format)
- Cow lincRNAs (in fasta format)

Pig (Sus scrofa) expression

- Pig FPKM expression levels, derived from ERA178851 using TopHat and Cufflinks
- Pig count expression levels, derived from ERA178851 using SummaryOverlaps

Chicken (Gallus gallus) expression

- Chicken FPKM expression levels, derived from SRA059960 using TopHat and Cufflinks
- Chicken count expression levels, derived from SRA059960 using SummaryOverlaps

Cow (Bos taurus) expression

- Cow FPKM expression levels, derived from SRA059960 using TopHat and Cufflinks
- Cow count expression levels, derived from SRA059960 using SummaryOverlaps

Fig 5. Screenshot of the BLAST and download pages. (A) The BLAST page. **(B)** The download page.

doi:10.1371/journal.pone.0124003.g005

using a BLAST search, and identify lincRNA sequences that closely resemble the query sequence. BLAST will meet users' requirements for finding homologous transcripts or genes of interest. On the BLAST page (Fig 5A), users can supply one or more query sequences by uploading or directly pasting them to search against the available databases using BLAST's default parameters. Users can also specify more parameters for a BLAST search to control its search sensitivity and result format, etc. The BLAST results will be displayed in another page in a pairwise format by default.

Analysis. Tissue-specific expression of genes tend to provide evidence for their function roles[44]. Users can find tissue-specifically expressed genes in the "Tissue-specific Expression" page, which were derived by using DESeq2 with the parameter *padj<0.1* according to their expression across various tissue samples.

The associations between lincRNAs and QTLs are presented in several forms: in trait hierarchies, by genes, by traits and by QTLs. These pages provide users with various information from different perspectives.

We blasted lncRNAs against one another among five species: pig, chicken, cow, human and mouse. Those transcripts with identical sequences and genomic locations were marked as "Identical" entries, and those with similar sequence information (query coverage>80%, subject coverage>80%, identity>0.8, e-value<1e-10) as "Similar" entries.

Downloads. The data of ALDB can be downloaded to perform local analysis. In the download page, users can download lincRNAs in FASTA, GFF3 and GTF formats. Data for downloads are divided per species and per content type, with information provided for the assembly of the species (Fig 5B).

We also provide the expression levels of lincRNA and Ensembl genes for downloads, which we derived from the datasets ERA178851 and SRA059960. The FPKM expression levels were obtained by TopHat (with the parameter—*no-novel-juncs*) and Cufflinks (with the parameter-*G*). The count expression levels were generated by summaryOverlaps with the parameter *Mode = "Union"*.

An up-to-date algorithm CNCI (Coding-Non-Coding Index) [45] is effective to differentiate coding and non-coding transcripts. For all the lincRNA transcripts archived in ALDB, we provided their CNCI scores [45] on the "Download" page of ALDB. Users can download these data and perform further analysis of lincRNA transcripts.

Case study

In Zhou *et al*.'s study [12], they found 30 pig lincRNA genes significant differentially expressed between the brains of domesticated pigs and wild boars. Of those genes, we arbitrarily take *linc-sscg4116* (chr9:2688331–2693356) as an example to demonstrate the usage of ALDB.

First, find a gene identical to or related to *linc-sscg4116* in ALDB. Get the genomic location of *linc-sscg4116* from the GTF file which is released by Zhou *et al.* [12]. In the "Search by Genomic Location" page, select the "Pig" option and enter the chromosome, start and end locations of *linc-sscg4116*. After clicking the "Search" button, the result page displays a gene (*ALDBSSCG0000006702*) and a transcript (*ALDBSSCT0000011043*) (Fig 4G). Click "*ALDBSSCT0000011043*" and its details will be displayed (Fig 4C). Comparing *linc-sscg4116* with these details, we find they are identical. Therefore, *linc-sscg4116* is same as *ALDBSSCG0000006702* in ALDB.

Second, in the pig database of GBrowse, enter "*ALDBSSCT0000011043*" in the "Landmark or Region" text box and click the search button (Fig 3A). Neighboring transcripts, genes and overlapped QTLs can be visualized. Click *ALDBSSCT0000011043*, and the details of *ALDBSSCT0000011043* will be displayed in a new window (Fig 3B). In addition, the nucleotide sequence of *ALDBSSCT0000011043* and its genomic structure are available in this window (Fig 3B).

Third, in the "Search by Id/Name" page, enter *ALDBSSCT0000011043* and click the search button. The basic information of the transcript *ALDBSSCT0000011043* and overlap with QTLs will be presented (Fig 4C). Search for the gene *ALDBSSCG0000006702*, its expression levels in 10 tissues of two individuals can be available in the "Expression profile" panel (Fig 4D). Also, its neighboring genes will be displayed in the "Neighboring genes" panel (Fig 4E).

Fourth, in the "BLAST" page, enter the nucleotide sequence in the big text box and select "Cow lncRNAs" or "Chicken lncRNAs" in the database drop box (Fig 5A). After clicking the "Basic Search" button, we can check if it is homogeneous with cow or chicken lncRNAs.

Finally, in the "Download" page, the FASTA, GFF3 and GTF files of lincRNAs can be downloaded (Fig 5B). We can extract genomic structure and nucleotide sequences from these files. Also, we can download expression measures in count or FPKM format. These expression data can be supplied to DESeq, edgeR and so on [38,39], to identify differentially expressed genes, or can be used for co-expression analysis.

Discussion

LincRNA genes have been identified to be associated with domestication and diseases [12,46]. Pig, chicken and cow are important domesticated animals and animal model in medical research. Therefore, it is necessary to identify lincRNA genes of these animals and construct a database to store them. To control the quality of transcripts, we integrated RNA-seq datasets and EST (UniGene) with a strict strategy to identify lincRNA genes. In pig, chicken and cow, we found 7,381, 6,151 and 6,213 lincRNA genes, respectively. To facilitate the study of these lincRNA genes, we built a database to store them and integrated expression levels and QTLs. Previous reports indicate that numerous lincRNA genes show significant tissue preferable expression and possibly interaction with chromatin proteins to positively or negatively regulate the expression of neighboring genes [15–17]. Thus, we presented tissue-specific genes, the expression levels of each lincRNA gene and its ten neighboring protein-coding genes in various tissues for pig, chicken and cow. These will help the function research of lincRNA genes in domesticated animals and the understanding of the domestication of pig, chicken and cow. Animal QTL database stores thousands of QTLs that are associated with the phenotypes of pig, chicken and cow. In this study, we found that most lincRNA genes were located in QTLs. Zhou *et al.* proposed that lincRNAs may contribute to the change of behavior, adipogenesis and muscle development in pigs [12] (Zhou *et al.* unpublished). Thus, to facilitate the trait research in lincRNA genes of domesticated animals, we integrated QTL datasets into ALDB. Users can conduct the analysis of lincRNAs' overlap with QTLs and visualize the QTL datasets in GBrowse. Users can also integrate the expression of lincRNA genes, the expression of their neighbor protein-coding genes and QTLs to help the function research in domesticated animals. From the above, we can conclude that ALDB is a useful database for the function research of lincRNA genes in domesticated animals.

Conclusions

The importance of domestic animals is unquestionable due to their global significance for agriculture and their particular relevance to biomedical research [47,48]. Recent reports suggest that lincRNAs have been implicated in muscle development, the modulation of pigmentation processes and domestication of domestic animals [7–10]. In this study, we constructed a unique and comprehensive database of domestic-animal lncRNAs, named ALDB, which currently embraces thousands of domestic-animal lincRNAs and other related data as well as several useful analysis tools.

12,103 pig lincRNAs, 8,923 chicken lincRNAs, and 8,250 cow lincRNAs were derived from a combination of RNA-seq and UniGene data. These lincRNAs were stored and well organized in ALDB. Domestic-animal lincRNA related data are also offered, such as expression levels and QTLs. Additionally, ALDB incorporates flexible search functionalities and integrates several analysis tools, such as BLAST and GBrowse.

In summary, ALDB provides a comprehensive ongoing platform specialized for domestic-animal lncRNA research. We hope that ALDB will become a valuable resource for the study of domestic animals, agriculture development and human diseases.

Availability and Future Developments

ALDB is publicly available at http://res.xaut.edu.cn/aldb/index.jsp. ALDB supports all major latest web-browsers, preferably Mozilla Firefox, Google Chrome or Apple Safari for visualization and performance purposes. The ALDB will continue to grow with the further addition of data, analytical tools and other functionalities. The ALDB will be ongoing updated when novel lincRNAs of other domestic animals, for example, dog and horse are available. Currently, we identified and curated only intergenic lncRNAs. In future, we plan to incorporate other kinds of lncRNA datasets, such as intronic lncRNAs and natural antisense lncRNAs.

Supporting Information

S1 File. Datasets used in identification of domestic-animal lncRNAs. Table A—Data used in previous domestic-animal lncRNA research. Table B—Summary statistics for all RNA-seq data used in this study. Table C—RNA-seq data used in this study. Table D—UniGene datasets used in this study.
(PDF)

S2 File. Entity-relation diagram of the ALDB database.
(PDF)

Acknowledgments

We thank Prof. Junying Zhang's team members for technical discussions and helpful comments. We also thank the anonymous reviewers of the manuscript for their helpful comments and suggestions.

Author Contributions

Conceived and designed the experiments: JZ AL. Performed the experiments: AL ZZ LW. Analyzed the data: AL ZZ Yajun Liu. Contributed reagents/materials/analysis tools: AL Yujuan Liu. Wrote the paper: JZ AL.

References

1. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell 2013; 154: 26–46. doi: 10.1016/j.cell.2013.06.020 PMID: 23827673
2. Li L, Eichten SR, Shimizu R, Petsch K, Yeh C-T, Wu W, et al. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol 2014; 15: R40. doi: 10.1186/gb-2014-15-2-r40 PMID: 24576388
3. Wang H, Chung PJ, Liu J, Jang I-C, Kean M, Xu J, et al. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome Res 2014: gr. 165555.165113.
4. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 2006; 312: 1653–1655. PMID: 16778056
5. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell 2010; 143: 46–58. doi: 10.1016/j.cell.2010.09.001 PMID: 20887892
6. Lee JT. Epigenetic regulation by long noncoding RNAs. Science 2012; 338: 1435–1439. doi: 10.1126/science.1231776 PMID: 23239728
7. Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics 2011; 12: 552. doi: 10.1186/1471-2164-12-552 PMID: 22067327
8. Li T, Wang S, Wu R, Zhou X, Zhu D, Zhang Y. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics 2012; 99: 292–298. doi: 10.1016/j.ygeno.2012.02.003 PMID: 22374175
9. 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. doi: 10.1186/1471-2164-14-789 PMID: 24225384
10. Qu Z, Adelson DL. Bovine ncRNAs are abundant, primarily intergenic, conserved and associated with regulatory genes. PLoS One 2012; 7: e42638. doi: 10.1371/journal.pone.0042638 PMID: 22880061
11. 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: 674–682. doi: 10.1111/j.1365-2052.2012.02325.x PMID: 22497321
12. Zhou ZY, Li AM, Adeola AC, Liu YH, Irwin DM, Xie HB, et al. Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs. Genome Biol Evol 2014; 6: 1387–1392. doi: 10.1093/gbe/evu113 PMID: 24891613
13. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, et al. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res 2014; 42: D98–D103. doi: 10.1093/nar/gkt1222 PMID: 24285305
14. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 2011; 39: D146–151. doi: 10.1093/nar/gkq1138 PMID: 21112873
15. 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: 1915–1927. doi: 10.1101/gad.17446611 PMID: 21890647
16. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet 2012; 8: e1002841. doi: 10.1371/journal.pgen.1002841 PMID: 22844254
17. Ørom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. Cell 2013; 154: 1190–1193. doi: 10.1016/j.cell.2013.08.028 PMID: 24034243
18. Jin J, Liu J, Wang H, Wong L, Chua NH. PLncDB: plant long non-coding RNA database. Bioinformatics 2013; 29: 1068–1071. doi: 10.1093/bioinformatics/btt107 PMID: 23476021
19. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res 2009; 37: D122–126. doi: 10.1093/nar/gkn617 PMID: 18829717

20. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 2013; 41: D983–986. doi: 10.1093/nar/gks1099 PMID: 23175614

21. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 2013; 41: D246–251. doi: 10.1093/nar/gks915 PMID: 23042674

22. Niazi F, Valadkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. RNA 2012; 18: 825–843. doi: 10.1261/rna.029520.111 PMID: 22361292

23. Yang J-H, Li J-H, Jiang S, Zhou H, Qu L-H. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res 2012.

24. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein—RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014; 42: D92–D97. doi: 10.1093/nar/gkt1248 PMID: 24297251

25. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res 2014; 42: D749–D755. doi: 10.1093/nar/gkt1196 PMID: 24316576

26. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 2010; 16: 1478–1487. doi: 10.1261/rna.1951310 PMID: 20587619

27. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, et al. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 2012; 24: 4333–4345. doi: 10.1105/tpc.112.102855 PMID: 23136377

28. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al. A comparison of brain gene expression levels in domesticated and wild animals. PLoS Genet 2012; 8: e1002962. doi: 10.1371/journal.pgen.1002962 PMID: 23028369

29. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature 2011; 478: 343–348. doi: 10.1038/nature10532 PMID: 22012392

30. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 2012; 338: 1593–1599. doi: 10.1126/science.1228186 PMID: 23258891

31. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science 2012; 338: 1587–1593. doi: 10.1126/science.1230612 PMID: 23258890

32. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2013; 41: D8–D20. doi: 10.1093/nar/gks1189 PMID: 23193264

33. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell 2009; 136: 629–641. doi: 10.1016/j.cell.2009.02.006 PMID: 19239885

34. 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: W345–349. PMID: 17631615

35. 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: 562–578. doi: 10.1038/nprot.2012.016 PMID: 22383036

36. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002; 12: 656–664. PMID: 11932250

37. Aboyoun P, Pages H, Lawrence M (2013) GenomicRanges: Representation and manipulation of genomic intervals.

38. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010; 11: R106. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621

39. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26: 139–140. doi: 10.1093/bioinformatics/btp616 PMID: 19910308

40. Consortium CT. The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet 2003; 4: 911. PMID: 14634638

41. Hu ZL, Park CA, Wu XL, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res 2013; 41: D871–879. doi: 10.1093/nar/gks1150 PMID: 23180796

42. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, et al. The generic genome browser: a building block for a model organism system database. Genome Res 2002; 12: 1599–1610. PMID: 12368253

43. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389–3402. PMID: 9254694

44. Warren ST, Schalling DEHcM. Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. Nat Genet 1993; 3.

45. 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: e166. doi: 10.1093/nar/gkt646 PMID: 23892401

46. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011; 29: 742–749. doi: 10.1038/nbt.1914 PMID: 21804560

47. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics 2004; 5: 202–212. PMID: 14970822

48. Roberts R, Smith G, Bazer F, Cibelli J, Seidel G, Bauman D, et al. Farm animal research in crisis. Science 2009; 324: 468. doi: 10.1126/science.1168521 PMID: 19390030