Comprehensive Identification and Annotation of Non-protein-coding Transcriptomes from Vertebrates Indicates Most ncRNAs are Regulatory

Zhipeng Qu

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics

School of Molecular and Biomedical Science

The University of Adelaide

October 2012

Table of Contents

Abstract

Non-coding RNAs (ncRNAs), in particular long ncRNAs, represent a significant proportion of the vertebrate transcriptome and probably regulate many biological processes. Initially, I developed a robust pipeline for the genome wide identification and annotation of ncRNAs and used publically available bovine ESTs (Expressed Sequence Tags) from many developmental stages and tissues as input. The pipeline yielded 23,060 annotated bovine ncRNAs, the majority of which (57%) were intergenic, and were only moderately correlated with protein coding genes. I then used this pipeline to annotate ncRNAs from human, mouse and zebrafish ESTs. Comparative analysis confirmed some previously described findings about intergenic ncRNAs, such as a positionally biased distribution with respect to regulatory or development related protein-coding genes, and weak but clear sequence conservation across species. Furthermore, comparative analysis of developmental and regulatory genes proximate to long intergenic ncRNAs indicated that the relationship of these genes to neighbor long ncRNAs was not conserved, providing evidence for the rapid evolution of species-specific gene associated long ncRNA. In addition, I built protein-coding and non-protein-coding gene co-expression networks based on available human transcriptome data. More than 30,000 human protein-coding and non-coding transcripts were annotated into tissue-specific co-expression sub-networks, indicating the possible regulatory connections between ncRNAs and protein-coding genes. In conclusion, I have reconstructed and annotated over 130,000 long ncRNAs, most of which are un-annotated, in human, mouse and zebrafish. Together with the annotated bovine ncRNAs, we provide a significantly expanded number of candidates for functional testing by the research community.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Zhipeng Qu and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed..... *Date*.....

List of Publications

Qu Z and Adelson DL (2012) Evolutionary conservation and functional roles of ncRNA. Front. Gene. 3:205. doi: 10.3389/fgene.2012.00205

Qu Z, Adelson DL (2012) Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes. PLoS ONE 7(8): e42638. doi:10.1371/journal.pone.0042638

Qu Z, Adelson DL (2012) Identification and Comparative Analysis of ncRNAs in Human, Mouse and Zebrafish Indicate a Conserved Role in Regulation of Genes Expressed in Brain. PLoS ONE 7(12): e52275. doi:10.1371/journal.pone.0052275

Acknowledgements

I would like to express my sincere gratitude to the following people:

Dave Adelson, for supervision and guidance, and for giving me the opportunity and support to do the PhD. I am so lucky to have such a nice supervisor. Dave, the knowledge that I learned from you is not just valuable for my PhD, and will support me for my whole academic career.

Jack Da Silva, my co-supervisor, for providing me the machine and office for the first several months of my PhD.

Joy Raison, for helping me to generate nice figures; Dan Kortschak, who helped me to read manuscripts and always provided valuable suggestions, and all other members of the Adelson lab, past and present, for making it such a supportive and enjoyable place to work.

Dong Wang in the Timmis lab, and everyone else in the Genetics Discipline and MLS building who has helped me along this journey.

Sandy McConachy, Richard Russell and Iris Liu, for always giving me support and advices in both research and life in Adelaide through this long journey.

My parents, for always encouraging and supporting me throughout my PhD, and for their endless love throughout my life. My sister and brother who have also always given me support and encouragement.

I would like to also specially thank the China Scholarship Council (CSC) and The University of Adelaide, who provided me this scholarship to support my PhD.

Chapter 1 Introduction

Evolutionary Conservation and Functional Roles of ncRNA

Zhipeng Qu and David L. Adelson

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

Frontiers in Genetics, October 2012, **3**:205, doi:10.3389/fgene.2012.00205

STATEMENT OF AUTHORSHIP

Evolutionary Conservation and Functional Roles of ncRNA

Frontiers in Genetics, 2012-October-9, **3**:205, doi:10.3389/fgene.2012.00205

Zhipeng Qu (Candidate)

Wrote the manuscript.

I hereby certify that the statement of contribution is accurate

Signed..... *Date*.....

David L. Adelson

Supervised development of work and assisted in writing the manuscript.

I hereby certify that the statement of contribution is accurate and I give permission for inclusion of the paper in the thesis

Signed..... *Date*.....

frontiers in GENETICS

REVIEW ARTICLE
published: 09 October 2012
doi: 10.3389/fgene.2012.00205

Evolutionary conservation and functional roles of ncRNA

Zhipeng Qu and David L. Adelson*

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

Edited by:
Peng Jin, Emory University, USA

Reviewed by:
Olivier Bensaude, École Normale Supérieure, France
Preethi Herat Gunaratne, University of Houston, USA

****Correspondence:***
David L. Adelson, School of Molecular and Biomedical Science, The University of Adelaide, North Terrace, Adelaide, SA, Australia.
e-mail: david.adelson@adelaide.edu.au

Non-coding RNAs (ncRNAs) are a class of transcribed RNA molecules without protein-coding potential. They were regarded as transcriptional noise, or the byproduct of genetic information flow from DNA to protein for a long time. However, in recent years, a number of studies have shown that ncRNAs are pervasively transcribed, and most of them show evidence of evolutionary conservation, although less conserved than protein-coding genes. More importantly, many ncRNAs have been confirmed as playing crucial regulatory roles in diverse biological processes and tumorigenesis. Here we summarize the functional significance of this class of "dark matter" in terms its genomic organization, evolutionary conservation, and broad functional classes.

Keywords: ncRNA, transcription, genetic, long ncRNA, evolution, molecular, gene regulation

INTRODUCTION

As the basis of genetics, the "central dogma" describes the genetic information flow of life (Crick, 1970). The functional roles of DNA as the repository genetic information, and protein as the functional incarnation of that information, have been viewed as the dominant molecular roles in the cell for nearly four decades, while RNA was subordinated as a temporary intermediate of this information flow. However, the hypothesis of an "RNA world" proposed by Gilbert (1986) challenged the "central dogma" view of the biological role of RNA. The RNA world theory proposed that the origin of life is based on RNA, which could both store genetic data and carry out functions such as catalysis. Although the RNA world hypothesis is debated, a hidden "RNA regulatory world" has been proposed in recent studies describing non-coding RNAs (ncRNAs). Thousands of pervasively transcribed ncRNAs have been identified in human, mouse, and other species. Furthermore, these ncRNAs also show clear evolutionary conservation. Many ncRNAs, especially recently identified long ncRNAs, have been shown to play key regulatory roles in diverse biological processes, including pathological processes such as tumorigenesis.

DEFINITION AND CLASSIFICATION OF ncRNAs

Previously ncRNA has been referred to by various names: non-protein-coding RNAs (npcRNAs; Mattick, 2003), intronic and intergenic ncRNAs (Louro et al., 2009), and mRNAs-like ncRNAs (Inagaki et al., 2005; Rymarquis et al., 2008). At present, ncRNAs are classified on the basis of their main functions: structural ncRNAs and regulatory ncRNAs (Mattick and Makunin, 2006). Structural ncRNAs include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), spliceosomal uRNAs (snRNAs), and snoRNAs. Most of these ncRNAs have well-established structural functions. ncRNAs with regulatory roles in gene expression are classified as regulatory ncRNAs, including small interfering RNA (siRNA), micro-RNAs (miRNAs), piwi-RNAs (piRNAs), long ncRNAs, and long intergenic ncRNAs.

Numerous studies in the past decade have focused on small ncRNAs. As a result of these studies it is now clear that this class of ncRNAs regulates almost every aspect of gene expression (Goodrich and Kugel, 2006).

In addition to small ncRNAs, large numbers of long ncRNAs have recently been revealed by large-scale transcriptome analyses. Although only a limited number of long ncRNAs have well-characterized structures and functions, many studies suggest that this class of ncRNA accounts for a large fraction of the transcriptome, and they are believed to play important roles in many key molecular regulatory processes (Yazgan and Krebs, 2007; Umlauf et al., 2008; Mercer et al., 2009). We will recapitulate the pervasive transcription and genome/transcriptome complexity of these regulatory ncRNAs, particularly with respect to long ncRNAs, and review their primary proposed functional models.

PERVASIVE TRANSCRIPTION OF ncRNAs

Rapid development in analytical technologies, such as whole genome tilling arrays, capped analysis of gene expression (CAGE), Chip-chip, Chip-seq, and RNA deep sequencing have revised people's views of eukaryotic genome/transcriptome complexity (Carninci, 2006; Gustincich et al., 2006). In the past decade, large-scale transcriptome analyses of several organisms indicate that genomes are pervasively transcribed and ncRNAs account for a large proportion of the whole transcriptome (Bertone et al., 2004; Birney et al., 2007).

THE HUMAN TRANSCRIPTOME IS MORE COMPLICATED THAN EXPECTED

It has been more than decade since the human genome was sequenced, yet the decoding of this information is far from complete. According to the statistics of the version 34b of the Ensembl Human Genome, there are about 20–25,000 protein-coding genes, with a total coding length of ∼34 Mb, which only occupies ∼1.2% of the whole genome. On the other hand, about 1,679 Mb non-coding sequences, accounting for more than half (∼57%) of the whole human genome, are believed to be transcribed. These non-coding sequences include introns, untranslated regions (UTRs), and other intronic and exonic sequences covered by spliced cDNAs/ESTs that are not annotated as protein coding. The

47:1 ratio of transcribed non-coding regions to coding regions indicates that ncRNAs represent a large share of the human transcriptome (Frith et al., 2005). Tiling array and other several large-scale analyses of the human genome have also provided strong support to this hypothesis. The large-scale transcriptional analysis of human chromosome 21 and 22 using oligonucleotide arrays showed that only 2.6% (26,516 of 1,011,768) probe pairs that interrogate approximately 35 Mb non-repetitive regions of these two chromosomes are detected inside the annotated exons of well-characterized genes. Ninety-four percent of the probes are expressed and located outside annotated exons in 1 of 11 detected cell lines, and the percentage is 88 for 5 of 11 cell lines. This indicates that some of non-coding transcripts are cell type-specific expressed (Kapranov et al., 2002). Further in-depth transcriptome analysis of human chromosome 21 and 22 provided similar results: nearly half of the studied transcripts originated outside of well-annotated exons, and these novel transcripts seem to have less variation and be cell type-specific in expression compared to well-characterized genes (Kampa et al., 2004). These results are reinforced by subsequent high-density genome tiling array studies of 10 human chromosomes (Cheng et al., 2005) and massively parallel signature sequencing (MPSS) analysis (Jongeneel et al., 2005). All of these results clearly demonstrate that the human genome is highly transcribed and the landscape of human ncRNAs is extremely complex.

ncRNAs ARE A MAJOR COMPONENT OF THE MOUSE TRANSCRIPTOME

Large-scale transcription analyses of the mouse genome have also revealed that ncRNAs are commonly transcribed. Early in 2002, a dataset of ncRNAs from the mouse transcriptome was proposed based on functional annotation of full-length cDNAs (also called FANTOM2). Over one-third (34.9%) of 33,409 "transcriptional units," clustered from 60,770 full-length cDNAs, were predicted as novel non-coding transcripts (Okazaki et al., 2002). According to the analysis of FANTOM3 in 2006, the number of predicted distinct non-coding transcripts had increased to 34,030, over threefold compared to FANTOM2 (Maeda et al., 2006). Further analysis of FANTOM3 by the FANTOM Consortium revealed that many putative ncRNAs were singletons in the full-length cDNA set but that 3,652 cDNAs, which were supported by overlapping with both the initiation and termination sites of ESTs, CAGE tags, or other cDNA clones, were identified as ncRNAs. In addition, 3,012 cDNAs that were previously regarded as truncated CDS were identified as genuine transcripts and were believed to be the ncRNA variants of protein-coding cDNAs (Carninci et al., 2005). Transcriptome sequencing of mouse embryonic stem cells also revealed 1,022 non-coding expressed transcripts, and some of them were shown to have expression levels correlated with differentiation state (Araki et al., 2006). The existence of large numbers of ncRNAs transcribed from the mouse genome was subsequently validated by RT-PCR, microarray, and northern blot analyses (Ravasi et al., 2006).

OTHER SPECIES ALSO EXPRESS LARGE NUMBERS OF ncRNAs

Although there have been fewer large-scale transcriptome studies of species other than human and mouse, they have confirmed the existence of ncRNAs. Seventeen distinct non-protein-coding polyadenylated transcripts were identified from the intergenic regions of the fly genome (Tupy et al., 2005). Moreover, 136 strong candidates for mRNA-like ncRNAs were screened from 11,691 fly full-length cDNAs, and 35 of them were expressed during embryogenesis. Of these 35 mRNA-like ncRNAs, 27 were detected only in specific tissues (Inagaki et al., 2005). These results indicate that many mRNA-like ncRNAs are expected to play important roles in the fly. In 2005, approximately 1,300 genes that produce functional ncRNAs were demonstrated in the worm *C. elegans* (Stricklin et al., 2005). However, the worm transcriptome is much more complicated than expected. The worm non-coding transcriptome mapped by whole-genome tiling array showed that at least 70% of the total worm genome was transcribed, and 44% of the total observed transcriptional output on the array was predicted to consist of non-polyadenylated transcripts without protein-coding potential. Seventy percent of these non-polyadenylated transcripts were shown to overlap with the coordinates of coding loci in complicated fashions (He et al., 2007). The prevalence of ncRNAs extends even further, as studies of *Saccharomyces cerevisiae* have also revealed large numbers of ncRNAs (Havilio et al., 2005; Miura et al., 2006).

EVIDENCE FROM WELL-CHARACTERIZED LONG ncRNA DATASETS

In past several years, our knowledge of long ncRNAs has been expanding thanks to the identification and annotation of diverse classes of long ncRNAs from human, mouse, and other species (**Table 1**). About 1,600 large intervening/intergenic ncRNAs (lincRNAs) were identified based on the chromatin-state maps from four mouse cell types (Guttman et al., 2009). Based on the same method, ∼3,300 lincRNAs were characterized according to the chromatin-state maps of various human cell types (Khalil et al., 2009). Moreover, a class of ∼3,200 enhancer-like long ncRNAs were discovered as a result of the ENCODE project (Orom et al., 2010). The rapid drop in price of next generation sequencing drove the generation of large amounts of RNA-seq data from a number of species. More than a thousand multi-exonic lincRNAs were revealed by reconstruction of transcriptomes from three mouse cell types (Guttman et al., 2010). Human transcriptome data from more sources (24 tissues and cell types), allowed the reconstruction of more than 8,000 human lincRNAs (Cabili et al., 2011). Large numbers of long ncRNAs were also found in zebrafish, fly, and worm transcriptomes based on RNA-seq data. A stringent set of 1,133 non-coding multi-exonic transcripts, including lincRNAs, intronic overlapping long ncRNAs, exonic antisense overlapping long ncRNAs, and precursors for small RNAs (sRNAs), were identified from transcriptome data of eight early zebrafish development stages (Pauli et al., 2011). Recently, 1,199 putative lincRNAs and more than 800 lincRNAs were annotated from fly and worm transcriptomes based on RNA-seq data (Nam and Bartel, 2012; Young et al., 2012).

EVOLUTIONARY CONSERVATION OF LONG ncRNAs

In contrast to well-conserved small ncRNAs, like miRNAs, the evolutionary sequence conservation of long ncRNAs is less pronounced. Most studies have shown that long ncRNAs are poorly conserved compared to protein-coding genes (Louro et al., 2009;

Table 1 | Recently well-characterized long ncRNA datasets.

Dataset	Number of ncRNAs	Source	Method	Reference
Chromatin-state-based lincRNAs (human)	4,860*	10 cell types	Chromatin signature identification (K4–K36 domain)	Khalil et al. (2009)
Enhancer-like long ncRNAs (human)	3,011	Multiple	Screening from GENCODE annotation	Orom et al. (2010)
RNA-seq-based lincRNAs (human)	8,195	24 tissues and cell types	Screening from assembled RNA-seq data	Cabili et al. (2011)
Chromatin-state-based lincRNAs (mouse)	2,127*	Four cell types	Chromatin signature identification (K4–K36 domain)	Guttman et al. (2009)
RNA-seq-based lincRNAs (mouse)	1,140	Three cell types	Screening from assembled RNA-seq data	Guttman et al. (2010)
RNA-seq-based long ncRNAs (zebrafish)	1,133	Eight embryonic stages	Screening from assembled RNA-seq data	Pauli et al. (2011)
RNA-seq-based lincRNAs (fruit fly)	1,119	30 developmental time points	Screening from assembled RNA-seq data	Young et al. (2012)
RNA-seq-based lincRNAs (*C. elegans*)	882	Multiple	Screening from assembled RNA-seq data	Nam and Bartel (2012)

**These are the exons identified by microarray from non-coding k4–k36 domains.*

Mercer et al., 2009). In a comparison between human and mouse long ncRNAs, Pang et al. (2006) found that the sequence homology of long ncRNAs was similar to that of introns (<70% between mice and humans) and a little less conserved than 5′ or 3′ UTRs. Thus the evolutionary constraints acting on long ncRNAs may differ from the constraints affecting small ncRNAs, allowing long ncRNAs to evolve faster than small RNAs. However, conservation analysis of long ncRNAs based on 50-nt window size revealed that many long ncRNAs may retain patches of higher conservation within their overall sequences, possibly representing interaction sites with RNA-binding proteins (Pang et al., 2006).

Recently, novel long ncRNA datasets identified from diverse species have confirmed that most long ncRNAs are less conserved than protein-coding genes while still showing clear conservation. Over 95% of the 1,600 mouse lincRNAs identified by chromatin-state maps showed clear evolutionary conservation (Guttman et al., 2009). Subsequent analysis of 3,300 human chromatin-state based lincRNAs also indicated that these lincRNAs were more conserved than intronic regions (Khalil et al., 2009). Analysis of human enhancer-like long ncRNAs also showed that the global conservation levels of these long ncRNAs were less than protein-coding genes, but higher than ancestral repeats (Orom et al., 2010). Long ncRNAs reconstructed from mouse RNA-seq data showed similar conservation levels compared to chromatin-state based lincRNAs (Guttman et al., 2010). In human, RNA-seq based long ncRNAs showed moderate conservation across different species (Cabili et al., 2011). The conservation of zebrafish RNA-seq derived long ncRNAs assessed by CBL score was substantially lower than protein-coding genes and comparable to intronic sequences (Pauli et al., 2011). Analysis from the fly RNA-seq based lincRNAs also showed that most of these ncRNAs, even for those expressed at low levels, have significantly lower nucleotide substitution rates compared with either untranscribed intergenic sequence or neutrally evolving short introns (Young et al., 2012). RNA-seq based lincRNAs identified from another invertebrate organism *C. elegans* were differentiated into two subclasses according to their conservation, non-conserved and moderately conserved. Similar to vertebrates, some of these *C. elegans* lincRNAs also tend to have short regions of conservation (Nam and Bartel, 2012).

Overall, while long ncRNAs identified from different species and based on different methods showed slightly different levels of conservation, it is clear that long ncRNAs are less conserved than protein-coding genes but still exhibit clear conservation compared to non-functional genomic elements. One widely accepted interpretation of poor sequence conservation for long ncRNAs is that long ncRNAs may function at the secondary structure level instead of the primary sequence level. This is in contrast to protein-coding, which genes require conserved nucleotide sequence to encode higher levels of structure with similar biological functions. Differently, the small conserved patches observed in some long ncRNAs might be sufficient to support the functions of these long ncRNAs, by binding with proteins, interacting with DNA promoters or with UTRs of mRNAs. Finally, the long ncRNA datasets described above were identified using different methods, possibly fostering bias for some classes of long ncRNAs, which might be subject to different selective pressure.

GENOMIC ORGANIZATION OF ncRNAs

Regulatory ncRNAs originate from different genomic regions (**Figure 1**). UTRs account for many of the regions encoding ncRNAs. Statistics from the UCSC human genome (NCBI build 35) show that total UTR sequences account for ∼1.1% of the whole human genome, nearly equivalent in length to protein-coding regions (32–34 Mb; Frith et al., 2005). This suggests that there may be unknown regulatory elements in these regions. Studies using CAGE, serial analysis of gene expression (SAGE), cDNA libraries, and microarray expression profiles have shown that there are independent transcripts expressed from 3′ UTRs. This class of independent transcripts has been termed "uaRNAs" (UTR-associated RNAs), some of which have been validated as being expressed in cell- and subcellular-specific fashion (Mercer et al., 2010).

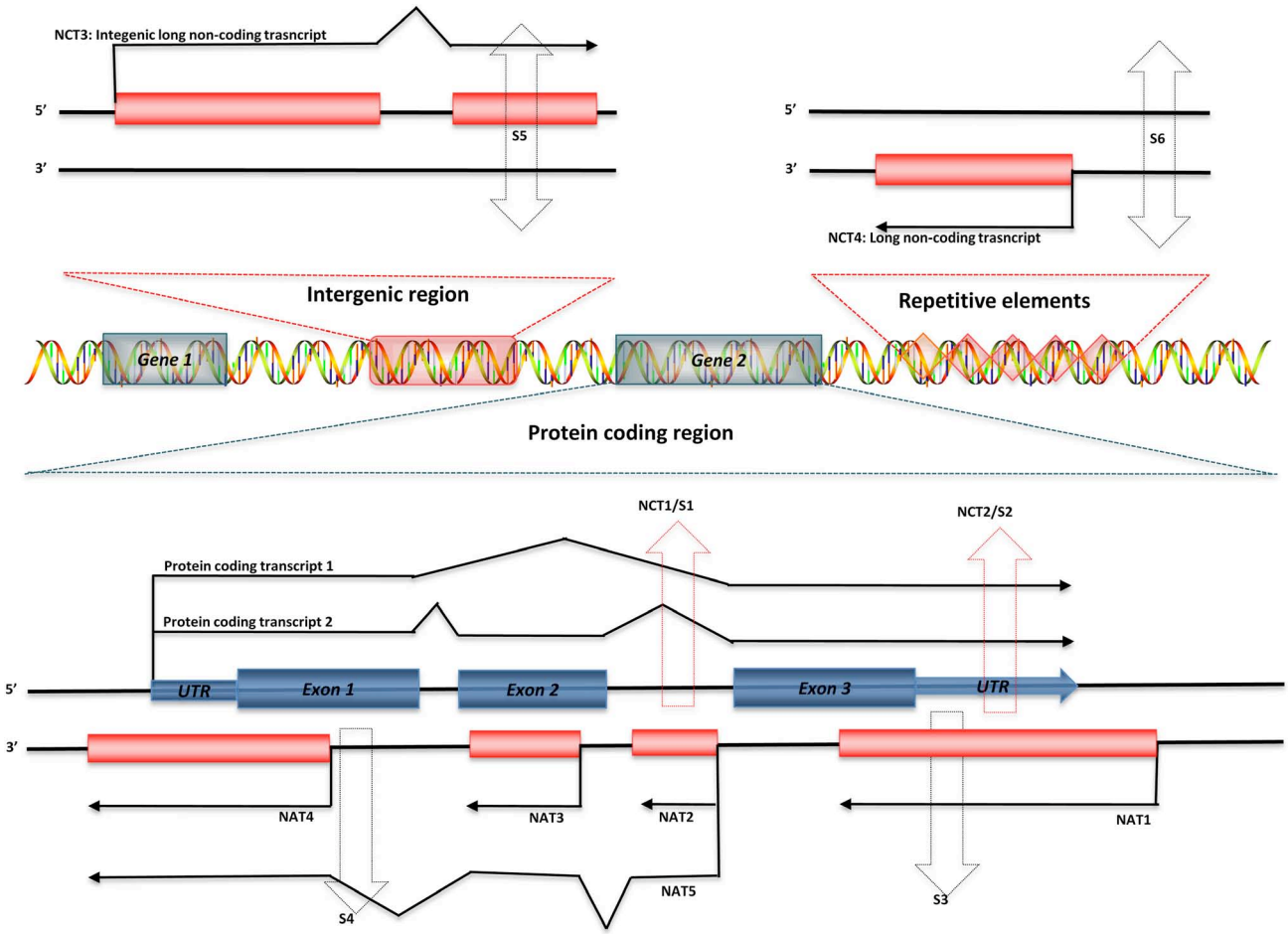

FIGURE 1 | Genomic organization of regulatory ncRNAs in mammalian genome. Green symbols represent protein-coding transcripts; red symbols represent non-coding transcripts; black dashed arrows represent small ncRNAs; and red dashed arrows represent shared regions transcribed as long ncRNAs or small ncRNAs. Long ncRNAs can be transcribed from: (1) Non-coding regions of protein-coding transcript: intron – NCT1; UTRs – NCT2. (2) Antisense of protein-coding transcripts: convergent (tail–tail) antisense transcript – NAT1; intronic antisense transcript – NAT2; contained antisense transcript – NAT3; divergent (head–head) antisense transcript – NAT4; mixed-model antisense transcript – NAT5. (3) Intergenic region: NCT3. (4) Repetitive elements: NCT4. Small ncRNAs can be transcribed from introns (S1) or UTRs (S2) of protein-coding genes, antisense region of UTRs (S3) or exons (S4), both strands of intergenic regions (S5), and both strands of repetitive elements (S6).

In addition to UTRs, other non-coding regions of genome, such as intronic sequences are also a potential source of functional ncRNAs. Over 30% of the human genome is made up of intronic sequences (Mattick and Gagen, 2001), and many highly conserved sequences have been identified in introns (Taft et al., 2007). Recent research has indicated that there are a large number of long intronic ncRNAs in both human and mouse (Nakaya et al., 2007; Louro et al., 2008, 2009). Long ncRNAs can also be derived from both the sense and antisense strands of various genomic regions, some of which overlap with or are within protein-coding genes. These results indicate that distinguishing between protein-coding and non-coding RNAs may be difficult in some circumstances (Dinger et al., 2008). Most importantly, Tens of thousands of long ncRNAs have been identified from intergenic regions (lincRNA), as discussed above. More and more lincRNAs have been validated and shown to possess important regulatory functions.

BROAD FUNCTIONALITY OF LONG ncRNAs

Recent reports have revealed the widespread functionality of long ncRNAs, ranging from epigenetic modification, to transcriptional and post-transcriptional regulation of protein-coding genes. These functions may only account for part of the functional repertoire of long ncRNAs, but they provide quite clear evidence supporting the functional significance of long ncRNAs.

CHROMATIN MODIFICATION

Many studies have shown that long ncRNAs play important roles in chromatin modification (Mattick, 2003; Costa, 2008). Dosage compensation achieved by X-chromosome inactivation (XCI) is a classic example of chromatin modification mediated by long ncRNAs in mammals (Leeb et al., 2009). There are two ncRNAs involved in this process. *Xist*, a 17-kb long ncRNA, initiates XCI, while *Tsix*, an antisense non-coding transcript to the *Xist* gene, opposes XCI. However, the exact mechanism of XCI mediated by

these two ncRNAs is still unclear. Ogawa et al. (2008) reported that murine *Xist* and *Tsix* may form *Tsix:Xist* duplexes and be processed into small RNAs by Dicer, then subsequently these small RNAs trigger the RNAi machinery to drive XCI. Another mechanism has been proposed to explain how *Xist* and *Tsix* regulate XCI. In this model, a 1.6-kb ncRNA (*RepA*) transcribed from *Xist* loci identifies and recruits polycomb repressive complex 2 (PRC2), whose catalytic subunit, Exh2, functions as the RNA binding subunit, initiating XCI. *Tsix* keeps the X chromosome active by inhibiting the interaction of *RepA* and PRC2 (Zhao et al., 2008). *HOTAIR* is another well-characterized long ncRNA that can alter chromatin structure by recruiting polycomb proteins. There are 39 human *HOX* genes which can be divided into four clusters (*HOXA-D*) based on their locations on different chromosomes (Woo and Kingston, 2007). A total of 231 *HOX* ncRNAs were identified from these human *HOX* loci. These *HOX* ncRNAs have specific sequence motifs, are spatially expressed along developmental axes, and their expression demarcates broad chromosomal domains of differential histone methylation and RNA polymerase accessibility. A 2.2-kb ncRNA in the *HOX* ncRNA cluster, called *HOTAIR*, can induce heterochromatin formation and repress transcription in *trans* by recruiting PRC2 to trimethylate the lysine-27 residues of Histone H3 in *HOXD* locus (**Figure 2**; Rinn et al., 2007). A common model of epigenetic control relies on ncRNAs acting as chromatin modifying complexes. Another example of this type of mechanism involves the imprinted ncRNA *Air*, which is required for allele-specific silencing of *cis*-linked *Slc22a3*, *Slc22a2*, and *igf2r* genes in mouse placenta. *Air* is believed to target repressive histone-modifying changes by interacting with the *Slc22a3* promoter chromatin and H3K9 histone methyltransferase G9a to epigenetically repress transcription (Nagano et al., 2008). A final example of this type of transcriptional control is driven by *Kcnq1ot1* an antisense ncRNA, that mediates lineage-specific transcriptional silencing patterns by recruiting chromatin-remodeling complexes (G9a and PRC2) to specific regions in the *Kcnq1* locus (Pandey et al., 2008).

TRANSCRIPTIONAL REGULATION

Many long ncRNAs can directly regulate gene expression at the transcriptional level. Specific mechanisms for direct regulation include transcriptional interference by binding to enhancers, promoters, and transcription factors, the latter being able to alter gene expression at a global level.

Transcriptional interference from long ncRNA has been shown for *SRG1* (*SER3* regulatory gene 1), a well-studied ncRNA in *S. cerevisiae*. The *SER3* gene encodes a serine biosynthesis related enzyme. This gene is strongly repressed and its regulatory region highly transcribed when *S. cerevisiae* are grown in a rich medium. The highly expressed transcript from the SER3 regulatory region was identified by northern blot analysis as *SRG1*, a 550-nt long polyadenylated ncRNA. Substitution analysis of a 150-bp sequence of *SRG1* revealed that *SRG1* can interfere with the activation of the *SER3* promoter to repress *SER3* gene expression (**Figure 3A**; Martens et al., 2004). In metazoa, the *bithoraxoid* (*bxd*) ncRNAs of the fly bithorax complex (BX-C) are a cluster of npcRNAs that have been shown to regulate gene expression by transcriptional interference. In this case, the transcription of several *bxd* ncRNAs are linked to the repression of the *Ubx* (*Ultrabithorax*)

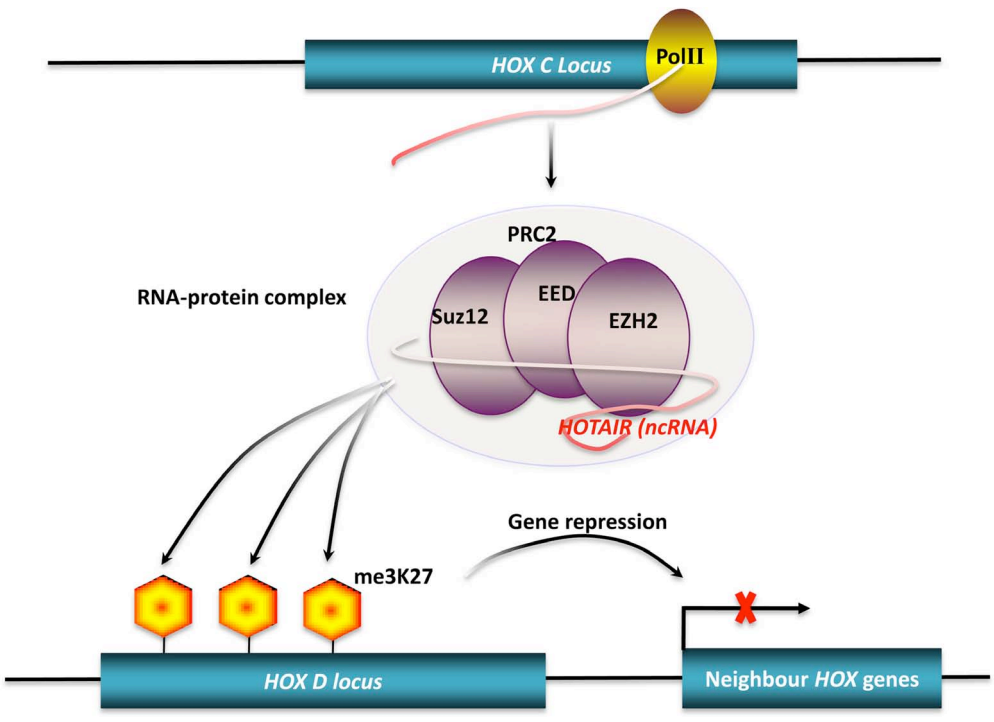

FIGURE 2 | Chromatin modification by long ncRNA.

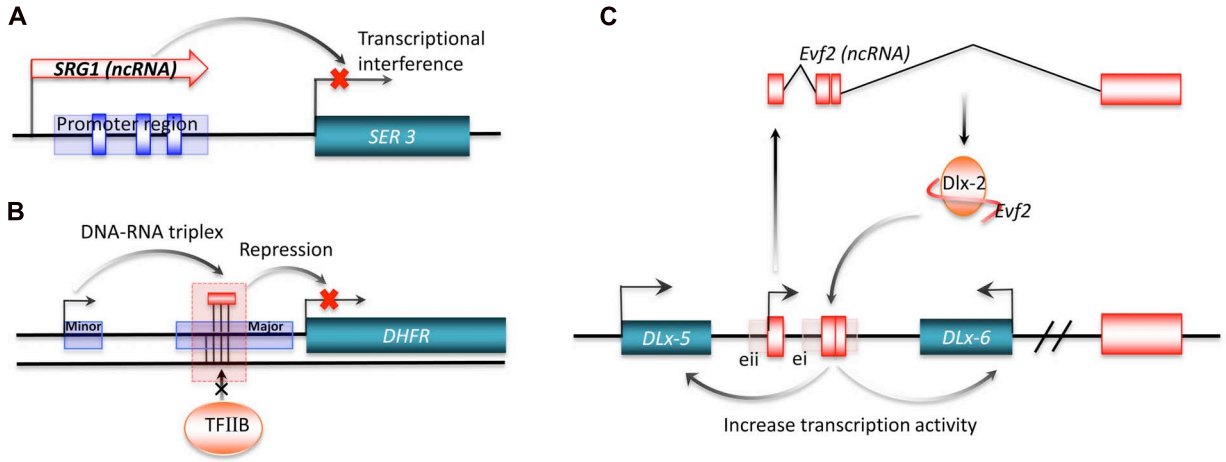

FIGURE 3 | Transcriptional regulation by long ncRNA. Long ncRNAs can conduct transcriptional regulation of target genes by: **(A)** transcriptional interference. The purple rectangle represents the promoter region for SER3, and the blue boxes are SER3 upstream activating sequence (UAS) elements; **(B)** interacting with promoters. The short purple rectangle represents the minor promoter, and the long purple rectangle represents the major promoter; **(C)** interacting with transcriptional factors. ei and eii are two ultraconserved enhancers.

protein-coding gene. Transcription of *bxd* ncRNAs represses *Ubx* expression in *cis*, where *Ubx* transcription is repressed by the transcriptional elongation of *bxd* ncRNAs. This is facilitated by the Trithorax complex TAC1, a transcriptional effector that binds to the *bxd* region (Petruk et al., 2006).

Interaction of promoters with long ncRNAs can also regulate gene expression. One example is a non-coding transcript initiated from the upstream minor promoter of the human dihydrofolate reductase (*DHFR*) gene, which represses gene expression by promoter inactivation. The *DHFR* locus has two promoters, with the downstream major promoter responsible for 99% of RNA transcription (Masters and Attardi, 1985). The upstream promoter generates a non-coding transcript that forms a stable complex with the major promoter by interacting with transcription factor II B (TFIIB). This complex acts by dissociating the pre-initiation complex from the major promoter (**Figure 3B**; Martianov et al., 2007). Another signal-induced low-copy-number ncRNA, over 200 nt long, named *ncRNA$_{CCND1}$s*, also mediates the repression of gene expression by promoter interaction (Wang et al., 2008). *ncRNA$_{CCND1}$s* recruits a key transcriptional sensor of DNA damage, the translocated in liposarcoma (TLS) RNA-binding protein, to the promoter region of *cyclin D1* (*CCND1*). The recruited TLS RNA-binding protein inhibits the histone acetyltransferase activities of CREB-binding protein (CBP) and p300. This leads to the repression of *CCND1* gene expression in human cell lines. Of particular interest is the signal-induced transcription of *ncRNA$_{CCND1}$s*, which may provide a novel understanding of stimulus-specific expression of ncRNAs (Wang et al., 2008).

In addition to promoter inactivation or activation, an increasing number of studies now suggest that ncRNAs also regulate gene expression by interacting with transcription factors. One example is *Evf-2*, which is a ~3.8-kb ncRNA transcribed from ei, one of the two *Dlx-5/6* ultraconserved intergenic regions (Zerucha et al., 2000). The ultraconserved region of *Evf-2* specifically interacts with the Dlx-2 protein to form a complex, which increases the transcriptional activity of the *Dlx-5/6* enhancer in a target and homeodomain-specific manner. The stable complex of *Evf-2* ncRNA and the Dlx-2 protein has been validated by *in vivo* assay, indicating that *Evf-2* ncRNA regulates transcriptional activity by directly affecting Dlx-2 activity (**Figure 3C**; Feng et al., 2006). The abundance of such ultraconserved sequences in vertebrate genomes suggests that this mechanism is a common strategy for the regulation of key developmental genes (Bejerano et al., 2004; Sandelin et al., 2004). Another example of this mechanism is *SRA* a ncRNA that interacts with MyoD, a transcription factor that regulates skeletal myogenesis. Through *in vitro* and *in vivo* experiments, Caretti et al. (2006) found that RNA helicases p68/p72, two MyoD-associated proteins, and *SRA* are co-activators of MyoD. The normal activation of muscle gene expression and cell differentiation are suppressed by RNA interference of *SRA*, implying that *SRA* plays an important role in the regulation of developmental gene expression.

Recent experimental evidence has indicated that long ncRNAs could contribute to the complexity of gene expression regulatory networks, where some long ncRNAs might alter global gene expression through a *trans*-acting mechanism. Using gene chip array analysis, Hill et al. (2006) proposed that human introns can coordinate the expression of a wide range of gene products at spatially diverse sites in the genome without miRNAs. Their experiments showed that extensive and specific transcriptional activities in epithelial cells (Hela) were influenced by the expression of three intronic sequences derived from the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, which was also abnormally expressed. A wide range of genes related to processes of epithelial differentiation and repair were affected as a result of these transcriptional changes, such as FOXF1, sucrase-isomaltase, collagen, interferon, complement, and thrombospondin 1. Hill et al. (2006) suggested that ncRNAs transcribed from intronic regions were responsible for these changes. In a similar vein, the

human *Alu* RNA, which is transcribed from short interspersed elements (SINEs), is recognized as a transacting transcriptional repressor which inhibits transcription by binding to RNA polymerase II (Pol II) complexes at promoters *in vitro* as a result of heat shock (Mariner et al., 2008). Because *Alu* elements are so abundant in the human genome, they may contribute to long ncRNA transcriptional repressor function (Amaral and Mattick, 2008).

POST-TRANSCRIPTIONAL REGULATION

There are many reports providing evidence that ncRNAs have the ability to regulate various aspects of post-transcriptional mRNA processing of, such as splicing, editing, transport, translation, or degradation. The significant role in post-transcriptional regulation of gene expression mediated by small regulatory ncRNAs has been well characterized in various species (see reviews Grishok, 2005; Kavi et al., 2005; Wienholds and Plasterk, 2005; Scherr and Eder, 2007; Filipowicz et al., 2008). Here we discuss how long ncRNAs can mediate post-transcriptional regulation via specific mechanisms.

Some antisense ncRNAs have been shown to regulate gene expression at the post-transcriptional level. For example, *SAF* is a long ncRNA transcribed from the antisense strand of intron 1 of the human *Fas* gene. The overexpression of *SAF* caused the proteins encoded by *Fas* to fail to anchor to the cell membrane and induce *Fas*-mediated apoptosis. It is believed that *SAF* regulates the expression of *Fas* alternative splicing forms through pre-mRNA processing (Yan et al., 2005). Another natural antisense transcript (NAT) of the *Snail1* gene can up-regulate gene expression by forming RNA duplexes in the following fashion. The expression of Zeb2, a transcriptional repressor of E-cadherin, requires an internal ribosome entry site (IRES) derived from a large intron located in the 5′ UTR of the *Snail1* gene, whose expression in epithelial cells triggers an epithelial–mesenchymal transition (EMT). The *Snail1* NAT overlaps with the 5′ splice site of the large intron and Beltran et al. (2008) found that overexpression of this NAT prevented the splicing of the *Zeb2* 5′-UTR, causing an increase in the expression level of the Zeb2 protein. Many antisense transcripts have been mapped to the introns of mammalian genomes (He et al., 2008; Li et al., 2008) indicating that this type of antisense regulation of alternative splicing may be quite common.

Another aspect of post-transcriptional regulation of gene expression mediated by long ncRNAs is the stabilization of protein-coding RNAs. Adenylate- and uridylate-rich (AU-rich) elements are specific *cis*-acting elements, found in the 3′ UTRs of many unstable mammalian mRNAs, controlling their half-lives (Bevilacqua et al., 2003). This *cis*-acting regulation can be inhibited, as shown by a *bcl-2*/IgH antisense transcript, formed by with *bcl-2*/IgH translocation, that up-regulates *bcl-2* mRNA expression. This hybrid antisense transcript masks AU-rich motifs present in the 3′ UTR of the *bcl-2* mRNA, increasing the stability of the protein-coding mRNA (**Figure 4**; Capaccioli et al., 1996). Although there is still little direct experimental evidence to identify all mechanisms involved, comparison of genome-scale expression profiles between protein-coding and non-protein-coding RNAs suggests that widespread post-transcriptional control of gene expression via the stabilization of protein-coding RNAs does occur (Nakaya et al., 2007).

CANCER AND LONG ncRNAs

Many ncRNAs play regulatory roles in cancer biology. Because they regulate cell differentiation and various developmental processes, the mis-expression of long ncRNAs can regulate clinically significant cancer genes.

A number of long ncRNAs have been associated with cancer development and progression. The antisense ncRNA *p15AS* epigenetically silences its sense target gene *p15* in leukemia (Yu et al., 2008). The expression of *p15AS* induces *p15* silencing in *cis* and *trans* through heterochromatin formation. *p15* silencing and increased cell growth were observed after differentiation of mouse embryonic stem cells induced by exogenous *p15AS* (Yu et al., 2008). *ANRIL* (antisense ncRNA from the *INK4A-ARF-INK4B* locus), which is regarded as an isoform of *p15AS*, interacts with chromobox homolog 7 (CBX7), a subunit of the PRC1 protein, and mediates the epigenetic transcriptional repression of its sense locus (Yap et al., 2010). Subsequent study revealed that this ncRNA binds to SUZ12 (suppressor of zeste 12 homolog), a component

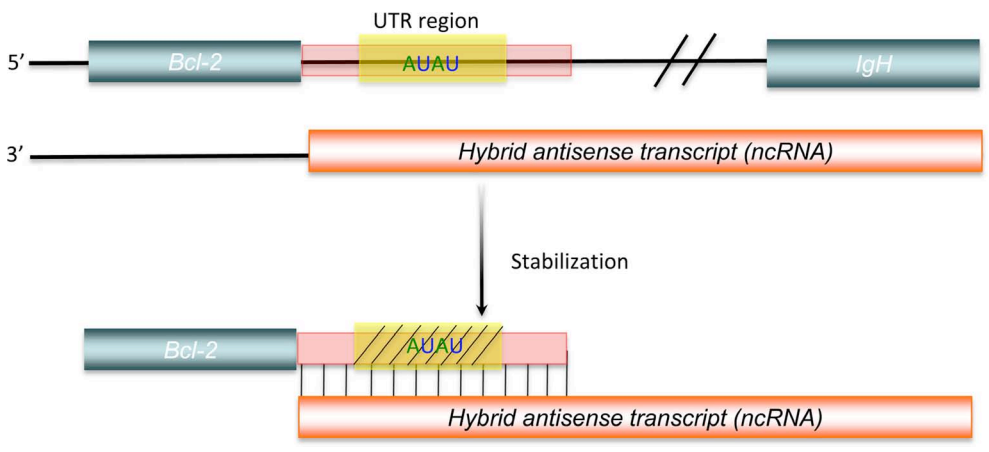

FIGURE 4 | Post-transcriptional regulation by long ncRNA.

of the PRC2, and recruits PRC2 to epigenetically repress *p15*^{INK4B} (Kotake et al., 2011).

In addition to acting as repressors of tumor suppressor genes, long ncRNAs also contribute to tumorigenesis via other mechanisms. *SRA* is a well-characterized ncRNA, which can co-activate the activity of a number of nuclear receptors in tumors. It can promote muscle differentiation and myogenic conversion of non-muscle cells through the co-activation of MyoD activity as discussed above (Caretti et al., 2006; Hube et al., 2011). Another long ncRNA *PCAT-1* (prostate cancer-associated transcript 1), which is over-expressed in a subset of prostate cancers, particularly metastatic tumors, is known to regulate cell proliferation in prostate cancer progression (Prensner et al., 2011). Moreover, long ncRNA *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1) was shown to be significantly associated with metastasis in non-small cell lung cancer patients (Ji et al., 2003). Subsequent analysis indicated that *MALAT1* was overexpressed in five other non-hepatic human carcinomas (Lin et al., 2007). *MALAT1* may play important roles in tumor cell invasion and formation of metastases (Tseng et al., 2009; Tano et al., 2010). In prostate cancer, a cDNA microarray analysis of intronic transcripts indicated that a high percentage (6.6%) of intronic transcripts were correlated with the degree of prostate tumor differentiation compared to transcripts from unannotated genomic regions (1%; Reis et al., 2004). In renal carcinoma cells (RCC) expression profiles also revealed that there are some non-coding intronic RNAs that are associated with malignant transformation of normal renal cells to tumor cells (Brito et al., 2008). As a result of these and similar observations, long ncRNAs have been used as diagnostic biomarkers because of their cell type-specific or stage-specific expression in different cancers (Mallardo et al., 2008; Reis and Verjovski-Almeida, 2012).

In addition to their functions contributing to tumorigenesis, many ncRNAs are known to act as tumor suppressors. One example is the imprinted gene *MEG3* (maternally expressed gene 3), which functions as a long ncRNA. Although *MEG3* has an open reading frame, it is the folding of *MEG3* RNA that activates p53 expression and selectively regulates p53 target gene expression (Zhou et al., 2007). In addition, *MEG3* can also inhibit cell proliferation via a p53-independent pathway. This evidence suggests that *MEG3* functions as a tumor suppressor in p53 dependent and independent fashion (Zhou et al., 2007; Zhang et al., 2010). Another long ncRNA, *Gas5* (growth arrest-specific 5), binds to the DNA-binding domain of the glucocorticoid receptor (GR), preventing the interaction of glucocorticoid response elements (GRE) with GR. The repression of GR suppresses the glucocorticoid-mediated induction of several genes, leading to apoptosis (Kino et al., 2010). Among the more than 1000 mouse chromatin-state based lincRNAs, one of them (*lincRNA-p21*) functions as a repressor of p53-dependent transcriptional response. *LincRNA-p21* is a transcriptional target gene of p53. It recruits a repressor complex, including heterogeneous nuclear ribonucleoprotein K (hnRNP-K), to a subset of previously active genes, mediating global gene repression and leading to apoptosis (Guttman et al., 2009; Huarte et al., 2010).

These results clearly illustrate the functional significance of long ncRNAs in tumorigenesis and cancer regulatory networks and transcriptional pathways. However, some mechanisms of long ncRNAs in cancer biology seem to be more complicated than expected. For instance, *lincRNA-p21* is transcribed from a region ∼15 kb upstream of p21 and mediates apoptosis in a p53-dependent manner upon DNA damage response as discussed above (Huarte et al., 2010). Another single exonic long ncRNA *PANDA* (P21 associated ncRNA DNA damage activated), is transcribed from the ∼5 kb upstream region of p21 in an antisense orientation to p21. The expression of *PANDA* is also induced by DNA damage and activated in a p53-dependent manner as *lincRNA-p21*. However, in contrast to *lincRNA-p21*, *PNADA* interacts with the transcription factor NF-YA to limit the expression of some pro-apoptotic genes (Hung et al., 2011). This is just one example of the complexity of cancer-related gene regulation by long ncRNAs. As more long ncRNAs become validated, we can imagine that more regulatory roles of long ncRNAs in tumorigenesis will be unveiled.

CONCLUSION

The recent explosion in studies of ncRNAs has fostered a new view of the RNA world. It is clear that gene regulation networks are more complicated than expected. And that in future, the central dogma may be challenged by more roles for ncRNAs. Genomes possess a high percentage of non-coding regions, and express a huge repertoire of ncRNAs, which probably contribute to cellular regulatory networks.

The functional significance of ncRNAs has been debated because of their perceived lack of evolutionary conservation. Lower conservation of ncRNAs (mostly for long ncRNAs) was regarded as an argument against functional importance and as a manifestation of transcriptional noise. But less conservation does not mean less function. Many studies indicate that evolutionary constraints on ncRNAs are different to protein-coding RNAs. These different constraints allow many ncRNAs to evolve more quickly subject to positive selection. The complexity underlying the evolutionary conservation of ncRNAs may be stem from the heterogeneity of ncRNAs. ncRNAs derived from different genomic regions may play different regulatory functions. In order to carry out those functions, each class of ncRNA from similar regions may share corresponding specific structures and characteristics, which undergo different evolutionary processes leading to different conservation patterns.

The ncRNA contribution to regulatory networks is complex. Many functional ncRNAs influence chromatin modification, and regulate gene expression at both transcriptional and post-transcriptional levels (Amaral et al., 2010). Although overwhelming evidence has shown that ncRNAs are pervasively expressed from different genomic regions, and possess a wide range of functionality in gene regulation, these discoveries still provide only a glimpse of the hidden ncRNA world. Well-annotated ncRNAs represent a small fraction of the available datasets and the majority of these annotations are structural. While continued advances in high-throughput sequencing will facilitate the discovery and elucidation of more regulatory ncRNAs, we will need a comparable revolution in high-throughput functional testing of ncRNAs to address the functions and mechanisms of long ncRNAs in regulatory networks.

REFERENCES

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2010). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39, D146–D151.

Amaral, P. P., and Mattick, J. S. (2008). Noncoding RNA in development. *Mamm. Genome* 19, 454–492.

Araki, R., Fukumura, R., Sasaki, N., Kasama, Y., Suzuki, N., Takahashi, H., et al. (2006). More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem Cells* 24, 2522–2528.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., et al. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325.

Beltran, M., Puig, I., Pena, C., Garcia, J. M., Alvarez, A. B., Pena, R., et al. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. *Genes Dev.* 22, 756–769.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242–2246.

Bevilacqua, A., Ceriani, M. C., Capaccioli, S., and Nicolin, A. (2003). Post-transcriptional regulation of gene expression by degradation of messenger RNAs. *J. Cell. Physiol.* 195, 356–372.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Brito, G. C., Fachel, A. A., Vettore, A. L., Vignal, G. M., Gimba, E. R., Campos, F. S., et al. (2008). Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. *Mol. Carcinog.* 47, 757–767.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927.

Capaccioli, S., Quattrone, A., Schiavone, N., Calastretti, A., Copreni, E., Bevilacqua, A., et al. (1996). A bcl-2/IgH antisense transcript deregulates bcl-2 gene expression in human follicular lymphoma t(14;18) cell lines. *Oncogene* 13, 105–115.

Caretti, G., Schiltz, R. L., Dilworth, F. J., Di Padova, M., Zhao, P., Ogryzko, V., et al. (2006). The RNA helicases p68/p72 and the noncoding RNA SRA are coregulators of MyoD and skeletal muscle differentiation. *Dev. Cell* 11, 547–560.

Carninci, P. (2006). Tagging mammalian transcription complexity. *Trends Genet.* 22, 501–510.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154.

Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene* 410, 9–17.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563.

Dinger, M. E., Pang, K. C., Mercer, T. R., and Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* 4, e1000176. doi: 10.1371/journal.pcbi.1000176

Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P., and Kohtz, J. D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* 20, 1470–1484.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9, 102–114.

Frith, M. C., Pheasant, M., and Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet.* 13, 894–897.

Gilbert, W. (1986). Origin of life: the RNA world. *Nature* 319, 618.

Goodrich, J. A., and Kugel, J. F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* 7, 612–616.

Grishok, A. (2005). RNAi mechanisms in *Caenorhabditis elegans*. *FEBS Lett.* 579, 5932–5939.

Gustincich, S., Sandelin, A., Plessy, C., Katayama, S., Simone, R., Lazarevic, D., et al. (2006). The complexity of the mammalian transcriptome. *J. Physiol.* 575, 321–332.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510.

Havilio, M., Levanon, E. Y., Lerman, G., Kupiec, M., and Eisenberg, E. (2005). Evidence for abundant transcription of non-coding regions in the *Saccharomyces cerevisiae* genome. *BMC Genomics* 6, 93. doi: 10.1186/1471-2164-6-93

He, H., Wang, J., Liu, T., Liu, X. S., Li, T., Wang, Y., et al. (2007). Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.* 17, 1471–1477.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science* 322, 1855–1857.

Hill, A. E., Hong, J. S., Wen, H., Teng, L., Mcpherson, D. T., Mcpherson, S. A., et al. (2006). MicroRNA-like effects of complete intronic sequences. *Front. Biosci.* 11, 1998–2006.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409–419.

Hube, F., Velasco, G., Rollin, J., Furling, D., and Francastel, C. (2011). Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. *Nucleic Acids Res.* 39, 513–525.

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., et al. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat. Genet.* 43, 621–629.

Inagaki, S., Numata, K., Kondo, T., Tomita, M., Yasuda, K., Kanai, A., et al. (2005). Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*. *Genes Cells* 10, 1163–1173.

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041.

Jongeneel, C. V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C. D., Khrebtukova, I., et al. (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* 15, 1007–1014.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., et al. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919.

Kavi, H. H., Fernandez, H. R., Xie, W., and Birchler, J. A. (2005). RNA silencing in *Drosophila*. *FEBS Lett.* 579, 5940–5949.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672.

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., and Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.* 3, ra8.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., et al. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* 30, 1956–1962.

Leeb, M., Steffen, P. A., and Wutz, A. (2009). X chromosome inactivation sparked by non-coding RNAs. *RNA Biol.* 6, 94–99.

Li, J. T., Zhang, Y., Kong, L., Liu, Q. R., and Wei, L. (2008). Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. *Nucleic Acids Res.* 36, 4833–4844.

Lin, R., Maeda, S., Liu, C., Karin, M., and Edgington, T. S. (2007). A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 26, 851–858.

Louro, R., El-Jundi, T., Nakaya, H. I., Reis, E. M., and Verjovski-Almeida, S. (2008). Conserved tissue

expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics* 92, 18–25.

Louro, R., Smirnova, A. S., and Verjovski-Almeida, S. (2009). Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* 93, 291–298.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engstrom, P. G., et al. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2, e62. doi: 10.1371/journal.pgen.0020062

Mallardo, M., Poltronieri, P., and D'Urso, O. F. (2008). Non-protein coding RNA biomarkers and differential expression in cancers: a review. *J. Exp. Clin. Cancer Res.* 27, 19.

Mariner, P. D., Walters, R. D., Espinoza, C. A., Drullinger, L. F., Wagner, S. D., Kugel, J. F., et al. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* 29, 499–509.

Martens, J. A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature* 429, 571–574.

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 445, 666–670.

Masters, J. N., and Attardi, G. (1985). Discrete human dihydrofolate reductase gene transcripts present in polysomal RNA map with their 5′ ends several hundred nucleotides upstream of the main mRNA start site. *Mol. Cell. Biol.* 5, 493–500.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939.

Mattick, J. S., and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18, 1611–1630.

Mattick, J. S., and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* 15, R17–R29.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long noncoding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159.

Mercer, T. R., Wilhelm, D., Dinger, M. E., Solda, G., Korbie, D. J., Glazov, E. A., et al. (2010). Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res.* 39, 2393–2403.

Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., et al. (2006). A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17846–17851.

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., et al. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322, 1717–1720.

Nakaya, H. I., Amaral, P. P., Louro, R., Lopes, A., Fachel, A. A., Moreira, Y. B., et al. (2007). Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.* 8, R43.

Nam, J. W., and Bartel, D. (2012). Long non-coding RNAs in *C. elegans*. *Genome Res.* doi: 10.1101/gr.140475.112 [Epub ahead of print].

Ogawa, Y., Sun, B. K., and Lee, J. T. (2008). Intersection of the RNA interference and X-inactivation pathways. *Science* 320, 1336–1341.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573.

Orom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58.

Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., et al. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* 32, 232–246.

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22, 1–5.

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2011). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* 22, 577–591.

Petruk, S., Sedkov, Y., Riley, K. M., Hodgson, J., Schweisguth, F., Hirose, S., et al. (2006). Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. *Cell* 127, 1209–1221.

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* 29, 742–749.

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* 16, 11–19.

Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V., Almeida, G. T., et al. (2004). Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23, 6684–6692.

Reis, E. M., and Verjovski-Almeida, S. (2012). Perspectives of long noncoding RNAs in cancer diagnostics. *Front. Genet.* 3:32. doi: 10.3389/fgene.2012.00032

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323.

Rymarquis, L. A., Kastenmayer, J. P., Huttenhofer, A. G., and Green, P. J. (2008). Diamonds in the rough: mRNA-like non-coding RNAs. *Trends Plant. Sci.* 13, 329–334.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., et al. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 5, 99. doi: 10.1186/1471-2164-5-99

Scherr, M., and Eder, M. (2007). Gene silencing by small regulatory RNAs in mammalian cells. *Cell Cycle* 6, 444–449.

Stricklin, S. L., Griffiths-Jones, S., and Eddy, S. R. (2005). "*C. elegans* noncoding RNA genes," in *WormBook*, ed. The C. elegans Research Community (WormBook). doi/10.1895/wormbook.1.1.1, available at: http://www.wormbook.org

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299.

Tano, K., Mizuno, R., Okada, T., Rakwal, R., Shibato, J., Masuo, Y., et al. (2010). MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Lett.* 584, 4575–4580.

Tseng, J. J., Hsieh, Y. T., Hsu, S. L., and Chou, M. M. (2009). Metastasis associated lung adenocarcinoma transcript 1 is up-regulated in placenta previa increta/percreta and strongly associated with trophoblast-like cell invasion in vitro. *Mol. Hum. Reprod.* 15, 725–731.

Tupy, J. L., Bailey, A. M., Dailey, G., Evans-Holm, M., Siebel, C. W., Misra, S., et al. (2005). Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5495–5500.

Umlauf, D., Fraser, P., and Nagano, T. (2008). The role of long noncoding RNAs in chromatin structure and gene regulation: variations on a theme. *Biol. Chem.* 389, 323–331.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., et al. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 454, 126–130.

Wienholds, E., and Plasterk, R. H. (2005). MicroRNA function in animal development. *FEBS Lett.* 579, 5911–5922.

Woo, C. J., and Kingston, R. E. (2007). HOTAIR lifts noncoding RNAs to new levels. *Cell* 129, 1257–1259.

Yan, M. D., Hong, C. C., Lai, G. M., Cheng, A. L., Lin, Y. W., and Chuang, S. E. (2005). Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum. Mol. Genet.* 14, 1465–1474.

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., et al. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674.

Yazgan, O., and Krebs, J. E. (2007). Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes. *Biochem. Cell Biol.* 85, 484–496.

Young, R. S., Marques, A. C., Tibbit, C., Haerty, W., Bassett, A. R., Liu, J. L., et al. (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol. Evol.* 4, 427–442.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P., et al. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451, 202–206.

Zerucha, T., Stuhmer, T., Hatch, G., Park, B. K., Long, Q., Yu, G., et al. (2000). A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory

interactions between Dlx genes in the embryonic forebrain. *J. Neurosci.* 20, 709–721.

Zhang, X., Gejman, R., Mahta, A., Zhong, Y., Rice, K. A., Zhou, Y., et al. (2010). Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression. *Cancer Res.* 70, 2350–2358.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322, 750–756.

Zhou, Y., Zhong, Y., Wang, Y., Zhang, X., Batista, D. L., Gejman, R., et al. (2007). Activation of p53 by MEG3 non-coding RNA. *J. Biol. Chem.* 282, 24731–24742.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 August 2012; accepted: 24 September 2012; published online: 09 October 2012.

Citation: Qu Z and Adelson DL (2012) Evolutionary conservation and functional roles of ncRNA. Front. Gene. ***3****:205. doi: 10.3389/fgene.2012.00205*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

Chapter 2

Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes

Zhipeng Qu and David L. Adelson

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

PLoS ONE, August 2012, **7**:8, e42638. doi:10.1371/journal.pone.0042638

STATEMENT OF AUTHORSHIP

Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes

PLoS ONE, August 2012, **7**:8, e42638. doi:10.1371/journal.pone.0042638

Zhipeng Qu (Candidate)

Performed experiments, analysed results and wrote the manuscript.

I hereby certify that the statement of contribution is accurate

Signed..... *Date*.....

David L. Adelson

Supervised development of work and assisted in analysing results and writing the manuscript.

I hereby certify that the statement of contribution is accurate and I give permission for inclusion of the paper in the thesis

Signed..... *Date*.....

OPEN ACCESS Freely available online

PLoS one

Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes

Zhipeng Qu, David L. Adelson*

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, South Australia, Australia

Abstract

It is apparent that non-coding transcripts are a common feature of higher organisms and encode uncharacterized layers of genetic regulation and information. We used public bovine EST data from many developmental stages and tissues, and developed a pipeline for the genome wide identification and annotation of non-coding RNAs (ncRNAs). We have predicted 23,060 bovine ncRNAs, 99% of which are un-annotated, based on known ncRNA databases. Intergenic transcripts accounted for the majority (57%) of the predicted ncRNAs and the occurrence of ncRNAs and genes were only moderately correlated ($r = 0.55$, p-value$<$2.2e-16). Many of these intergenic non-coding RNAs mapped close to the 3′ or 5′ end of thousands of genes and many of these were transcribed from the opposite strand with respect to the closest gene, particularly regulatory-related genes. Conservation analyses showed that these ncRNAs were evolutionarily conserved, and many intergenic ncRNAs proximate to genes contained sequence-specific motifs. Correlation analysis of expression between these intergenic ncRNAs and protein-coding genes using RNA-seq data from a variety of tissues showed significant correlations with many transcripts. These results support the hypothesis that ncRNAs are common, transcribed in a regulated fashion and have regulatory functions.

Citation: Qu Z, Adelson DL (2012) Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes. PLoS ONE 7(8): e42638. doi:10.1371/journal.pone.0042638

Editor: Leonardo Mariño-Ramírez, National Institutes of Health, United States of America

Received June 9, 2012; **Accepted** July 11, 2012; **Published** August 6, 2012

Funding: This study was funded by the University of Adelaide. ZQ thanks the China Scholarship Council (CSC) for funding. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: david.adelson@adelaide.edu.au

Introduction

As a result of advances in DNA sequencing technologies, a number of mammalian genomes have been sequenced and assembled. The impetus for sequencing mammalian genomes is to use comparative genomics to identify important, evolutionarily conserved sequences, such as protein-coding genes. While protein-coding genes are considered the most important elements of the genome, they only account for a small fraction of the genome sequence or the mammalian transcriptome. This indicates that the complexity of the mammalian genome, especially the transcriptome, cannot be interpreted merely according to the central dogma of molecular biology "DNA-RNA-protein" [1,2,3,4,5]. In human, only about 1–2% of the entire genome is transcribed as protein-coding RNAs, while more than half (~57%) of the human genome is transcribed as "non-protein-coding" RNAs (ncRNAs) [3]. Furthermore, studies from the FANTOM consortium have also confirmed that the majority of the mouse genome is transcribed, commonly from both strands. Most of these transcripts cannot be annotated as protein-coding RNAs [4]. These findings are evidence of a hidden, non-protein-coding transcriptome in mammals.

At present there is debate about the true nature of the non-protein-coding transcriptome. Some believe that most ncRNAs are "transcriptional noise" associated with protein coding genes and have no function [6]. But this may not be the whole story. Apart from well-studied small non-protein-coding RNAs, like miRNAs, siRNAs, snoRNAs and piRNAs, other classes of abundant functional ncRNAs have been demonstrated in recent studies. Guttman *et al.* identified over a thousand highly conserved large intergenic non-coding RNAs (lincRNAs) in the mouse by analysing chromatin signatures [7]. Subsequent experimental analysis confirmed that one of these lincRNAs serves as a repressor in p53-dependant transcriptional responses [8]. Recently, another class of long non-coding RNAs was discovered in the human. Some of these thousand or so long ncRNAs were shown to have an un-anticipated enhancer-like role in activation of critical regulators of development and differentiation [9]. Furthermore, new types of small ncRNAs, like tiRNAs (tiny RNAs) [10], PASRs (Promoter-Associated Short RNAs) [11], TASRs (Termini-Associated Short RNAs) [11], and aTASRs (antisense Termini-Associated Short RNAs) [12], have been discovered in mammals. It is now clear that evidence confirms that there are indeed many functional sequences in the non-protein-coding transcriptome.

To characterize the non-coding transcriptome at genome scale, we built a computational pipeline to identify non-protein-coding transcripts from Expressed Sequence Tags (ESTs), which were originally designed to identify and annotate protein-coding genes. ESTs have the advantage of being readily available from public repositories, and are generally far longer than the RNA-seq tags generated by current high throughput DNA sequencers. The latter allows confident reconstruction of much longer transcripts. We used the bovine genome as a starting point for three main reasons: it has a large number of ESTs sampled from many tissues and developmental stages, the protein coding gene annotations are robust and based on thorough comparative genomic analysis and we had already exhaustively annotated the repetitive component

of the genome [13]. We were thus able to reconstruct many long transcripts and unambiguously map them to either protein-coding genes or non-repetitive, non-protein-coding regions of the genome. In this report we have identified thousands of non-coding RNAs (ncRNAs), the vast majority of which were previously un-annotated. We have also characterized the genomic distribution of these ncRNAs, compared to protein-coding genes and carried out conservation analyses to detect evidence of potential conserved function. Our analyses show that most ncRNAs were transcribed from clearly conserved genomic regions. A predominant class of intergenic ncRNAs were transcribed from the proximate flanking regions of genes, leading us to hypothesize that they play *cis*-regulatory roles in the regulation of their neighbour genes and/or *trans*-regulatory roles elsewhere in the genome. Taken together, our findings provide a general view of the composition, distribution, and conservation of a mammalian non-protein-coding transcriptome at genomic scale, sampled across a wide selection of tissues and developmental stages, and support the idea that most ncRNAs are of potential functional importance.

Materials and Methods

Databases

All data used in this research were sourced from public databases. Bovine ESTs were retrieved from dbEST of NCBI [14]. The information from source libraries is shown in Table S1. Two different bovine repeat databases were used: the first was developed by Adelson *et al.* [13]; the other was a custom-built repetitive protein database generated according to Smith *et al.*'s method [15]. The genome assembly of bosTau4 and its corresponding RefSeq dataset (as of September of 2009) was downloaded from NCBI. The Swiss-Prot protein reference database (as of September of 2009) was also obtained from NCBI.

Several known ncRNA databases were used to annotate ncRNAs. The miRNA database, miRBase release 14, which included 10,566 mature miRNAs and 10,867 pre-miRNAs, was obtained from miRBase (http://www.mirbase.org/) [16]. Rfam9.1, which contained tRNAs, rRNAs, snoRNAs, miRNAs, and other ncRNA models, was obtained from http://rfam.janelia.org/ [17]. NONCODE2.0 was obtained from http://www.noncode.org/ [18].

Programs used to develop the pipeline of ncRNA identification

All programs used in the pipeline of ncRNA identification can be freely accessed from the Internet (Table S2). All of them are stand-alone versions running under the Linux environment. Perl was used to link them into a pipeline. All Perl scripts are available upon request.

Annotation of ncRNAs

Several methods were used to annotate bovine ncRNAs. Similarity search was used to identify miRNAs from bovine ncRNAs. Blastn of ncRNAs against both mature miRNA and pre-miRNA databases was used to find transcripts of significant similarity to known mature miRNAs (identity >95%, coverage = 100%) and primary miRNAs (identity >95%, coverage >95%). Two steps were used to validate tRNAs from bovine ncRNAs. tRNAscan_SE was used to generate a list of tRNA candidates [19]. Only the candidates subsequently validated by Rfam were classified as known tRNAs [17].

The Stand-alone Rfam search was performed by a Perl script Rfam_scan.pl provided with Rfam [17]. Additionally, BLASTN against NONCODE2.0 was used to identify long known ncRNAs and piRNAs [18].

Distribution analysis of ncRNAs

All 23,060 ncRNAs and 24,373 RefSeqs were mapped to the bosTau4 assembly. The numbers of ncRNAs and RefSeqs in 1 MB non-overlapping bins were counted to determine the density distribution. The Spearman correlation coefficient between the densities of ncRNAs and RefSeqs per 1 MB bin across the whole genome was calculated using the R package (v2.12.0).

Positional bias analysis of intergenic ncRNAs

For each ncRNA, the closest gene model, either upstream or downstream, was defined as the nearest neighbour. The intergenic region of two nearby genes was defined as the gene interval.

To maximize the number of intergenic ncRNAs annotated in this step, the transcription orientations of intergenic ncRNAs were determined by the union, instead of the intersection of the two methods used to determine the transcription orientation of ESTs in the step of *cis*-NATs (Natural Antisense Transcripts) identification.

Functional over-representation of intergenic ncRNAs' neighbour genes

All neighbour genes with intergenic ncRNAs in 5 kb flanking upstream or downstream regions were identified. 3,166 unique genes with intergenic ncRNAs in 5′ flanking regions were identified, and 741 unique genes were identified with intergenic ncRNAs in 3′ flanking regions. The intersection of these two gene lists resulted in 183 unique genes. The GO (Gene Ontology) functional annotation and clustering were conducted using DAVID (Database for Annotation, Visualization and Integrated Discovery) [20,21]. Over-represented GO terms were filtered to contain at least 5 genes and FDR (False Discovery Rate)<0.05. Ten control gene lists for 5′ and 3′ neighbour gene lists were generated respectively. For each control list for 5′ end intergenic ncRNA, 741 genes were randomly selected from all the genes with 5′ intergenic regions. For each control list for 3′ end intergenic ncRNA, 3,166 genes were randomly selected from all the genes with 3′ intergenic regions. All over-represented GO terms (≥5 genes and FDR<0.05) were highlighted as yellow in Table S3.

Analysing the sequence conservation of predicted ncRNAs

Conservation analysis based on phastCons score [22]: The reference phastCons score files containing the phastCons scores for multiple alignments of 4 other vertebrate genomes (Dog, May 2005, canFam2; Human, Mar 2006, hg18; Mouse, July 2007, mm9; Platypus, Mar 2007, ornAna1) to the reference of cow genome (Oct 2007, bosTau4) were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/phastCons5way/). Each base in the EST or RefSeq was assigned a phastCons score according to the reference files. The bases that were not included in the conserved elements of the reference files were given phastCons scores of "0". For a given sequence, the mean phastCons score was calculated by normalizing the sum of phastCons scores against the length of the sequence.

Conservation analysis based on GERP++ score [23]: GERP++ is another tool that uses maximum likelihood evolutionary rate estimation for position-specific scoring. It calculates the RS (rejected substitution) score based on multiple alignments and a phylogenetic tree. The 5-way multiple alignment file for cow (the same species and genome assemblies used for phastCons scores)

and the corresponding phylogenetic tree were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/multiz5way/). A PERL script was created to convert the default multiple alignment file format into the file format that can be fed into GERP++. The GERP++ score for each base of bosTau4 was calculated using GERPv2.1 (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html). Mean GERP++ scores were calculated in the same way as mean phastCons scores.

24,000 genomic fragments, which ranged in size from 500 bp to 15,000 bp, were randomly extracted from un-transcribed regions of bosTau4 as the control dataset. The cumulative frequency for each dataset was calculated and plotted using the R package.

Identification of sequence specific motifs from intergenic ncRNAs

Bovine gene expression profiles were generated based on transcriptome data from 95 samples (92 adult, juvenile and fetal cattle tissues and 3 cattle cell lines) [24].

FIRE was used to predict sequence motifs from bovine intergenic ncRNAs [25]. Bovine intergenic ncRNAs located in 5 kb of upstream or downstream gene flanking regions were used as motif prediction pools. Intergenic ncRNAs were converted as sense RNAs according to their transcription orientation. The motif-identification mode was set as "DNA", which means motif sequence can be predicted from both strands of intergenic ncRNAs. FIRE was run against 5′ end and 3′ end intergenic ncRNAs according to 95 individual gene expression profiles respectively.

The comparison of predicted RNA sequence motifs against known DNA motifs was performed using the TOMTOM web server [26].

Expression correlation analysis based on bovine MPSS data

The expression profiles of intergenic ncRNAs and bovine RefSeqs were calculated based on the MPSS (Massively Parallel Signature Sequencing) tags mapped to the 3′ most end of each transcript [24]. The tag count for each transcript was normalized according to the library size. Transcripts mapped with less than 3 tags were removed from the expression profile. The MIC score was generated by MINE based on the expression of intergenic ncRNA and RefSeq pairs [27]. Only intergenic ncRNAs/RefSeqs with expression (read counts) in at least 3 libraries were used to perform expression correlation analysis.

Results

The development of ncRNAs identification pipeline

We identified ncRNAs from bovine ESTs, by developing a computational pipeline based on public software and Perl scripts (Figure 1). A total set of 1,517,143 bovine ESTs (as of 30th September, 2009), extracted from the dbEST of NCBI, was processed as the input dataset for the pipeline. After quality control, repeat filtration and EST assembly, we identified 216,095 unique transcripts. We opted for stringent mapping criteria (coverage $\geq$90% and identity $\geq$95%) and as a result, 69,099 unique transcripts were unable to be mapped to the BosTau4 assembly and were therefore discarded. Of the mapped sequences, 3,121 were classified as putative *cis*-NATs, 74 of which were subsequently manually checked on UCSC genome browser (Materials S1). The remaining 143,875 mapped unique transcripts were further analysed to annotate and characterize the bovine transcriptome.

Of the 143,875 mapped unique transcripts, 87,373 were very similar to bovine RefSeqs (E-value$<$1e-3), and 48,773 of them shared similarity over more than 90% of their length with 14,962 RefSeqs and were denoted as known gene transcripts. Of the 38,600 sequences that shared similarity with RefSeqs over less than 90% of their length, more than one third (13,035) were un-spliced.

There were 1,856 transcripts, which we were unable to annotate based on similarity search against bovine RefSeqs, but were identified by BLAST in the Swiss-Prot database at the amino acid level. These sequences may represent novel un-annotated bovine protein-coding genes that are conserved across taxa.

The resulting set of sequences, filtered with respect to sequence similarity to repeats, protein-coding transcripts and *cis*-NATs was then further scrutinized by checking the length of predicted ORFs (Open Reading Frames). As a result, 31,586 unique sequences were removed from the 54,646 "protein-coding gene filtered unique transcripts" because they contained either long predicted ORFs ($\geq$100 amino acids) or shorter ORFs ($\geq$50 amino acids) at the ends. These "ORF-containing sequences" may include transcripts from un-annotated, novel protein-coding genes. The large number of these transcripts raises the possibility that there are still significant numbers of protein-coding genes in the bovine genome that remain undiscovered.

As a result of this highly stringent filtering against known protein-coding genes and the exclusion of ORF containing transcripts we were left with 23,060 ncRNAs (Table S4), which accounted for ~15.5% (23,060 out of 143,875) of the mapped bovine unique transcripts. These ncRNAs were then analysed to identify previously annotated ncRNAs.

Few well-characterized ncRNAs were identified

The annotation of the 23,060 ncRNAs was carried out using several different methods (See methods for detailed procedures). As a result of this effort we determined that only 77 of these sequences had been previously identified as ncRNAs, either as miRNAs, snoRNAs, tRNAs, rRNAs, mRNA-like ncRNAs, piRNAs and other ncRNAs (Materials S1, Table S5 and Table S6). One additional class of ncRNAs that we identified were *cis*-NATs. We identified 74 *cis*-NAT*s* distributed on 28 different chromosomes (Materials S1 and Table S7 and Figure S1).

Whilst our results showed that ESTs could be used to identify ncRNAs by rational and stringent sequence similarity searches, the vast majority of the ncRNAs we identified could not be annotated based on previously well-characterized ncRNAs.

Genome-wide distribution of ncRNAs

To understand the distribution of predicted ncRNAs in the genome, our 23,060 predicted ncRNAs mapped onto BosTau4 were compared to the mapped locations of 24,373 bovine RefSeqs. Figure 2 shows the density distributions of ncRNAs and RefSeqs in 30 bovine chromosomes (29 autosomes and X). Together with the relative frequencies of the densities of ncRNAs and RefSeqs, which are shown in Figure 3, it is obvious that the "gene poor regions" (with fewer than 10 genes in 1 Mb) are more abundant than "ncRNA poor regions" (less than 10 ncRNA s in 1 Mb) in the bovine genome. Furthermore, 288 gene deserts (no gene in 1 Mb) were identified compared to 156 ncRNA deserts (no ncRNA in 1 Mb). At the other end of the gene density spectrum, 21 regions were found with more than 50 genes/Mb, but no comparable regions were found for ncRNAs. These results showed that ncRNAs were more evenly distributed than protein-coding genes across the genome. A correlation analysis of the densities of protein-coding genes and ncRNAs per 1 Mb revealed only a

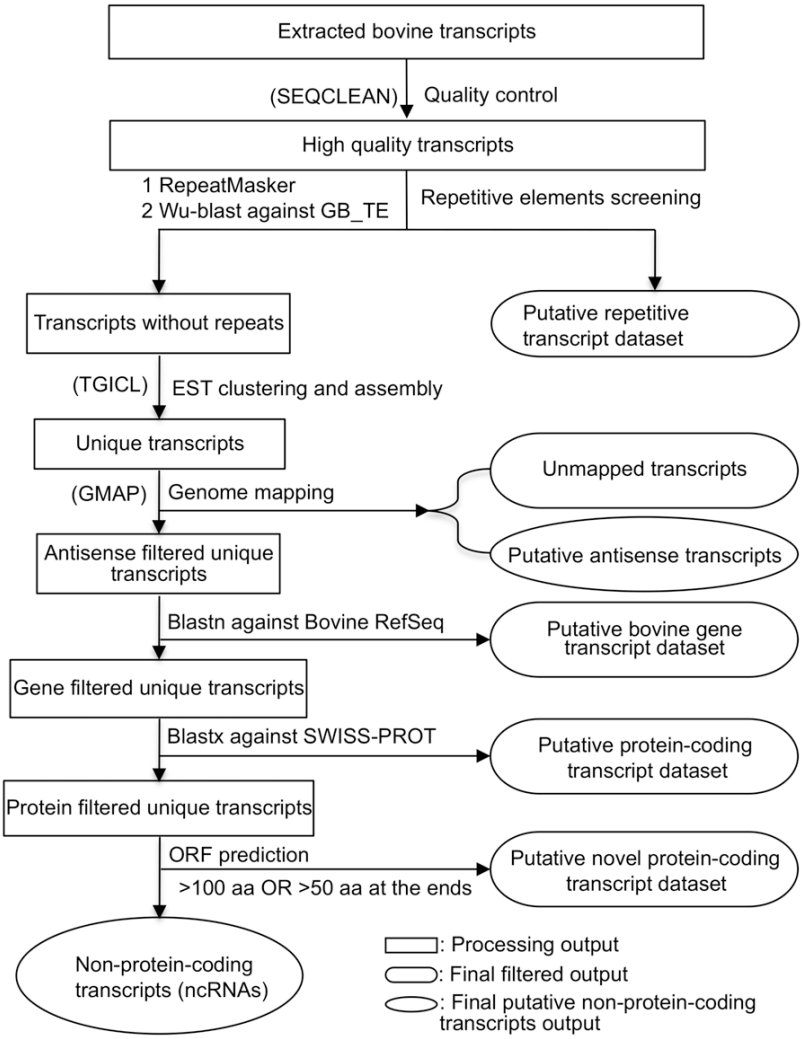

Figure 1. Flowchart describing the pipeline for ncRNA identification.
doi:10.1371/journal.pone.0042638.g001

moderate correlation between these two transcriptome sets at the whole genome level ($r = 0.5528816$, $p<2.2e\text{-}16$).

We further classified our ncRNAs with respect to neighbour protein-coding genes to analyse the potential transcriptional overlap with RefSeq genes. Our classification scheme for ncRNAs is shown in Figure 4. Excluding 952 ncRNAs mapped to uncharacterized genomic locations, there were three main types of ncRNAs based on this classification and their relative proportions are shown in Figure 5. The majority of the ncRNAs in our dataset were intergenic transcripts (57% intergenic compared to 42% intronic). We also noticed that most ncRNAs were singletons (72.2% out of intergenic, 81.1% out of intronic and 71.3% out of overlapped ncRNAs respectively)(Table 1). The data in Table 1 also showed that the vast majority of ncRNAs (both intergenic and intronic) were apparently unspliced transcripts.

Detailed inspection of overlapped ncRNAs revealed that 98 of them overlapped with their corresponding genes by less than 50 basepairs; 85 of them at the 3′ end, and the rest at the 5′ end of the genes. These ncRNAs may represent unannotated UTRs or 5′ and 3′ extensions of genes [28], but there is the possibility that some of them, especially 5′ overlapped ncRNAs, were transcribed as functional ncRNAs, like PASRs, tiRNAs or uaRNAs [10,11,29,30]. Our result did show that there are antisense transcripts among these overlapped ncRNAs (10 of 85 at 3′ end and 3 of 13 at 5′ end).

Most ncRNAs were of intergenic origin

Most bovine ncRNAs mapped to intergenic regions (Figure 5). To get a better understanding of these intergenic ncRNAs, we plotted the frequency distribution of intergenic ncRNAs as a function of their distance and transcriptional orientation to the nearest neighbour genes (Figure 6). About 67.4% (8,500 out of 12,614) of intergenic ncRNAs had a neighbour gene within 20 kb, with a significant concentration of intergenic ncRNAs in the 5 kb flanking regions of genes. Beyond 10 kb, the number of intergenic ncRNAs decreased very gradually as a function of distance. It was also apparent from Figure 6A that intergenic ncRNAs were more prevalent at the 3′ end of genes than at the 5′ end. The intergenic ncRNAs closest to the 5′ end of a gene also tended to be within

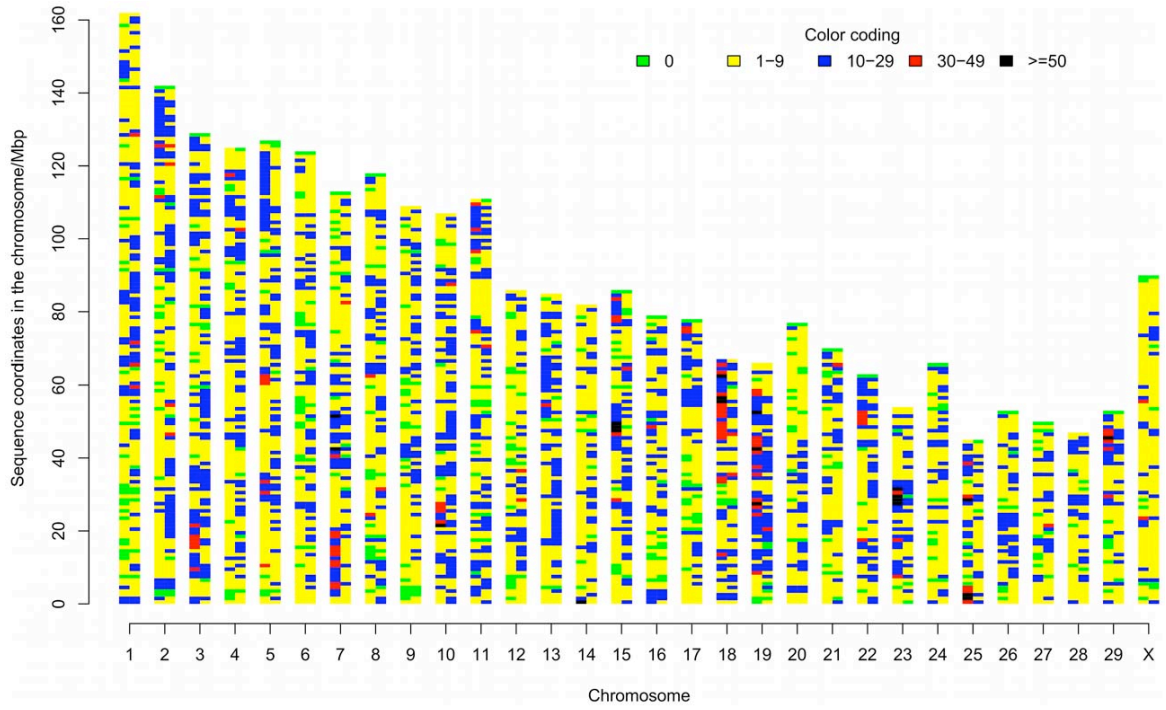

Figure 2. Distribution of genes and ncRNAs in the bovine genome. Chromosomes are on the X axis, and sequence coordinates on the Y axis, with the "top" of the chromosome at the Y axis origin. All cattle autosomes are acrocentric. Each chromosome is represented by two vertical bands, the left band shows gene number and the right band shows ncRNA number, per 1 Mb bin. The legend shows the band colour coding for numbers per 1 Mb bin.
doi:10.1371/journal.pone.0042638.g002

5 kb of the gene, but this localization was not significantly different to the control frequency distribution calculated using gene to gene nearest neighbour distances, where the majority of intergenic distances were less than 5 kb. We were able to determine transcriptional orientation of 10,969 of 12,614 intergenic ncRNAs based on their dbEST annotation. When we compared the transcriptional orientation of these intergenic ncRNAs to their closest gene neighbour, we observed that most of them closest to the 3′ end of genes were transcribed from the same strand as the gene (Figure 6B). There were four times more ncRNAs in the same transcriptional orientation when they were 3′ to the closest gene (6,296 to 1,433). This difference in transcriptional orientation for the ncRNAs 5′ of the closest gene was also observed, but not to the same degree (1,931 same to 1,309 reverse). The intergenic ncRNAs, transcribed from the same strand as the closest gene, might be extensions of the UTRs produced by alternative transcription start or termination sites of protein-coding genes, but many of them were at significant distances from these genes making this an unlikely possibility.

To determine the likelihood that these intergenic ncRNAs were potential gene UTRs, we compared them against the annotated UTR database (including human, mammals and vertebrates) [31]. 3,168 of these intergenic ncRNAs were highly similar to 3′ UTRs (E-value<1e-3), while only 198 were highly similar to 5′ UTRs (E-value<1e-3). Together with 2,516 intergenic ncRNAs which are located in the proximal 1 kb of gene flanking regions (5′ end or 3′ end), we classified these 4,584 intergenic ncRNAs as UTR-Related RNAs (Table S4), which are named to differentiate them from uaRNAs (UTR-associated RNAs), a class of previously annotated independent ncRNAs transcribed from UTRs [30]. The reasonably large number of intergenic ncRNAs transcribed in the opposite orientation to their nearest gene (1,309 from the 5′ end and 1,433 from the 3′ end), raised the possibility that there might be transcriptional antisense regulation associated with these elements.

The spatial clustering of all predicted intergenic ncRNAs with respect to protein coding genes suggested a *cis*-regulatory relationship to us. To understand the potential biological significance of such a relationship, we functionally clustered the neighbour genes within 5 kb flanking regions of intergenic ncRNAs according to GO [32]. We found that regulatory genes were over-represented in the neighbour genes of these intergenic ncRNAs (Table S3), but the gene count of these over-represented GO terms was very small, most likely because of the poor functional annotation of bovine reference genes in GO. The functional clustering of control gene lists (see methods) indicated these over-representations were not chance occurrences (Table S3). When we differentiated the neighbour genes according to the position of their nearby intergenic ncRNAs, we observed that positive regulatory genes were over-represented in the neighbour genes with intergenic ncRNAs in their 5′ flanking regions (Table S3). Assessment of neighbour gene function based on regulatory-related keywords searching of the subset of 183 genes flanked at both ends by intergenic ncRNAs revealed that 85 (46.4%) of these genes were involved in either transcriptional regulation, signal transduction or encoded domains consistent with these functions. By comparison, only 8,087 (33.2%) of all 24,373 RefSeq genes were annotated as regulatory genes based on the same keywords searching. This indicated that the purely GO-based results were probably a significant underestimate of the regulatory potential of

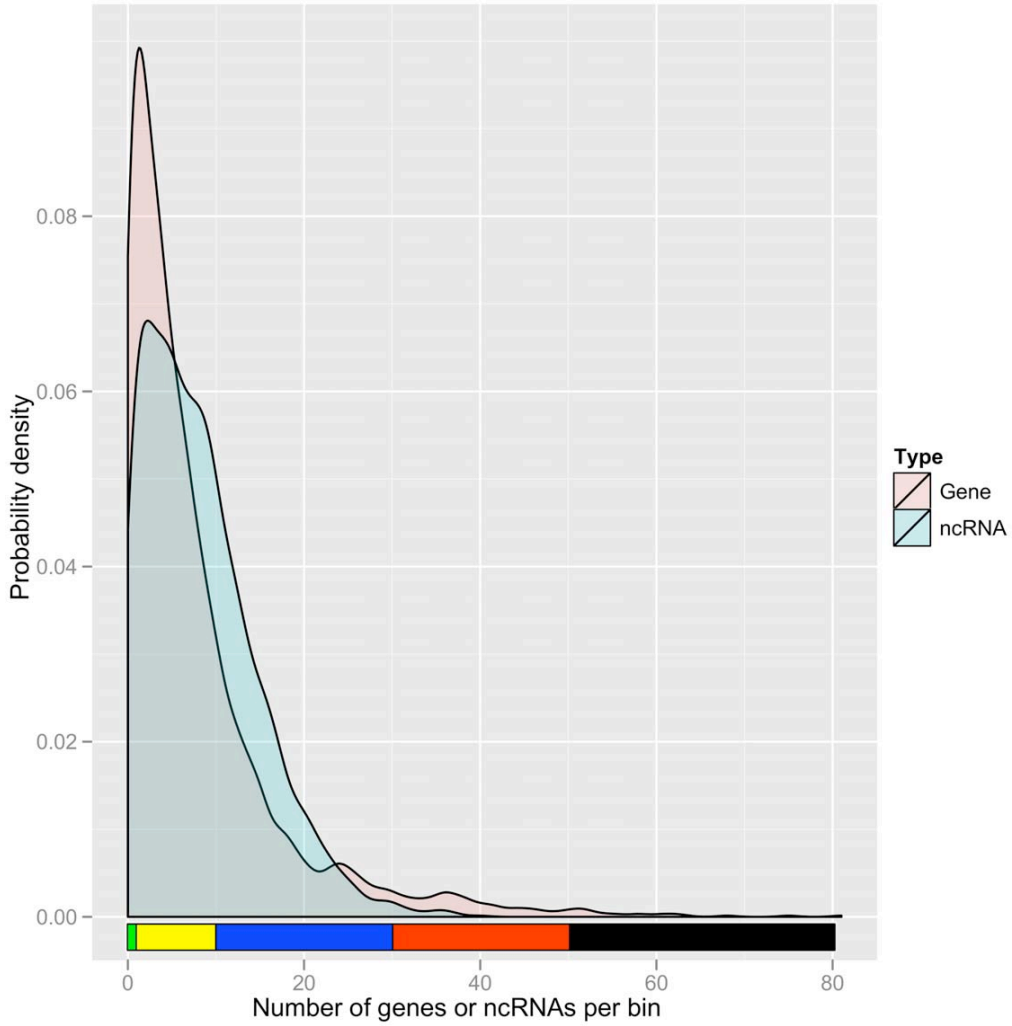

Figure 3. Probability densities of genes and ncRNAs per 1 Mb bin. NcRNAs have similar genomic densities compared to protein coding genes, but with fewer extreme density regions. The colour coding is consistent with Figure 2.
doi:10.1371/journal.pone.0042638.g003

these neighbour genes. In summary, we hypothesize that our gene-proximate intergenic ncRNAs are potentially *cis*-regulatory and tend to regulate regulatory genes. Confirmation of this hypothesis will have to await specific, functional perturbation experiments, but is consistent with published data from small numbers of intergenic ncRNAs.

Evolutionary conservation of bovine ncRNAs

To assess whether ncRNAs were under selective constraint, we used two different methods to assess the degree of sequence conservation as shown in Figure 7. Figure 7A shows the degree of conservation based on phastCons score; ncRNAs were clearly conserved compared to control sequences, which were selected at random from un-transcribed regions of the bovine genome, but were less conserved compared to protein-coding genes. When we compared the degree of sequence conservation between intergenic and intronic ncRNAs according to phastCons score (Figure 7B), intergenic ncRNAs were more conserved than intronic ones. When we further refined this to assess the sequence conservation of intergenic ncRNAs according to their relationships with protein-coding genes, we observed that intergenic ncRNAs closest to the 3′ end of genes were more conserved than those closest to the 5′ end of genes. And when we took into the consideration the transcriptional orientation of these ncRNAs with respect to their closest gene, the "sense" intergenic ncRNAs, which are transcribed from the same strand as their neighbour genes, were more conserved than the "antisense" intergenic ncRNAs, regardless of whether they were closest to the 5′ or 3′ end of protein-coding genes (Figure 7C).

We were able to confirm these observations regarding the conservation level of ncRNAs using GERP++ [23], based on a different statistical model. If we only consider the sequences that were under a substitution deficit (positive score), the conservation level of ncRNAs was between protein-coding genes and un-transcribed genomic fragments, which was consistent with the phastCons result. Nearly 40% of ncRNAs had a substitution deficit, compared to ~80% of protein-coding genes and less than 20% of un-transcribed genomic fragments. On the other hand, for sequences that showed a substitution surplus (negative score), the divergence level of ncRNAs was more pronounced than for

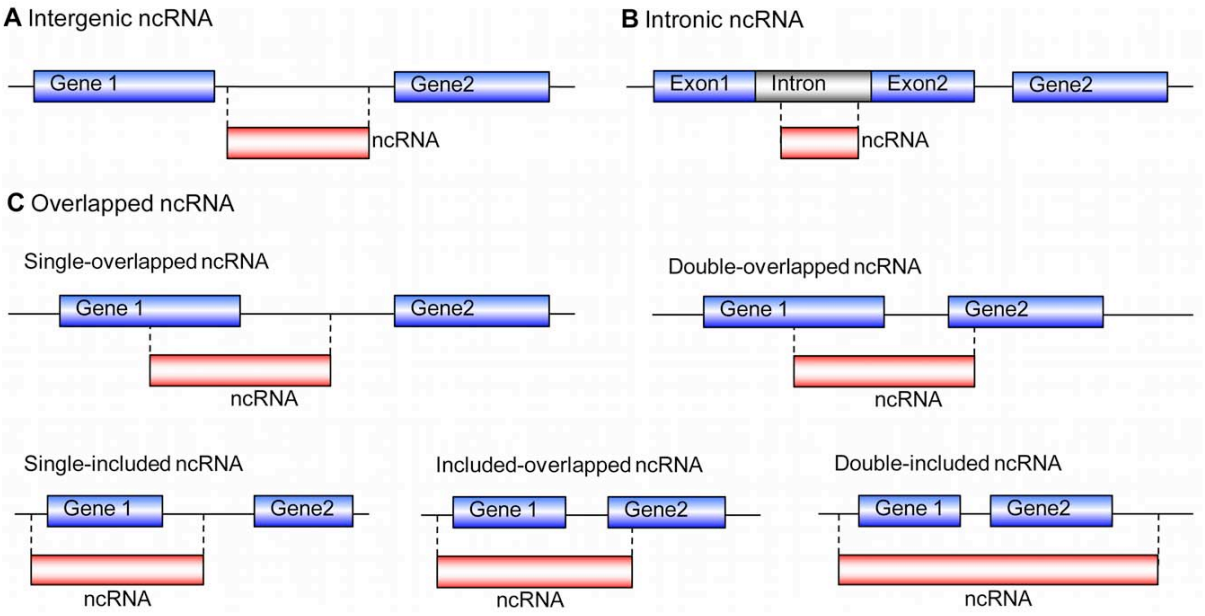

Figure 4. Classification of ncRNAs in relation to protein-coding genes. (A) The entire EST is transcribed from an intergenic region, regardless of the transcription orientation. (B) The entire EST is transcribed from an intron, regardless of the transcription orientation. (C) Single-overlapped ncRNA: EST partially overlapped with a gene; Double-overlapped ncRNA: Both ends of the EST overlapped with two genes and spanned an intergenic region; Single-included ncRNA: The gene was fully included inside the EST; Included-overlapped ncRNA: One gene was fully included within the ncRNA, and the ncRNA spanned the intergenic region and overlapped with a neighbour gene; Double-included ncRNA: More than one genes were fully included within the EST.
doi:10.1371/journal.pone.0042638.g004

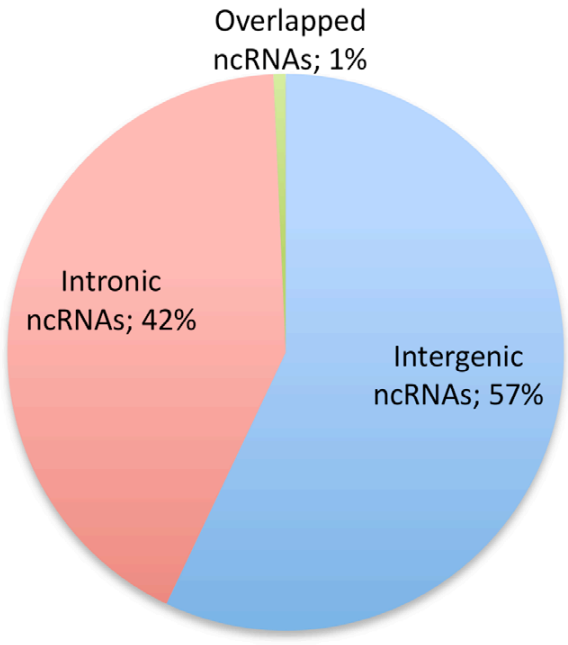

Figure 5. Relative abundance of the three main classifications of ncRNAs. Almost 60% of ncRNAs are long intergenic non-coding RNAs (intergenic ncRNAs).
doi:10.1371/journal.pone.0042638.g005

protein-coding genes and un-transcribed genomic fragments (Figure 7D). The results of the GERP++ score for the intergenic and intronic ncRNAs, as well as the different intergenic classes were also consistent with their respective phastCons score results (Figure 7E and Figure 7F).

When we removed all UTR-related RNAs from 23,060 ncRNAs, the remaining sequences still showed clear conservation compared to un-transcribed control fragments (Figure S2). The highly conserved UTR-related RNAs is consistent with these being part of poorly annotated UTRs or independent transcripts from UTRs, as UTRs across different species are often well conserved (Figure S2).

Table 1. Summary of transcriptional redundancy and splicing information of three types of ncRNAs.

Class of ncRNAs	Number	Singleton		Unspliced	
		Count	Fraction	Count	Fraction
Intergenic	12,614	9,113	72.2%	9,852	78.1%
Intronic	9,337	7,571	81.1%	8,085	86.6%
Overlapped	157	112	71.3%	80	51.0%
–Single-overlapped	138	96	69.6%	78	56.5%
–Double-overlapped	2	2	100%	0	0
–Single-included	10	9	90%	1	10%
–Included-overlapped	2	2	100%	0	0
–Double-included	5	3	60%	1	20%

– denotes subclass of Overlapped.
doi:10.1371/journal.pone.0042638.t001

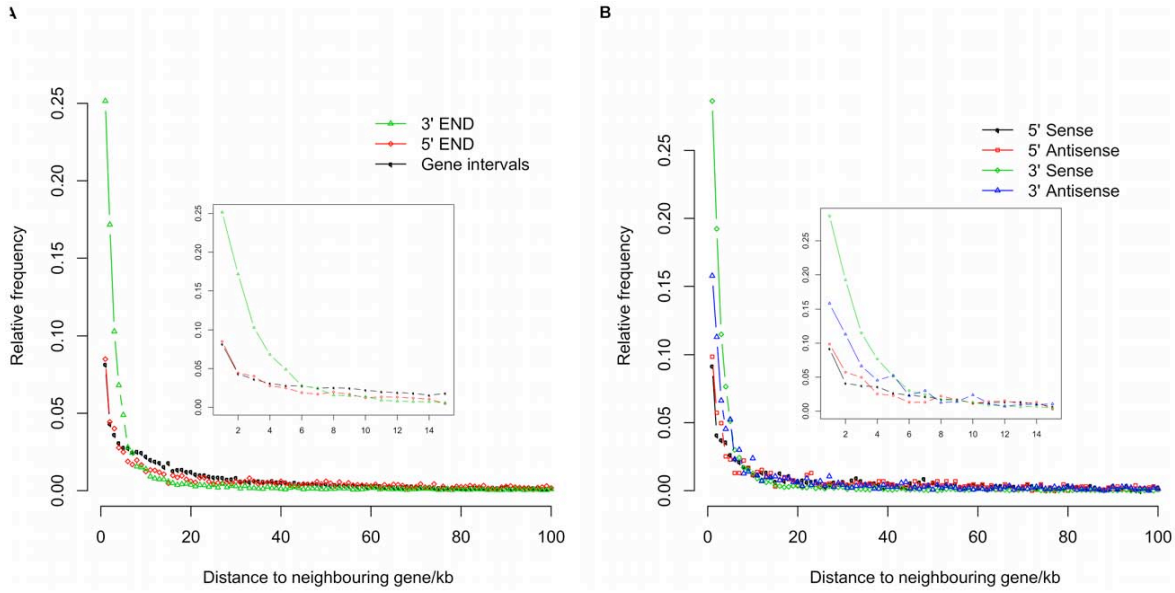

Figure 6. Positional bias distribution of ncRNAs with respect to neighbour genes. (A) Relative frequencies of ncRNAs with respect to the distance from neighbour genes. 100 kb adjacent to TSS or TTS of genes is shown in these plots. 3′ END means the ncRNA is located in the 3′ flanking region of its neighbour gene. 5′ END means the ncRNA is located in the 5′ flanking region of its neighbour gene. "Gene intervals" refers to the intergenic region of two adjacent genes. (B) Relative frequencies of ncRNAs from neighbour genes partitioned with respect to transcription orientation. The internal boxes represent the zoom in view of the relative frequencies from 5 kb to 20 kb.
doi:10.1371/journal.pone.0042638.g006

Identification of sequence motifs from intergenic ncRNAs

Based on the gene expression profiles generated from 95 bovine transcriptome libraries, we identified 21 sequence specific motifs from 5′ intergenic ncRNAs and 29 from 3′ intergenic ncRNAs (Table S8, A & B). By comparison against known DNA motif databases using TOMTOM, we found that 2 motifs, "160_1_5END" from 5′ end intergenic ncRNAs and "086_1_3END" from 3′ end intergenic ncRNAs, showed significant similarity against known DNA motifs "ste11" and "ARF" respectively (p-value<1e-04 and FDR<0.05) (Figure 8 and Table S8). It is interesting to note that the number of "sense" sequence motifs of "ste11" (the motif is the same as the intergenic ncRNA strand) is almost equal to the number of "antisense" "ste11" motifs (the motif is complementary to the intergenic ncRNA strand) (Table S8, A & B). 3 other motifs from 5′ intergenic ncRNAs and 4 from 3′ intergenic ncRNAs also showed strong similarity (p-value<1e-04, FDR<0.5) against known DNA motifs (Figure S3 and Figure S4). The numbers of "sense" and "antisense" sequence sites in intergenic ncRNAs are almost equal for most of the identified motifs (Table S8, A & B and Figure S5).

After we removed all UTR-related RNAs from the 5 kb intergenic ncRNAs and re-ran the motif identification procedure with the same expression profiles and parameters, we still found 15 and 17 motifs from the remaining 5′ and 3′ intergenic ncRNAs. However, all of these novel 32 motifs were different to the 50 originally identified motifs (Table S8, C & D). Only one novel 3′ motif (136-1, [ACT]AG[AC]CATA[AGT]) showed similarity with a known DNA motif FOXL1, which was also the best hit for an originally identified 3′ end motif (119_1_3END, [ACT]AAA[CT]ATA[GT]).

Expression correlation and functional significance

Most of the identified intergenic ncRNAs reported from other species were directly or indirectly involved in gene regulatory networks. To understand whether there are correlations between the expression of intergenic ncRNAs and corresponding neighbour genes, we identified all intergenic ncRNA and neighbour gene pairs with expression in at least one library based on the 95 bovine MPSS transcriptome data. Globally, there was no clear correlation between the expression of intergenic ncRNAs and corresponding neighbour genes no matter whether intergenic ncRNAs were at the 5′ end or 3′ end of the genes (Figure 9). Because many intergenic ncRNAs containing sequence motifs are also close to regulatory genes, we checked the expression of these "motif and regulatory" intergenic ncRNAs across different libraries (Figure S6). Some of these intergenic ncRNAs showed negative expression correlation with neighbour genes. One of these intergenic ncRNAs is the antisense transcript of protein-coding gene "*ZNFX1*" (Figure S6). In human, the antisense transcript of "*ZNFX1*" has been annotated as "*ZNFX1-AS1*" [33]. This antisense transcript in bovine might be the homolog of the human "*ZNFX1-AS1*". This bovine *"ZNFX1-AS1"* does not show high sequence conservation with 4 different human transcript variants (Figure S7). It is also the host transcript of two possible snoRNAs (SNORD12 and SNORD12B), which is consistent with human *"ZNFX1-AS1"* (Figure S8) [33].

To understand the associations between the expression of intergenic ncRNAs with other protein-coding genes, we used MINE (Maximal Information-based Nonparametric Exploration) to analyse the correlations between each intergenic ncRNA and all RefSeq genes [27]. For most intergenic ncRNAs detected by the RNA-seq data (191 out of 389 at 5′ end and 1,678 out of 2,673 at 3′ end), we identified significantly associated protein-coding genes based on MIC (Maximal Information Coefficient) score, with FDR≤0.05 after multiple testing (Table S9), and many of these showed significant associations with multiple protein-coding genes in terms of their expression, with 35 out of 191 5′ intergenic ncRNAs and 425 of 1,678 3′ end intergenic ncRNAs correlated

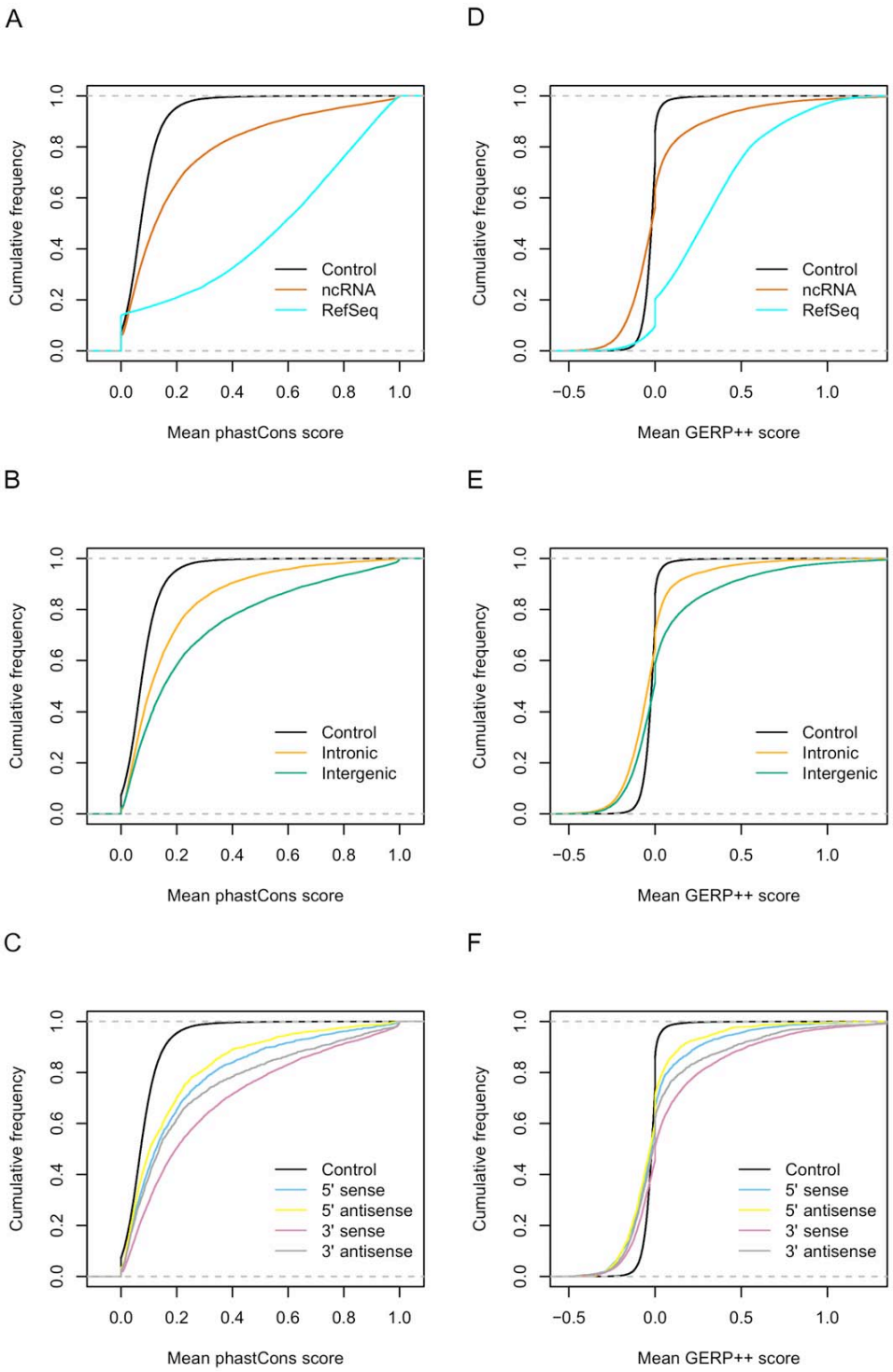

Figure 7. Sequence conservation analysis of ncRNAs. (A, B & C) are based on phastCons score. (D, E & F) are based on GERP++ score. The control line is based on a similar number of randomly selected non-transcribed genomic regions. A & D – ncRNAs compared to RefSeqs, B & E – intergenic ncRNAs compared to intronic and C& F – 5′ vs 3′ ncRNAs and transcriptional orientation with respect to nearest neighbour genes.
doi:10.1371/journal.pone.0042638.g007

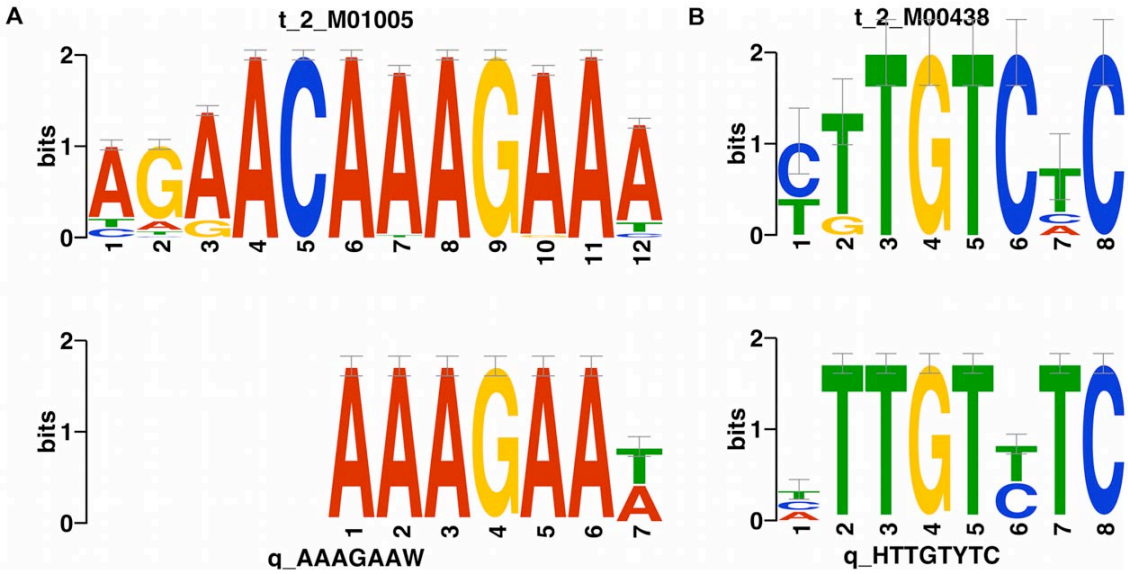

Figure 8. Two sequence motifs from intergenic ncRNAs with significant similarity against known DNA motifs. For each comparison, the upper one is the known DNA motif, and the lower one is the intergenic ncRNA sequence motif.
doi:10.1371/journal.pone.0042638.g008

with their neighbour genes (Table S9). 78 of the 191 5′ intergenic ncRNAs and 1,124 of the 1,678 3′ end intergenic ncRNAs were UTR-related RNAs.

Discussion

Identification of ncRNAs

While increasing numbers of studies have confirmed that ncRNAs possess significant regulatory functions in different biological pathways, their computational identification can be very challenging. One current approach is to identify ncRNA based on homology searches, such as sequence-based, profile HMM and structure enhanced methods [34,35,36]. Compared to these methods, our pipeline for ncRNA identification has two advantages [37]. First, our ncRNAs were identified from transcriptome data. Most homology-search-based methods use the entire genome sequence as the starting point, so it is not obvious if the ncRNAs identified by these methods are transcribed functional elements. Normally, further experiments are required to validate the expression of these functional elements. Second, most of the homology search methods are based on multi-alignments or taking known ncRNAs as a training set, so the output generated by these programs tends to identify only conserved ncRNAs. Conservation of ncRNAs is not as obvious as mRNAs. Some ncRNAs, like miRNAs, are indeed under strong selective constraint, but more ncRNAs, especially long ncRNAs, seem to be less conserved than protein-coding RNAs. By using stringent filters in our pipeline, we effectively removed the protein-coding transcripts, and identified different kinds of ncRNAs, which were not restricted to conserved ncRNAs. For the time being we have ignored ncRNA transcribed from repetitive elements, mostly retrotransposons, because it is virtually impossible to map such sequences to a unique genomic location and conservation scores for such sequences are only available for ancestral retrotransposon insertions. However retrotransposon ncRNAs may also be functional, as previous investigators have shown that transcripts of retrotransposon origin are differentially regulated during development [38].

The existence of well-characterized ncRNAs in our ncRNA dataset indicated that our pipeline was effective but also illustrated how few ncRNAs were conserved on the basis of sequence similarity. To avoid false positives, we relied on stringent criteria. For example, when mapping transcripts to the genome, only transcripts mapped with more than 90% coverage and greater than 95% identity were kept for further analyses. This explains why approximately 32% of the unique transcripts were classified as "un-mapped" transcripts. These criteria ensured that we removed contaminating and error rich sequences. Subsequently, when filtering protein-coding genes using BLAST, transcripts with hits (E-value$<$1e-5), regardless of coverage or percent identity in bovine RefSeq or Swiss-Prot databases, were discarded. This ensured that un-annotated distant paralogs or pseudogenes along with protein-coding ESTs were removed from our ncRNA set.

As a result, our pipeline provides a tool to mine the abundance of ESTs, which were originally used to identify protein-coding genes. Many studies have confirmed that ESTs can be used to detect ncRNAs. The most important evidence is the FANTOM ncRNA dataset, which are mRNA-like ncRNAs identified from mouse cDNAs [4]. NcRNAs identified from ESTs have also been reported in other organisms [39,40]. Recently, a class of human long ncRNAs with enhancer-like function was identified from GENCODE annotation that, in part, relied on ESTs mapped to non-protein-coding regions [9]. Because our analyses were based on such stringent criteria, it is quite likely that our results represent a conservatively low estimate of the number of long ncRNAs in a mammalian transcriptome.

The genome-wide distribution of ncRNAs

According to previous RNA-seq and tiling-array studies, more reads can be mapped to intronic than intergenic regions [5]. In contrast, our data showed that there were more intergenic than intronic ncRNAs in the bovine non-protein-coding transcriptome. Introns are known to be rich sources of both small and long ncRNA transcripts [41], but the larger number of conserved intergenic ncRNAs that we identified indicated that there might be

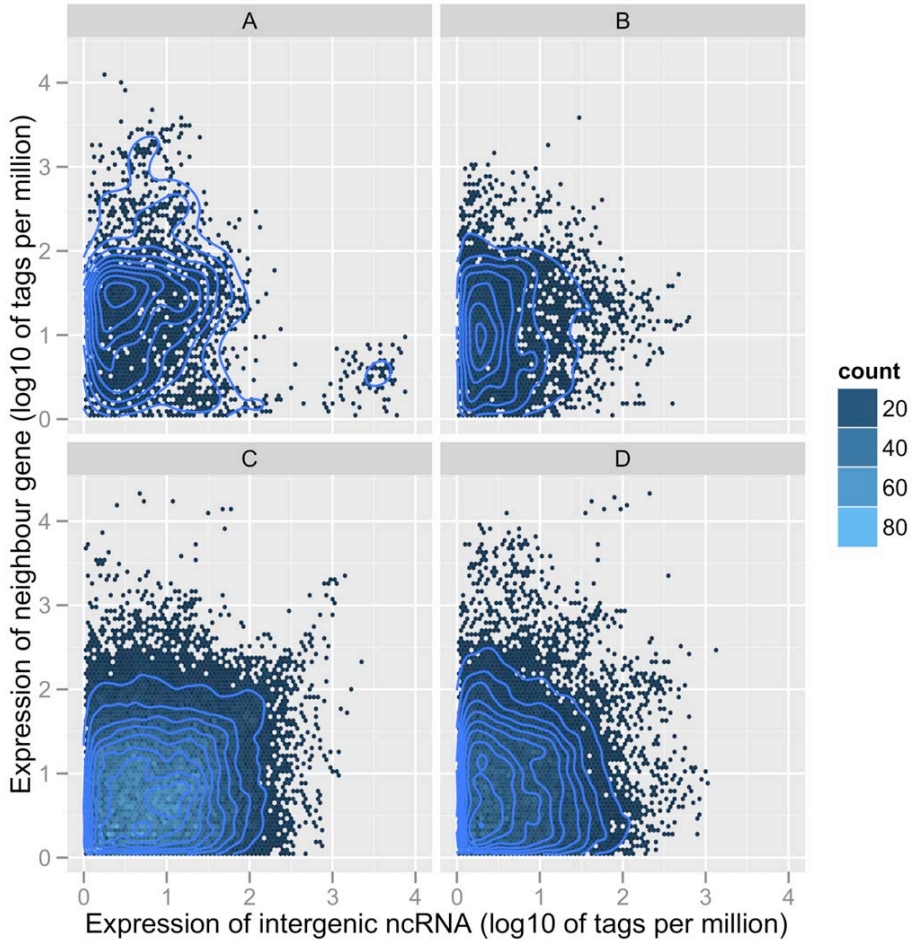

Figure 9. Scatter plot for the log10 ratio of expressions of intergenic ncRNAs and corresponding neighbour genes. Dots were binned into 80*80 hexagons across the plot area. Different colours represent the dot count in each bin. A represents the expression of 5′ end UTR-related RNAs and neighbour genes. B represents the expression of 5′ end intergenic ncRNAs with UTR-related RNAs removed and neighbour genes. C represents the expression of 3′ end UTR-related RNAs and neighbour genes, and D represent the expression of 3′ end intergenic ncRNAs with UTR-related RNAs removed and corresponding neighbour genes.
doi:10.1371/journal.pone.0042638.g009

more functional regulatory transcripts embedded in the intergenic regions of bovine genomes.

Previous research has shown that many ncRNAs are expressed in tissue-specific fashion or are restricted to certain developmental stages [42,43,44], which would likely manifest as singletons in the pooled tissue, normalized EST libraries that account for almost all of the bovine ESTs we analysed. Furthermore, the prevalence of unspliced transcripts (Table 1) was also reported in ncRNAs by Khachane *et al.* in a dataset of functional long ncRNAs [45]. These features may explain that why ncRNAs are not as easily detected as protein-coding genes in many situations.

The genome-wide map of ncRNA distribution in bovine demonstrates that ncRNAs are more evenly spread throughout the genome than protein-coding genes. This may mean that ncRNAs have evolved differently to protein-coding genes, which can form gene-rich regions by gene duplication [46]. This might also partially explain the poor conservation of ncRNAs. The different genomic distributions of ncRNA compared to genes is reflected in the moderate correlation between the densities of ncRNAs and protein-coding genes, indicating that many ncRNAs may function as remote regulatory elements rather than regulating their neighbour genes in some proximity based fashion. Previously, ncRNAs have been experimentally demonstrated to regulate gene expression by influencing the transcription process or chromatin structure in *trans*-acting fashion [47,48,49]. Some of these newly discovered enhancer-like long ncRNAs activate distant genes rather than surrounding ones, at distances in excess of 300 kb [9].

The moderate correlation of ncRNA density with gene density is also reflected in the fact that most bovine intergenic ncRNAs were transcribed from regions near protein-coding genes, especially from the 3′ end. This distribution bias has been observed previously in RNA-seq and tiling array expression experiments [4,29,50]. Our results however, were based on long reads from most tissues and developmental stages and were therefore unlikely to result from short, ragged ends of run-on transcripts. Furthermore, while many of these transcripts were found very near to genes, significant numbers were also found thousands to tens of thousands of base pairs away. Even in the UTR-related RNAs that we classified, there are still a proportion (492 of 4,584) transcribed from the antisense strand of protein-coding genes. Therefore, most of the intergenic

ncRNAs, which were transcribed from both strands near protein-coding genes were inconsistent with trivial explanations such as transcriptional noise or mis-annotated UTRs. We therefore need to consider that these gene proximate intergenic ncRNAs may function as either *cis*-regulatory elements of their neighbour genes or as *trans*-acting regulatory sequences. Previous studies have confirmed that there are functional ncRNAs transcribed from the promoter, transcription start and terminal regions of protein-coding genes in sense orientation [10,11]. Evidence for antisense ncRNAs comes from a recent study, using tSMS (true Single Molecule Sequencing) technology [12,29]. In this study, a novel RNA copying mechanism was proposed, capable of producing antisense poly(U) small RNAs from the transcription start or terminal regions of genes, confirming that some human ESTs result from this process [12]. This is consistent with our results, where a significant fraction of the gene-proximate antisense ncRNAs were mapped very close to the 3′ ends of genes. However, while the functional significance of such antisense transcripts is unknown, this copying mechanism does not explain the significant fraction of gene proximate ncRNAs originating from the antisense strand much further away from the 3′ ends of genes. Even for the intergenic ncRNAs close to 3′ end neighbour protein-coding genes, in the same transcriptional orientation, which might be transcribed from potential un-characterized UTRs, there is also the possibility that they are independent functional transcripts, which have been observed mostly in human, mouse and fly genomes, and classified as uaRNAs [30]. On balance it is difficult to come up with a reasonable, consistent and trivial explanation for the occurrence of non-coding transcripts such as our ncRNAs leading us to conclude that they have a biological purpose.

Conservation level of ncRNAs

The vast majority of the ncRNAs we have identified did not have detectable sequence similarity with well-annotated ncRNAs. However, in general, the conservation analysis of bovine ncRNAs based on phastCons and GERP++ score showed that ncRNAs were less conserved than protein-coding genes, while still exhibiting strong selection signatures. Our result was consistent with previous studies, which demonstrated that ncRNAs might experience different selective constraints compared to protein-coding genes [7,9,51]. Our result was also consistent with the possibility that ncRNAs might represent different ncRNA categories, each manifesting different levels of sequence conservation.

We observed that intergenic ncRNAs were slightly more conserved than intronic ones. This finding indicated that there might be more functional elements transcribed from the intergenic regions of the genome, such as recently discovered novel ncRNAs, including uaRNAs, PASRs, lincRNAs and enhancer-like RNAs, identified from intergenic regions [7,9,10,11,30].

Sequence specific motifs identified from intergenic ncRNAs

Previous studies have reported that there are small or long ncRNAs transcribed from gene regulatory elements, like promoter regions. A report from Hans *et al.* showed that there are ncRNAs transcribed from promoter regions, which were named promoter-associated RNAs [52]. These promoter-associated RNAs function as recognition motifs to direct epigenetic silencing complexes to the promoter regions of target genes. Promoter-associated RNAs can also interact with transcription factor recognition sites to form DNA:RNA triplexes, which then interact with the rDNA promoter, mediating recruitment of DNMT3b and silencing rRNA genes by epigenetic regulation [53]. The location of these 5′ end bovine intergenic ncRNAs with respect to their corresponding neighbour genes and the existence of common sequence motifs indicate that these sequence motifs from intergenic ncRNAs may function as recognition sites for RNA-binding proteins, which form an RNA-protein complex to modulate target gene expression. Some sequence motifs from our 5′ end intergenic ncRNAs showed strong similarity with known DNA motifs and the almost equal numbers of sense and antisense motifs distributed in these transcribed 5′ end intergenic ncRNAs indicated that they might be compatible with different regulatory models. Both the sense and antisense sequence motifs could bind with known DNA motifs to form DNA:RNA triplexes that regulate gene expression as above. Alternatively, it could also be the transcription of the intergenic ncRNAs themselves that interferes with the binding of transcription factors to target sites in promoter regions. It has been reported that sequence motifs are widely distributed in the 3′ UTRs of protein-coding genes. They tend to be recognition sites of RNA-binding proteins or target sites of miRNAs, which play important function in mRNA stability or degradation [54]. The existence of sequence motifs in intergenic ncRNAs indicates that a similar regulatory system may also involve non-coding RNAs.

Expression correlation and functional significance

The poor expression correlation between intergenic ncRNAs and their neighbour genes does not mean that they lack functional significance. There are three arguments that support this view. First the observed dynamic range of MPSS tag abundance for intergenic ncRNAs was very similar to that of RefSeq tags. This implies that similar levels or types of regulation exist for intergenic ncRNAs and mRNAs. Second, the bovine MPSS expression profiles we analysed were generated from multiple sources, including different tissues/cell lines, different developmental stages and different sexes [24]. Studies have confirmed that intergenic ncRNAs tend to be expressed in tissue-specific or development-specific ways [55,56]. Intergenic ncRNAs in different tissues or developmental stages may be either repressed or activated. This will make the expression correlation fuzzy and unpredictable when these stages are pooled for analysis. Third, intergenic ncRNAs might represent a wide spectrum of functional non-coding RNAs. Different classes of ncRNAs use different mechanisms to regulate gene expression. Some intergenic ncRNAs that are *cis*-regulators might have strong correlations with their neighbour genes. While intergenic ncRNAs functioning in *trans* might show poor correlation with their neighbour genes. The MIC scores for each intergenic ncRNA with all RefSeqs confirmed that many intergenic ncRNAs showed strong correlations with a number of non-neighbour protein-coding genes, which indicated that intergenic ncRNAs might have multiple targets and be involved in multiple gene-regulation networks. In human, mouse and zebrafish, studies based on RNA-seq have also shown that there is no strong expression correlation between intergenic ncRNAs and neighbour genes at the global level [55,56].

In conclusion, we have demonstrated that EST data sets can be useful for identifying ncRNAs or ncRNA precursors. Genomic distribution and conservation analysis of ncRNAs suggested that these transcripts were not of trivial origin and most originated from genomic regions exhibiting signatures of negative selection or conservation. Our results support the view that most ncRNAs are functional in the context of the regulon hypothesis [57] and that further studies should be aimed at validating this experimentally. Finally we speculate that some of the gene proximate ncRNAs we have identified may act as *cis*-regulatory gene expression elements of regulatory genes through some as yet unknown mechanism(s), but that most of them may be *trans*-acting.

Supporting Information

Materials S1 Supporting results and methods.
(DOCX)

Figure S1 Classification of *cis*-NATs identified by pipeline. The top line denotes three sequentially distributed gene models, in which arrows represent the direction of transcription.
(TIF)

Figure S2 Most ncRNAs are still conserved after removed UTR-related RNAs. "URTs" represent "UTR-related RNAs", which include 4,584 intergenic ncRNAs.
(TIF)

Figure S3 Three sequence motifs from 5′ intergenic ncRNAs with strong similarity against known DNA motifs. For each comparison, the upper motif is the known DNA motif, and the lower one is the sequence motif from intergenic ncRNA.
(TIF)

Figure S4 Four sequence motifs from 3′ intergenic ncRNAs with strong similarity against known DNA motifs. For each comparison, the upper motif is the known DNA motif, and the lower one is the sequence motif from intergenic ncRNA.
(TIF)

Figure S5 The sequence motifs identified from intergenic ncRNAs tend to have equal numbers of sense and antisense target sites. The target site means the sequence region of the motif in its host intergenic ncRNA.
(TIF)

Figure S6 Expression profiles of "motif and regulatory" intergenic ncRNAs and corresponding neighbour genes across different libraries. The "motif and regulatory" represents intergenic ncRNA with motif(s) and regulatory neighbour gene. The dots linked with coloured line represent the expresion of one intergenic ncRNA and its neighbour gene across different libraries. A represents 5′ end UTR-related RNAs. B represent 5′ end intergenic ncRNAs with UTR-related RNAs removed. C represent 3′ end UTR-related RNAs, and D represent intergenic ncRNAs with UTR-related RNAs removed.
(TIF)

Figure S7 Sequence alignment of bovine "*ZNFX1-AS1*-like" ncRNA and four different human "*ZNFX1-AS1*" transcript variants.
(PDF)

Figure S8 Genomic overview of bovine "*ZNFX1-AS1*-like" intergenic ncRNA. The genomic location of bovine "*ZNFX1-AS1*-like" intergenic ncRNA and corresponding protein-coding gene "*ZNFX1*" is shown in A. The zoomed in view of "*ZNFX1-AS1*-like" ncRNA is shown in B.
(TIF)

Table S1 Library information of all bovine ESTs. This table contains a detailed description of bovine EST libraries downloaded from NCBI.
(XLSX)

Table S2 Summary of the programs used in the pipeline.
(DOCX)

Table S3 Functional over-representation of the neighbour genes of intergenic ncRNAs. This table contains the over-represented GO terms for the 5′ end and 3′ end neighbour genes as well as 10 control gene sets for each end.
(XLSX)

Table S4 Genome coordinates of predicted bovine ncRNAs. This excel table contains two sheets: The first one is the genomic coordinate file with PSL format and based on genome assembly bosTau4; the second one is the annotation for the intergenic ncRNAs.
(XLSX)

Table S5 Summary of annotated known ncRNAs.
(DOCX)

Table S6 Known ncRNAs identified by Rfam and NONCODE2.0. This excel table contains ncRNA annotation based on Rfam and NONCODE2.0.
(XLSB)

Table S7 Summary of identified *cis*-NATs. This excel table contains all the known *cis*-NATs that were identified from bovine ESTs.
(XLSX)

Table S8 Summary of the motifs identified from intergenic ncRNAs. This excel table contains 4 sheets: motifs identified from all 5′ end intergenic ncRNAs with neighbour genes in less than 5 kb distance; motifs identified from all 3′ end intergenic ncRNAs with neighbour genes in less than 5 kb distance; motifs identified from 5′ end intergenic ncRNAs with UTR-related RNAs removed; motifs identified from 3′ end intergenic ncRNAs with UTR-related RNAs removed.
(XLSX)

Table S9 Summary of significantly correlated genes with 5′ end intergenic ncRNAs and 3′ end intergenic ncRNAs. This excel table contains results of genome wide MINE correlation analysis for the 5′ end and 3′ end intergenic ncRNAs.
(XLSX)

Acknowledgments

The authors wish to thank Dan Kortschak, Udaya DeSilva and Jerry Taylor for valuable discussions and critical reading of drafts.

Author Contributions

Conceived and designed the experiments: DLA. Performed the experiments: ZQ. Analyzed the data: ZQ DLA. Wrote the paper: ZQ DLA.

References

1. Carninci P (2006) Tagging mammalian transcription complexity. Trends Genet 22: 501–510.
2. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, et al. (2006) The complexity of the mammalian transcriptome. J Physiol 575: 321–332.
3. Frith MC, Pheasant M, Mattick JS (2005) The amazing complexity of the human transcriptome. Eur J Hum Genet 13: 894–897.
4. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
5. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
6. van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "Dark Matter" Transcripts Are Associated With Known Genes. Plos Biology 8: –.
7. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.

8. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, et al. (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142: 409–419.
9. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46–58.
10. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, et al. (2009) Tiny RNAs associated with transcription start sites in animals. Nat Genet 41: 572–578.
11. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.
12. Kapranov P, Ozsolak F, Kim SW, Foissac S, Lipson D, et al. (2010) New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. Nature 466: 642–646.
13. Adelson DL, Raison JM, Edgar RC (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci U S A 106: 12855–12860.
14. Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST–database for "expressed sequence tags". Nat Genet 4: 332–333.
15. Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, et al. (2007) Improved repeat identification and masking in Dipterans. Gene 389: 1–9.
16. Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32: D109–111.
17. Griffiths-Jones S (2005) Annotating non-coding RNAs with Rfam. Curr Protoc Bioinformatics Chapter 12: Unit 12 15.
18. He S, Liu C, Skogerbo G, Zhao H, Wang J, et al. (2008) NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 36: D170–172.
19. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
20. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
21. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
22. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
23. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6: e1001025.
24. Harhay GP, Smith TP, Alexander LJ, Haudenschild CD, Keele JW, et al. (2010) An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. Genome Biol 11: R102.
25. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28: 337–350.
26. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8: R24.
27. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science 334: 1518–1524.
28. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.
29. Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, et al. (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature 457: 1028–1032.
30. Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, et al. (2011) Expression of distinct RNAs from 3′ untranslated regions. Nucleic Acids Res 39: 2393–2403.
31. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, et al. (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 38: D75–80.
32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
33. Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, et al. (2011) SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. RNA 17: 878–891.
34. Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2: 8.
35. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102: 2454–2459.
36. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35: W345–349.
37. Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17: 117–125.
38. Mourier T, Willerslev E (2009) Retrotransposons and non-protein coding RNAs. Brief Funct Genomic Proteomic.
39. Xue C, Li F (2008) Finding noncoding RNA transcripts from low abundance expressed sequence tags. Cell Res 18: 695–700.
40. Seemann SE, Gilchrist MJ, Hofacker IL, Stadler PF, Gorodkin J (2007) Detection of RNA structures in porcine EST data and related mammals. BMC Genomics 8: 316.
41. Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L, et al. (2011) Critical association of ncRNA with introns. Nucleic Acids Res.
42. Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, et al. (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol 11: R72.
43. Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, et al. (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 8: R43.
44. Amaral PP, Mattick JS (2008) Noncoding RNA in development. Mamm Genome 19: 454–492.
45. Khachane AN, Harrison PM (2010) Mining mammalian transcript data for functional long non-coding RNAs. PLoS One 5: e10316.
46. Hancock JM (2005) Gene factories, microfunctionalization and the evolution of gene families. Trends Genet 21: 591–595.
47. Li JT, Zhang Y, Kong L, Liu QR, Wei L (2008) Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. Nucleic Acids Res 36: 4833–4844.
48. Reiner R, Ben-Asouli Y, Krilovetzky I, Jarrous N (2006) A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription. Genes Dev 20: 1621–1635.
49. Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, et al. (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456: 130–134.
50. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
51. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22: 1–5.
52. Han J, Kim D, Morris KV (2007) Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. Proc Natl Acad Sci U S A 104: 12422–12427.
53. Schmitz KM, Mayer C, Postepska A, Grummt I (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev 24: 2264–2269.
54. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434: 338–345.
55. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915–1927.
56. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503–510.
57. Keene JD (2007) RNA regulons: coordination of post-transcriptional events. Nat Rev Genet 8: 533–543.

Chapter 3

Identification And Comparative Analysis Of ncRNAs In Human, Mouse And Zebrafish Indicate A Conserved Role In Regulation Of Genes Expressed In Brain

Zhipeng Qu and David L. Adelson

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

PLoS ONE **7**:12: e52275. doi:10.1371/journal.pone.0052275

STATEMENT OF AUTHORSHIP

Identification And Comparative Analysis Of ncRNAs In Human, Mouse And Zebrafish Indicate A Conserved Role In Regulation Of Genes Expressed In Brain

PLoS ONE **7**:12: e52275. doi:10.1371/journal.pone.0052275

Zhipeng Qu (Candidate)

Designed and performed experiments, analysed results and wrote the manuscript.

I hereby certify that the statement of contribution is accurate

Signed..... *Date*.....

David L. Adelson

Supervised development of work and assisted in analysing results and writing the manuscript.

I hereby certify that the statement of contribution is accurate and I give permission for inclusion of the paper in the thesis

Signed..... *Date*.....

OPEN ACCESS Freely available online

PLOS ONE

Identification and Comparative Analysis of ncRNAs in Human, Mouse and Zebrafish Indicate a Conserved Role in Regulation of Genes Expressed in Brain

Zhipeng Qu, David L. Adelson*

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia

Abstract

ncRNAs (non-coding RNAs), in particular long ncRNAs, represent a significant proportion of the vertebrate transcriptome and probably regulate many biological processes. We used publically available ESTs (Expressed Sequence Tags) from human, mouse and zebrafish and a previously published analysis pipeline to annotate and analyze the vertebrate non-protein-coding transcriptome. Comparative analysis confirmed some previously described features of intergenic ncRNAs, such as a positionally biased distribution with respect to regulatory or development related protein-coding genes, and weak but clear sequence conservation across species. Significantly, comparative analysis of developmental and regulatory genes proximate to long ncRNAs indicated that the only conserved relationship of these genes to neighbor long ncRNAs was with respect to genes expressed in human brain, suggesting a conserved, ncRNA cis-regulatory network in vertebrate nervous system development. Most of the relationships between long ncRNAs and proximate coding genes were not conserved, providing evidence for the rapid evolution of species-specific gene associated long ncRNAs. We have reconstructed and annotated over 130,000 long ncRNAs in these three species, providing a significantly expanded number of candidates for functional testing by the research community.

Citation: Qu Z, Adelson DL (2012) Identification and Comparative Analysis of ncRNAs in Human, Mouse and Zebrafish Indicate a Conserved Role in Regulation of Genes Expressed in Brain. PLoS ONE 7(12): e52275. doi:10.1371/journal.pone.0052275

Editor: Leonardo Mariño-Ramírez, National Institutes of Health, United States of America

Received October 16, 2012; **Accepted** November 12, 2012; **Published** December 20, 2012

Funding: Funding from the University of Adelaide and a PhD scholarship to ZQ from the China Scholarship Council supported this research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: david.adelson@adelaide.edu.au

Introduction

Protein-coding genes account for only a small proportion of vertebrate genome complexity, specifically, only ~2% of the human genome [1]. With better and more sensitive methods for studying gene expression, such as genome tiling arrays and deep RNA sequencing, we now know that vertebrate "RNA-only" transcriptomes are much more complex than their protein-coding transcriptomes [2,3,4,5]. Studies of some vertebrate genomes have indicated that there are tens of thousands of ncRNAs (non-coding RNAs) [6,7,8], including structural RNAs, such as ribosomal RNAs, transfer RNAs and small non-coding regulatory transcripts such as siRNAs (small interfering RNAs), miRNAs (micro RNAs) and piRNAs (piwi-interacting RNAs) [9]. In addition to these well-characterized ncRNAs, there are a substantial number long ncRNAs, only a few of which have been functionally characterized [10,11,12,13,14].

The few functionally characterized long ncRNAs have various regulatory roles ranging from gene imprinting [15,16], to transcriptional activation/repression of protein-coding genes [17,18]. Specific long ncRNAs have been found with roles in neural development [19] and cell pluripotency [20,21]. Long ncRNAs have also been implicated in pathological processes resulting from aberrant gene regulation [13,22,23]. But not all long ncRNAs are the same and a number of different methods have been used to discover and annotate them. Guttman *et al.* identified thousands of lincRNAs (large intervening/intergenic non-coding RNAs) in mouse using chromatin signatures [10], and Khalil *et al.* extended the catalog of human chromatin-signature-derived lincRNAs to ~3,300 using the chromatin-state maps of 6 human cell types [11]. Many more lincRNAs have been reconstructed from RNA-seq data from multiple sources in human, mouse and zebrafish [12,14,24] and over a thousand long ncRNAs, some of which showed enhancer-like activity, were characterized based on GENCODE annotation [25].

Extrapolation from the limited set of experimentally validated long ncRNAs supports the idea that long ncRNAs are a "hidden" layer of gene regulation. Two lines of evidence supporting this view are their (modest) level of evolutionary sequence conservation and spatial association with regulatory genes. In this report we present the first systematic and methodologically comparable evolutionary analysis of ncRNAs.

In order to determine the full extent of evolutionary conservation of ncRNAs, we used a pipeline built for identifying bovine ncRNAs, particularly long ncRNAs, at genome scale from public EST (Expression Sequence Tag) data. By using ESTs, we were able to get comprehensive datasets of long ncRNAs from both sexes, in many different tissues, cell types, developmental stages, and experimental treatments. In this report we have used this pipeline to analyse all publically available human, mouse and zebrafish ESTs and we present the first global and systematic comparative analysis of non-protein-coding transcriptomes across different species.

We have found large numbers of novel long ncRNAs, many of which originate from the flanking regions of protein-coding genes. Furthermore, we have also shown that gene flanking, intergenic RNAs show sequence conservation compared to non-transcribed genomic regions and are preferentially found near regulatory/developmental protein-coding genes in a species-specific fashion.

Results

1 Genome-wide Exploration of ncRNAs from Human, Mouse, and Zebrafish ESTs

We used a previously described pipeline [26] to screen non-protein-coding transcripts from all publically available human, mouse and zebrafish ESTs and identified over 130,000 ncRNAs (Table 1 and Table S1, http://share:sharingisgood@genomes.ersa.edu.au/ncRNA_pub/). The large numbers of predicted long ncRNAs from human, mouse and zebrafish, together with previously identified bovine ncRNAs, confirm and significantly extend previous reports of pervasive transcription from these four organisms [1,27,28].

Our long ncRNAs fell into 3 categories based on their genomic coordinates with respect to protein-coding genes; intergenic ncRNAs, intronic ncRNAs and overlapped ncRNAs, which overlapped by a small number of base pairs with exons of protein-coding genes [26]. In human and mouse, more than 50% of long ncRNAs were intronic (Figure 1 and Table 2), consistent with previous studies based on other methods [8]. In zebrafish, intergenic ncRNAs were far more numerous than intronic transcripts (Figure 1), but because of the much smaller number of zebrafish intergenic ncRNAs compared to human and mouse (Table 2) it is difficult to be sure that this difference in relative abundance of intergenic ncRNAs is real.

Because many intergenic ncRNAs have been validated as functional elements from different species [10,12,14,25,29], we focused our analyses on all predicted intergenic ncRNAs. The distribution of intergenic ncRNAs with respect to protein-coding genes was the first question we addressed. In all three species, intergenic ncRNAs showed a biased distribution with respect to protein-coding genes at both 5′ and 3′ ends (Figure 2). This is consistent with our previous observation in cow [26] and previous observations in human and mouse based on tiling array and RNA-seq analyses [30,31]. Furthermore, we know that many functional transcripts are located in these regions [8,31].

Larger proportions of sense-strand intergenic ncRNAs were transcribed near the 3′ end of protein-coding genes than antisense ncRNAs in all three species (Figure 2), but the positional distributions of intergenic ncRNAs at the 5′ end of protein-coding genes showed a slightly larger proportion of antisense-strand intergenic ncRNAs, compared to sense intergenic ncRNAs in human and mouse. We considered the possibility that gene-proximate 3′ transcripts were un-annotated UTRs (Untranscribed regions) or alternative transcripts, so we classified these ncRNAs into two subcategories: UTR-related RNAs, that shared high sequence similarity with annotated UTRs or located within 1 kb of protein-coding genes, and "true" intergenic ncRNAs. These results are summarized in Table 2. Some the UTR-related ncRNAs were transcribed from the antisense strand of nearby protein-coding genes, and these may correspond to uaRNAs (UTR-associated RNAs), which are independent transcripts with potential functional significance [32].

2 Problems in the Annotation of Long ncRNA Datasets

Different methods have been used to identify several classes of long ncRNAs, especially lincRNAs, in human [10,11,24,25], mouse [12] and zebrafish [14]. We compared the genomic coordinates of our long ncRNAs from all available tissues and developmental stages in human, mouse and zebrafish, with previously annotated long ncRNA datasets in order to determine the degree of overlap in ncRNAs identified by different methods. The number of EST-based ncRNAs that overlapped with three different human ncRNA datasets was very limited (Figure 3). Only 2,585 ncRNAs in our dataset had overlap with transcripts in at least one of the three known ncRNA datasets (Figure 3A). 1,597 of them overlapped with ~16% (2,296 out of 14,353) of RNA-seq-

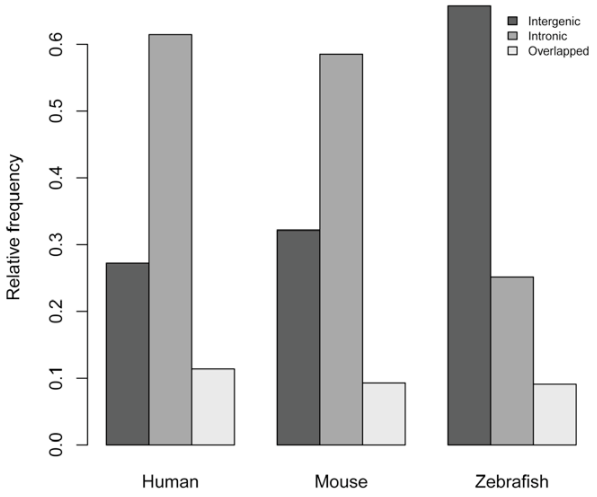

Figure 1. Percentage of intergenic, intronic and overlapped ncRNAs in human, mouse and zebrafish.
doi:10.1371/journal.pone.0052275.g001

Table 1. Summary of procedures for ncRNA identification in human, mouse and zebrafish.

Species	Number of ESTs	Number of assembled transcripts	Mapped to RefSeqs	Mapped to Swiss-Prot	With long ORFs	Putative ncRNAs	Reconstructed ncRNAs
Human*	8,314,483	1,037,755*	44,245*	135,073	130,291	105,994	87,173
Mouse	4,853,460	1,356,763	382,852	3,911	60,342	45,975	36,280
Zebrafish	1,481,936	262,387	117,337	1,828	10,778	11,323	9,877

*Due to the large number of ESTs from human, we ran BLAST for all ESTs against human RefSeqs before assembly and removed all high confident ESTs (coverage >90% and identity >90%). This makes the "Number of assembled transcripts" and "Mapped to RefSeqs" smaller than expected.
doi:10.1371/journal.pone.0052275.t001

Table 2. Classification of ncRNAs.

Species	Number of UTR-related ncRNAs	Number of intergenic ncRNAs	Number of intronic ncRNAs	Number of overlapped ncRNAs
Human	3,438	20,268	55,601	10,724
Mouse	2,179	9,490	21,541	4,414
Zebrafish	2,031	4,464	2,514	1,010

doi:10.1371/journal.pone.0052275.t002

based lincRNAs, and 1,009 overlapped with ~28% (854 out of 3,011) of enhancer-like long ncRNAs. However, only 435 of them overlapped with ~10% (508 out of 4,860) of chromatin-based lincRNAs (Table 3). The intersection of all four of these long ncRNA datasets contained only 25 transcripts, but this is to be expected if previously annotated ncRNAs were present in RefSeq, which we used to screen out known genes transcripts from our EST input data. We confirmed the small number of overlaps between our mouse ncRNAs with four other annotated mouse long ncRNA datasets (Figure 3B and Table 3). In order to confirm that this lack of overlap between our results and previously reported long ncRNAs was attributable to this screening process, we aligned them to the ESTs we used as a starting point for ncRNA identification. Depending on the dataset, we found between 46% and 99% of previously reported human ncRNAs in the EST data (Figure 4 and Table S2). We discuss this further below. Because gene models are continuously being revised, we found that some of our non intergenic ncRNAs overlapped with ncRNAs previously described as intergenic (Table 3).

3 Evolutionary Conservation of ncRNAs in Human, Mouse and Zebrafish

Most protein-coding genes are strongly conserved across different species, as judged by sequence alignment, and this characteristic is exploited to predict genes in newly sequenced organisms. However simple comparison of sequence alignment is insufficient to identify sequence conservation in ncRNAs because they are much less conserved than protein-coding genes. To analyze the evolutionary conservation of predicted ncRNAs, we used a maximum likelihood based method (GERP++ score) [33]. Overall, ncRNAs were conserved, compared to randomly selected un-transcribed genomic fragments, but they were less conserved than protein-coding genes (Figure 5). This result is consistent with previous observations [10,25,26,34]. We also found that many ncRNAs (~50% in human and ~60% in mouse, based on GERP++ score) exhibited positive selection compared to control, randomly selected un-transcribed genomic regions (Figure 5A and 5C). Comparison of specific ncRNA subclasses showed that UTR-related RNAs were more conserved than intergenic ncRNAs, which in turn, were more conserved than intronic ncRNAs (Figure 5B, 5D and 5F). These observations were confirmed using two other methods, phastCons and phyloP (Figure S1 and Figure S2).

To compare the sequence conservation of our predicted ncRNAs with previously annotated long ncRNAs, we calculated the GERP++, phastCons and phyloP scores for human chromatin-based, enhancer-like and RNA-seq-based long ncRNAs (Figure S3, Figure S4 and Figure S5). Our predicted ncRNAs showed similar, but slightly more conserved cumulative conservation curves compared to all three known ncRNA datasets.

4 Intergenic ncRNAs are Preferentially Transcribed Proximate to Regulatory or Developmental Genes

Many ncRNAs, particularly intergenic ncRNAs can regulate gene transcription via different mechanisms [13,20,25,35], including *cis*-regulatory mechanisms. We previously showed that intergenic ncRNAs were more likely to be close to regulatory genes [26]. We used the same methods to analyze the functional classification of human, mouse and zebrafish neighbor genes of gene-proximate intergenic ncRNAs. We chose intergenic ncRNAs located within 5 kb gene-flanking regions as "gene-proximate intergenic ncRNAs", and used GO (Gene Ontology) to functionally classify these neighbor genes in human, mouse and zebrafish [36].

We found that genes with regulatory roles and/or associated with development were enriched in these neighbor genes across all three species with either 5′ end or 3′ end intergenic ncRNAs (Figure 6, Figure 7, Figure S6 and Figure S7). But very few of these neighbor genes were conserved across species, as confirmed by "Gene Symbol" comparison (Figure 8). However, 12 neighbor genes with 5′ proximate ncRNAs in human were found to have sequence-conserved correspondents in mouse and zebrafish neighbor genes, and 96 with 3′ proximate ncRNAs had sequence-conserved correspondents (Identity >60% and coverage >60%) (Table 4, Table S3). Significantly the vast majority of these neighbor genes with conserved proximate ncRNAs are expressed in human brain, suggesting a conserved *cis*-regulatory role for ncRNAs in brain gene expression. To determine if there was a biased functional distribution of protein-coding genes, many of which are 5 kb away from other protein-coding genes, we analyzed human GO annotation for all protein-coding genes with neighbor genes within 5 kb. We found no over-representation of regulatory or developmental genes in this set, indicating that a biased distribution of protein-coding genes did not affect our finding of enriched developmental and regulatory annotation for genes neighboring intergenic ncRNAs (Figure S8).

In order to determine if common GO terms were enriched across species, we compared all the significantly over-represented GO terms (p-value <0.05) across all three species. For genes with 5′ proximate intergenic ncRNAs, we found 19 over-represented terms in common, mostly concerning regulation of different biological pathways (Table 5). Specific molecular function terms enriched in all three species were "transcription factor activity" and "transcription regulator activity" (Table 5). In 3′ end neighbor genes, we found 34 significantly over-represented common GO terms, and the majority of them were "regulation" associated functional enrichments, also including "transcription factor activity" and "transcription regulator activity" (Table 6).

Taken together, these results indicated that many intergenic ncRNAs were transcribed proximate to regulatory or developmental genes in human, mouse and zebrafish. This positional bias and functional classification of neighbor genes indicated a potential

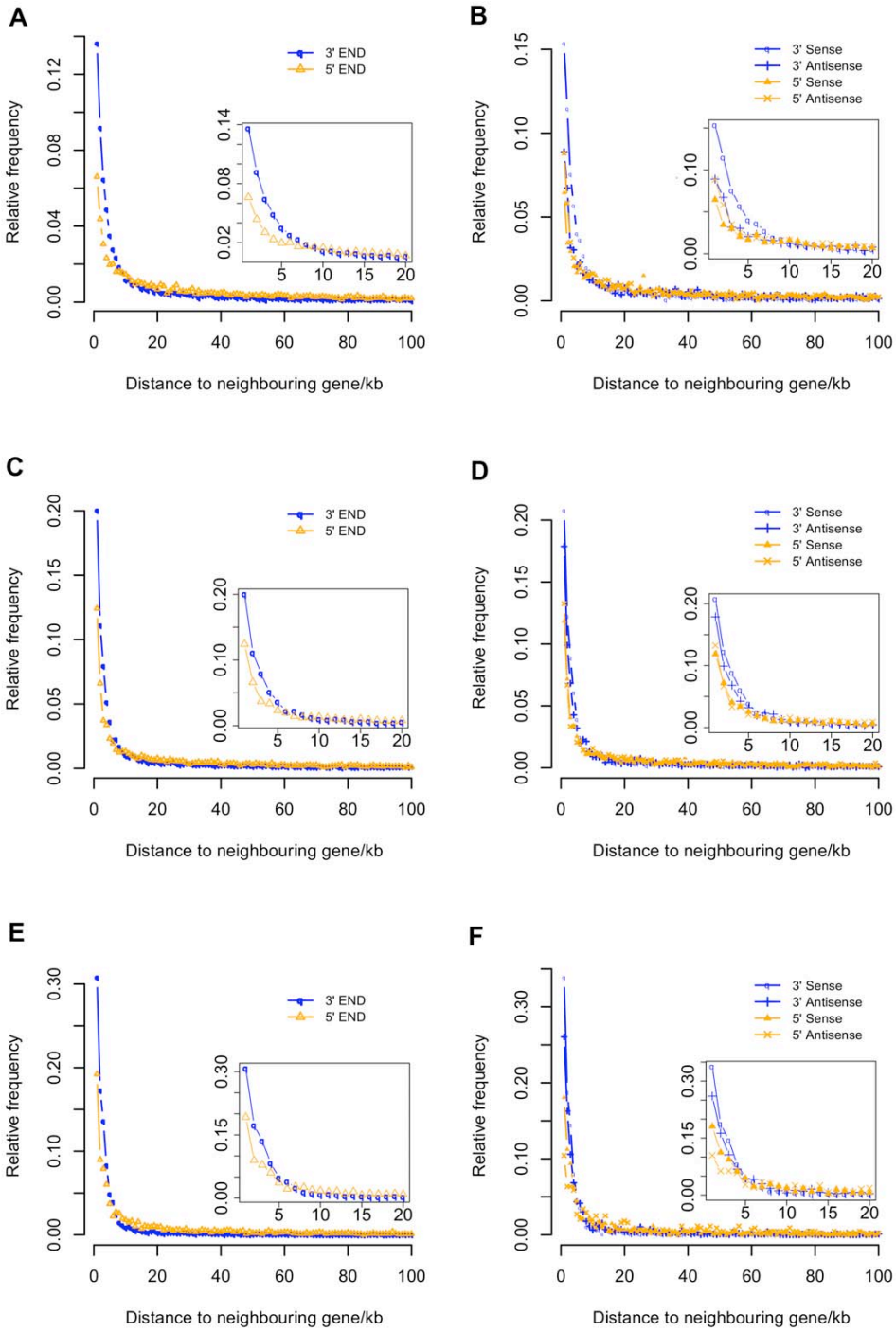

Figure 2. Biased positional distribution of intergenic ncRNAs with respect to neighbor protein-coding genes in human, mouse and zebrafish. The top 2 panels (A & B) are from human, the middle 2 panels (C & D) are from mouse and the bottom 2 panels (E & F) are from zebrafish. A, C and E show the positional distribution of 5′ or 3′ end ncRNAs. B, D and F show the positional distribution of ncRNAs in terms of transcription orientation compared to neighbor genes.
doi:10.1371/journal.pone.0052275.g002

cis-regulatory role for intergenic ncRNAs in the transcription of protein-coding genes.

Discussion

We have assembled and annotated the non-protein-coding transcriptome from human, mouse and zebrafish in a stringent

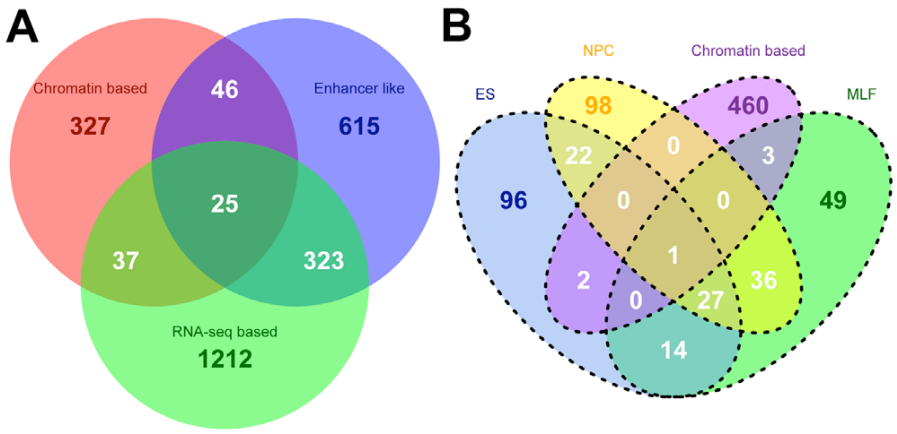

Figure 3. Overlap of our predicted ncRNAs with known human or mouse long ncRNAs from different datasets. A shows the overlap of our ncRNAs with three different human lincRNA datasets. B shows the overlap of our ncRNAs with mouse long ncRNA datasets. "Chromatin based": lincRNAs identified based on chromatin-state maps [10,11]. "Enhancer like": long intergenic ncRNAs identified based on GENCODE [25]. "RNA-seq based": long ncRNAs identified by reconstruction of RNA-seq data in human. "ES", "NPC" and "MLF": long ncRNAs identified by construction of RNA-seq data from 3 different mouse cell types.
doi:10.1371/journal.pone.0052275.g003

and comprehensive fashion using all publically available ESTs. Our results increase the number of annotated ncRNAs by more than an order of magnitude and are robust and highly significant for the following reasons. First, ESTs used to assemble long ncRNAs were generated from multiple libraries from a broad spectrum of tissues/cell types, developmental stages or biological circumstances. Second, robust, highly stringent selection procedures used to assemble long ncRNAs enabled us to remove possible sequencing artifacts. Third, ESTs generated by traditional sanger sequencing technology gave longer raw reads and could be assembled into longer and more accurate consensus transcripts than possible with short read sequencing technologies used in previous studies [12,14,24]. In spite of these positive attributes we also have to acknowledge the potential shortcomings of our reconstructed long ncRNAs. First, many ESTs were archived without transcription orientation, thus it was difficult to deduce transcription orientations for some reconstructed ncRNAs. Second, reconstruction of ESTs from different libraries might have resulted in loss of alternative transcripts. Third, although longer raw reads enabled us to build long consensus transcripts with high accuracy, many reconstructed transcripts are possibly still not full-length. One limitation of our results stemmed from our decision to specifically exclude repetitive ESTs from our analysis because they confounded our sequence reconstructions. This means that repeat containing ncRNAs were not included in our results.

Intergenic ncRNAs from all three species showed the same positional bias in their distribution with respect to protein-coding genes, consistent with previous observations in cow [26]. Because this positional bias was also previously reported in long intergenic ncRNAs identified using quite different methods [27,30,31,37], we propose that this is a common property for intergenic ncRNAs across vertebrate species. This biased genomic distribution could result from two possible scenarios: First, the observed positional bias is a functional attribute for intergenic ncRNAs because they

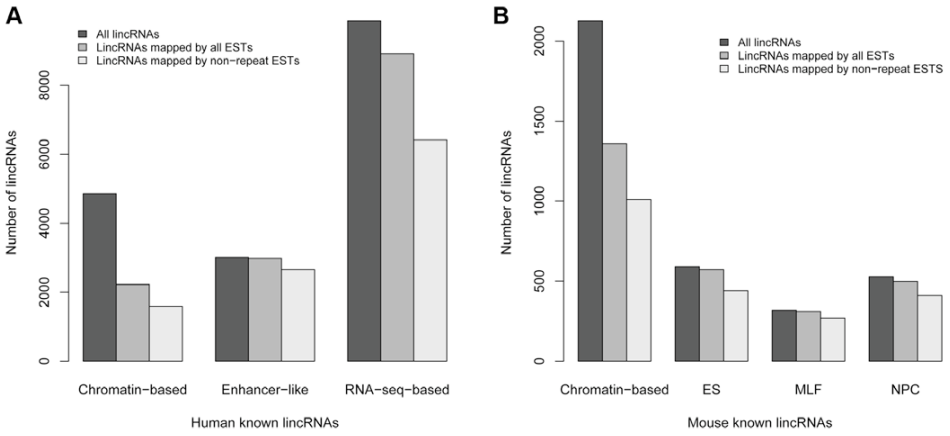

Figure 4. Comparisons of known long ncRNAs mapped by ESTs or non-repeat ESTs in human and mouse. "Chromatin based": lincRNAs identified based on chromatin-state maps [10,11]. "Enhancer like": long intergenic ncRNAs identified based on GENCODE [25]. "RNA-seq based": long ncRNAs identified by reconstruction of RNA-seq data in human. "ES", "NPC" and "MLF": long ncRNAs identified by construction of RNA-seq data from 3 different mouse cell types.
doi:10.1371/journal.pone.0052275.g004

Table 3. Overlap of EST-based ncRNAs with previously identified ncRNAs*.

Dataset	Number of intronic ncRNAs	Number of overlapped ncRNAs	Number of UTR-related RNAs	Number of intergenic ncRNAs (Percentage**)	In total
Chromatin-based lincRNAs (human)	21	8	15	391/1.93%	435
Enhancer-like long ncRNAs (human)	22	10	32	945/4.66%	1,009
RNA-seq-based lincRNAs (human)	11	19	83	1,484/7.32%	1,597
LincRNAs from ES (mouse)	26	13	15	108/1.14%	162
lincRNAs from MLF (mouse)	40	9	11	70/0.74%	130
LincRNAs from NPC (mouse)	30	14	15	125/1.32%	184
Chromatin-based lincRNAs (mouse)	27	87	59	293/3.09%	466
RNA-seq-based long ncRNAs (zebrafish)	16	12	28	105/2.36%	161

*Numbers in this table are shown as our EST-based ncRNAs.
**The percentage is based on the number of all intergenic ncRNAs as shown in table 2.
doi:10.1371/journal.pone.0052275.t003

cis-regulate nearby protein-coding genes through a number of possible mechanisms. Many long intergenic ncRNAs, such as enhancer-like ncRNAs and promoter-associated ncRNAs, have been validated as *cis*-regulators of nearby protein-coding genes [25,38,39]. The transcription of these long intergenic ncRNAs may remodel the chromatin status of surrounding regions, including the promoters of protein-coding loci [18,40,41,42]. Another possibility is that transcription of long ncRNAs from promoter regions of protein-coding genes competes for the transcription-binding complex between long ncRNAs and nearby genes, thus balancing their transcription [17,43,44]. Although many long ncRNAs have been experimentally validated and fed into different gene regulation models, more functional manipulations of long ncRNAs are required to test different regulatory models. The second scenario is that these ncRNAs are fragments of un-annotated UTRs or alternative splicing isoforms. Current ncRNA identification methods are heavily reliant on the available gene models, which may be incomplete. This possibility has some support because some gene-proximate intergenic ncRNAs were similar to UTRs. Because of this possibility, all functional classifications in our analysis were based on stringent intergenic ncRNAs (all UTR-related RNAs removed). However we also observed a large number of antisense transcripts within the gene-proximate intergenic ncRNAs, which cannot be categorized as possible UTRs. Moreover, many studies have identified pervasive, independent functional non-coding transcripts from gene-proximate regions, even in UTRs of protein-coding genes [32]. We conclude that our gene-proximate intergenic ncRNAs are most likely functional, but that we need to wait for further experimental testing to understand how they work [45]. We put forward our ncRNAs as good starting points for functional screening.

Long ncRNAs are pervasively transcribed across genomes in different species [1,46,47]. However, the true number of long ncRNAs is still not known. Previous studies using whole-genome tiling arrays demonstrated that the majority of the human genome was transcribed [2,3,48]. The FANTOM project also revealed thousands of long ncRNAs based on cDNAs in mouse [6]. In the past few years, different categories of long ncRNAs, particularly lincRNAs, have been annotated using a variety of methods [10,11,12,14,24,25]. Our ncRNAs are novel because we screened out ESTs with significant similarity to RefSeqs (coding and non-coding). This novelty is confirmed by the limited overlap of our ncRNAs with previous ncRNAs. In order to assess our methodology vis a vis previous methods, we aligned previously reported ncRNAs against the raw EST data we used as input for our pipeline (See Material S1). Generally ncRNAs from other datasets based on transcriptome data were present in the ESTs, but this was not the case with ncRNAs based on prediction from chromatin state [10,11]. When we assessed the expression of previously reported ncRNAs from chromatin state [10,11] we found that many of these predicted ncRNAs showed no evidence of transcription based on ESTs. These ncRNAs were validated by using tiling array based expression analysis with reported expression levels of 70% within single tissues/cell types [11]. Because we found no more than 46% of these in the raw human EST data (Figure 4, Table S2 and Material S1), we re-visited the tiling arrays reported for the validation. Most of the chromatin state based predicted ncRNAs contained repeats and about 38% of the tiling array probes used to validate them also contained repetitive sequence (Material S1). It is likely that the reported tiling array validation of 70% of the chromatin state predicted ncRNAs is an inflated estimate, as many transcripts contain repeats in their UTRs which would cross-hybridize to these probes, providing false positive signals. On the whole, the number of ncRNAs that were not found in ESTs was a tiny fraction of the total number of ncRNAs included in previous publications and in the present report. We conclude that the number of ncRNAs, particularly for intergenic, repeat containing ncRNAs, is significantly underestimated based on our current knowledge.

Sequence conservation is an important functional signature of genomic transcripts. Many of the ncRNAs that we identified, even though they are clearly less conserved than protein-coding genes, show clear sequence conservation compared to randomly selected, un-transcribed genomic fragments. Furthermore, intergenic ncRNAs are more conserved than intronic ncRNAs in all three species. This weak but significant purifying selection of lincRNAs was observed in a previous study [49] and these results are also consistent with the conservation levels of ncRNAs previously identified from cow [26], as well as previously reported long ncRNA datasets [10,12,14].

Sequence conservation is not the only benchmark for functional significance, as we also observed a small number of

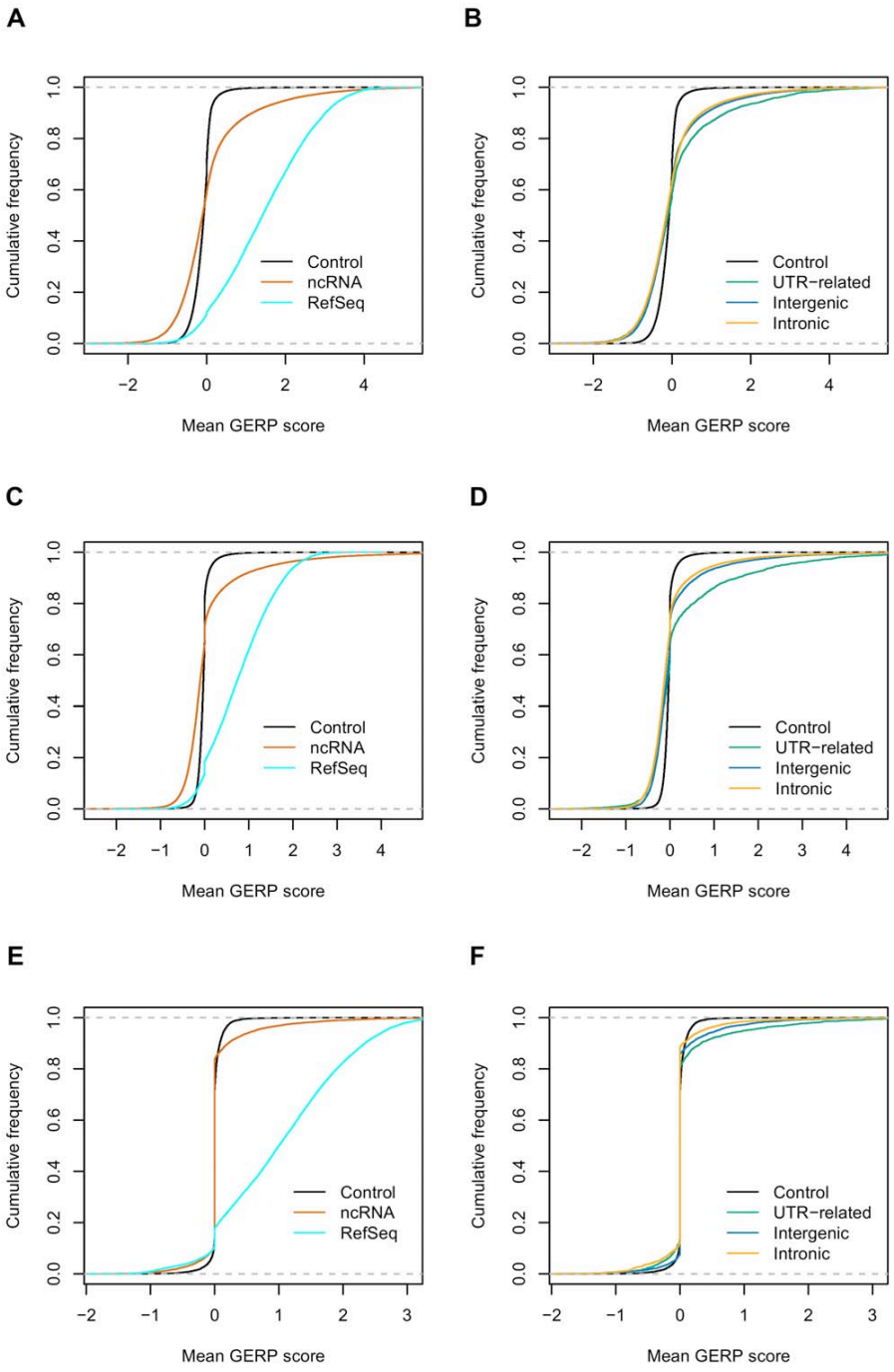

Figure 5. GERP++ score for ncRNAs identified from human, mouse and zebrafish. A and B are from human. C and D are from mouse. E and F are from zebrafish.
doi:10.1371/journal.pone.0052275.g005

protein-coding genes under positive selection. Genes for ncRNAs probably evolve more rapidly than protein-coding genes, which are constrained by triplet codons to maintain the conserved functions of translated proteins. For functional ncRNAs, such as microRNAs, conserved secondary structures have been identified as functional elements required to regulate gene expression. Conserved secondary structures may be more important than conserved primary sequence for long ncRNAs [34]. Furthermore, because many long ncRNAs are transcribed in tissue/cell-type specific fashion [12,14,24,50,51] we suggest that many ncRNAs might be species-specific. The overall lack of correspondence between neighbor genes with proximate intergenic ncRNAs across species supports the idea that ncRNAs evolve rapidly, generating species-specific patterns of

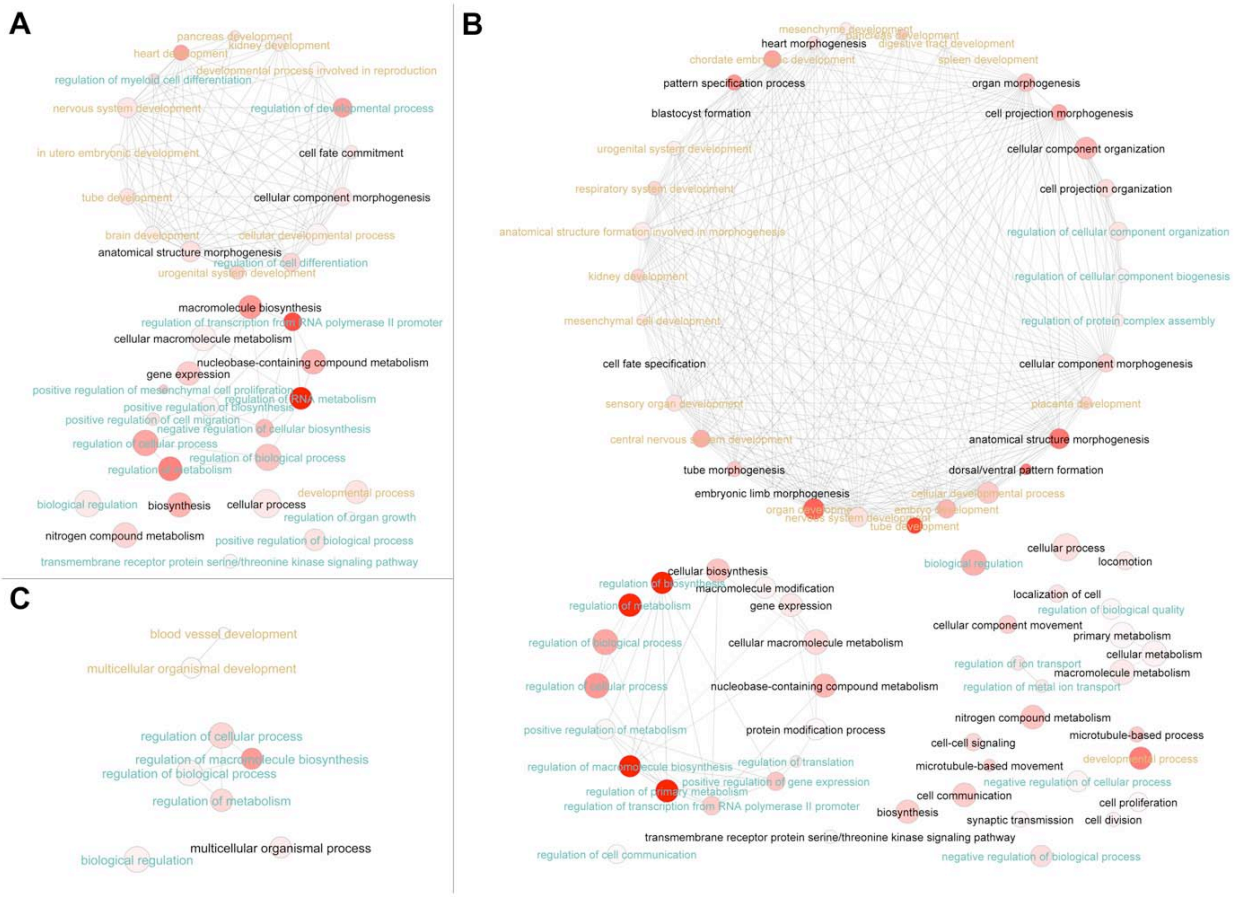

Figure 6. Over-represented GO terms of neighbor genes of 5′ end gene-proximate intergenic ncRNAs in human (A), mouse (B) and zebrafish (C). The bubble color indicates the P-value (EASE score from DAVID); bubble size indicates the frequency of the GO term in the underlying GOA database. Highly similar GO terms are linked by edges in the graph. Regulatory GO terms were highlighted with cyan-like colors, and developmental-associated GO terms were highlighted with gold colors.
doi:10.1371/journal.pone.0052275.g006

tissue specific, developmental regulation. ncRNAs undergoing positive selection might represent novel tissue/cell-type/species specific regulatory transcripts. A significant exception to the lack of correspondence between neighbor genes and proximate intergenic ncRNAs was the conservation of 108 genes with proximate ncRNAs in human, mouse and zebrafish. 97 of these genes are expressed in human brain, suggesting a conserved *cis*-regulatory role for ncRNAs in brain development. Previously, Chodroff *et al.* [52] showed that four conserved long ncRNAs also had conserved expression in brain across a range of amniotes. Our results indicate that conservation of ncRNA association with protein-coding genes expressed in brain also occurs (Table 4, Table S3), suggesting the vertebrates possess a conserved co-expression or *cis*-regulatory network of ncRNA/gene pairs.

As discussed above, the biased positional distribution of intergenic ncRNAs suggested *cis*-regulatory functions. The functional annotation of neighbor genes with nearby intergenic ncRNAs supports this hypothesis. Many intergenic ncRNAs are preferentially transcribed from regions adjacent to regulatory and developmental genes as seen in this report and on a smaller scale by others [10,24,38].

In conclusion, we present a significantly expanded set of ncRNAs that suggests that ncRNAs, while exhibiting sequence conservation, evolve rapidly in terms of their association with neighboring regulatory and developmental genes. The exception to this rapid evolution appears to be with respect to a subset of genes expressed in brain. Long ncRNAs, such as intergenic ncRNAs, may function through different mechanisms as genome wide regulatory elements in many biological pathways, including brain development [53].

Methods

1 ncRNA Identification from Human, Mouse and Zebrafish

ncRNA identification was performed using a previously built pipeline [26]. First, all available ESTs were extracted from dbEST (NCBI). After removing low quality sequences and ESTs composed mostly of repetitive elements, all remaining ESTs were clustered and assembled into longer unique consensus transcripts. Protein-coding genes were removed from the unique transcripts based on similarity searches against RefSeqs and Swiss-Prot databases. As a final step, transcripts were checked for ORFs to remove potential un-annotated protein-coding genes. This left a set

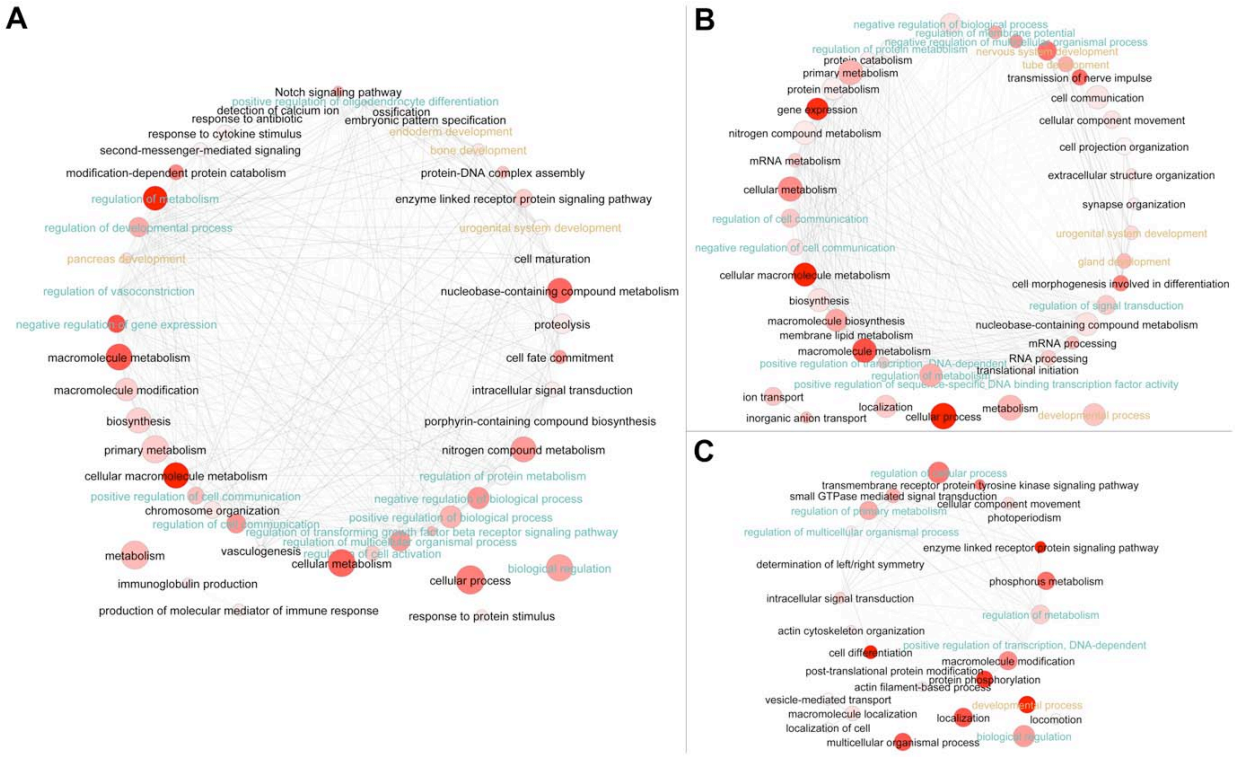

Figure 7. Over-represented GO terms of neighbor genes of 3′ end gene-proximate intergenic ncRNAs in human (A), mouse (B) and zebrafish (C). The bubble color indicates the P-value (EASE score from DAVID); bubble size indicates the frequency of the GO term in the underlying GOA database. Highly similar GO terms are linked by edges in the graph. Regulatory GO terms were highlighted with cyan-like colors, and developmental-associated GO terms were highlighted with gold colors.
doi:10.1371/journal.pone.0052275.g007

of long ncRNAs. To further reduce the redundancy of these long ncRNAs, we reconstructed all putative long ncRNAs based on their genomic coordinates using inchworm [54].

The classification of ncRNAs into three different categories, intronic, intergenic and overlapped ncRNAs with respect to protein-coding genes was performed with R as previously described [26]. The intergenic ncRNAs that were located within 1 kb of the 5′ and 3′ ends of protein-coding genes, or with sequence similarity against known UTRs, were further classified as UTR-related RNAs. All remaining intergenic ncRNAs were classified as *bona fide* intergenic ncRNAs.

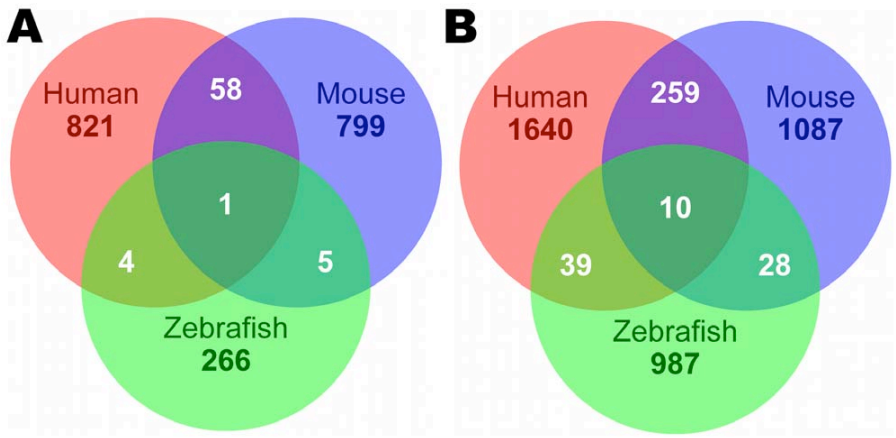

Figure 8. Venn diagrams show the conserved neighbor genes proximate to intergenic ncRNAs from human, mouse and zebrafish. A shows the intersection of neighbor genes with ncRNAs at their 5′ end. B shows the intersection of neighbor genes with ncRNAs at their 3′ end.
doi:10.1371/journal.pone.0052275.g008

Table 4. Human genes conserved in mouse and zebrafish with proximate intergenic ncRNAs at their 5′ end (<5 kb).

Official_gene symbol	Expression in brain (Human)*	Aliases & Descriptions	Diseases disorders*	Related ncRNAs
MAN1A1	Yes	Processing alpha-1,2-mannosidase IA | MAN9 | processing alpha-1,2-mannosidase IA | mannosyl-oligosaccharide 1,2-alpha-mannosidase IA | mannosidase, alpha, class 1A, member 1 | Man(9)-alpha-mannosidase | man(9)-alpha-mannosidase | Mannosidase alpha class 1A member 1 |HUMM3 | alpha-1,2-mannosidase IA | Alpha-1,2-mannosidase IA | Man9-mannosidase | HUMM9 |EC 3.2.1.113	Mannosidase deficiency disease	N/A
MAN1A2	Yes	mannosidase, alpha, class 1A, member 2 |alpha-1, 2-mannosidase IB | Mannosidase alpha class 1A member 2 | mannosyl-oligosaccharide 1,2-alpha-mannosidase IB | alpha1,2-mannosidase | Processing alpha-1,2-mannosidase IB | processing alpha-1,2-mannosidase IB | MAN1B | Alpha-1,2-mannosidase IB |EC 3.2.1.113	N/A	N/A
ONECUT2	Yes	OC2 | hepatocyte nuclear factor 6-beta |ONECUT-2 homeodomain transcription factor | HNF6B | One cut homeobox 2 | HNF-6-beta | Hepatocyte nuclear factor 6-beta | onecut 2 | OC-2 | one cut domain, family member 2 | transcription factor ONECUT-2 | one cut domain family member 2 | Transcription factor ONECUT-2 | one cut homeobox 2	Oral cancer	Target of miR-9
PANK2	Yes	hPanK2 | pantothenate kinase 2 | FLJ11729 | neurodegeneration with brain iron accumulation 1 (Hallervorden-Spatz syndrome) | NBIA1 |Hallervorden-Spatz syndrome | HARP | HSS | Pantothenic acid kinase 2 | C20orf48 | pantothenic acid kinase 2 | PKAN | pantothenate kinase 2, mitochondrial |EC 2.7.1.33	Hallervorden-Spatz syndrome |dementia |dystonia	Host of miR-103
KCNJ4	Yes	IRK-3 | hIRK2 | IRK3 | inward rectifier K(+) channel Kir2.3 | Potassium channel, inwardly rectifying subfamily J member 4 | HRK1 | HIRK2 | potassium channel, inwardly rectifying subfamily J member 4 |hippocampal inward rectifier potassium channel | potassium inwardly-rectifying channel, subfamily J, member 4 | Hippocampal inward rectifier | inward rectifier K+ channel Kir2.3 | HIR | inward rectifier potassium channel 4 | Kir2.3 | Inward rectifier K(+) channel Kir2.3	N/A	N/A
PDCD6IP	Yes	apoptosis-linked gene 2-interacting protein X | dopamine receptor interacting protein 4 | ALIX | programmed cell death 6 interacting protein | ALG-2-interacting protein 1 | programmed cell death 6-interacting protein | PDCD6-interacting protein | Hp95 | KIAA1375 | Alix | HP95 | AIP1 |ALG-2 interacting protein 1 | DRIP4	N/A	Target of miR-1225-5P
SNX14	Yes	sorting nexin 14 | RGS-PX2 |sorting nexin-14	N/A	N/A
TUBB2B	Yes	tubulin beta-2B chain | tubulin, beta polypeptide paralog | MGC8685 | bA506K6.1 | tubulin, beta 2B class IIb | DKFZp566F223 | tubulin, beta 2B | class IIb beta-tubulin |class II beta-tubulin isotype	Lissencephaly	N/A
ZNF41	Yes	TUBB |class IIa beta-tubulin | tubulin, beta 2A class IIa | TUBB2 | tubulin, beta polypeptide 2 | tubulin, beta 2 | TUBB2B | dJ40E16.7 | tubulin beta-2A chain | tubulin, beta polypeptide | tubulin, beta 2A	Aland Island eye disease | mental disorder |intellectual disability	N/A
ZNF595	Yes	MRX89 |MGC8941 | zinc finger protein 41	N/A	N/A
ZNF676	Yes	FLJ31740 | zinc finger protein 595	N/A	N/A
ZNF761	No	zinc finger protein 676	N/A	N/A

*The expression and disease annotation were based on GeneCards V3 [57].

doi:10.1371/journal.pone.0052275.t004

2 Neighbor Genes and Transcription Orientation of ncRNAs with Respect to Neighbor Genes

The closest protein-coding gene to an intergenic ncRNA was chosen as the neighbor gene of this intergenic ncRNA. The transcriptional orientation of ncRNAs was determined based on two criteria: First, many ESTs extracted from NCBI have cloning and sequencing information, which was used to determine the transcription orientation of both singletons and contigs. Second, the transcription orientation of spliced long ncRNAs was deduced from splicing information when they were mapped onto the genome. The "sense" intergenic ncRNAs were defined as transcribing from the same strand as neighbor genes, and *vice versa*.

Table 5. GO terms in common from human, mouse and zebrafish neighbor genes within 5kb of proximate ncRNAs at their 5′ end.

Category	Term	*P value (human)	P value (mouse)	P value (zebrafish)
Molecular Function	GO:0003700~transcription factor activity	6.88E-07	0.001685935	0.002045234
Molecular Function	GO:0030528~transcription regulator activity	2.80E-06	2.50E-05	0.001720193
Biological Process	GO:0006355~regulation of transcription, DNA-dependent	4.53E-06	0.000108619	0.02130028
Biological Process	GO:0051252~regulation of RNA metabolic process	7.91E-06	0.000178503	0.023870388
Biological Process	GO:0010556~regulation of macromolecule biosynthetic process	8.37E-06	4.96E-07	0.000915362
Biological Process	GO:0060255~regulation of macromolecule metabolic process	5.89E-05	7.41E-06	0.00691373
Biological Process	GO:0045449~regulation of transcription	6.20E-05	2.37E-06	0.001790827
Biological Process	GO:0031326~regulation of cellular biosynthetic process	8.41E-05	1.10E-06	0.001054761
Biological Process	GO:0009889~regulation of biosynthetic process	0.000119902	1.33E-06	0.001088173
Biological Process	GO:0080090~regulation of primary metabolic process	0.000146447	6.89E-07	0.002903755
Biological Process	GO:0010468~regulation of gene expression	0.000154686	1.42E-06	0.002943972
Biological Process	GO:0031323~regulation of cellular metabolic process	0.00015819	4.08E-06	0.002422663
Biological Process	GO:0019219~regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.000321532	7.14E-06	0.002751033
Biological Process	GO:0051171~regulation of nitrogen compound metabolic process	0.000343647	6.14E-06	0.002831208
Biological Process	GO:0019222~regulation of metabolic process	0.000349372	1.09E-05	0.011044253
Biological Process	GO:0050794~regulation of cellular process	0.001348476	0.000766239	0.009737321
Biological Process	GO:0050789~regulation of biological process	0.00433817	0.001382295	0.033481278
Biological Process	GO:0065007~biological regulation	0.022428992	0.002031998	0.031603795
Biological Process	GO:0007275~multicellular organismal development	0.035916788	0.000243142	0.043621824

*The GO terms were ordered by p-value in human.

doi:10.1371/journal.pone.0052275.t005

3 Comparisons with Known Well-characterized Long ncRNAs in Human, Mouse and Zebrafish

The sources and summary information for previously characterized ncRNAs are shown in Table 7. For chromatin-based lincRNAs in human and mouse, we used the exons instead of the long chromatin regions as the known lincRNAs. The overlap of our EST-based ncRNAs with these known long ncRNA datasets were analyzed with the "GenomicFeatures" R package.

4 Conservation Analyses of ncRNAs

Three different conservation scores were used to analyze the sequence conservation of ncRNAs. The GERP++ scores for human and mouse were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/. For zebrafish, the GERP++ scores were calculated with GERP++ tool based on the multiple alignments of 7 genomes (hg19/GRCh37, mm9, xenTro2, tetNig2, fr2, gasAcu1, oryLat2) with danRer7 of zebrafish. The phastCons scores and phyloP scores for human, mouse and zebrafish were downloaded from UCSC based on genome assembly hg19/GRCh37 (human), mm9 (mouse) and danRer7 (zebrafish) respectively. The mean GERP++/phastCons/phyloP score for each ncRNA/RefSeq/control sequence was calculated by normalizing the sum of GERP++/phastCons/phyloP scores against the length of the sequence. All RefSeqs excluding "NR" and "XR" entries (non-coding transcripts) were used as the protein-coding gene dataset. The same number of

Table 6. GO terms in common from human, mouse and zebrafish neighbor genes within 5kb of proximate ncRNAs at their 3′ end.

Category	Term	*P value (human)	P value (mouse)	P value (zebrafish)
Molecular Function	GO:0003677~DNA binding	2.52E-07	0.001016369	0.022517442
Biological Process	GO:0019222~regulation of metabolic process	5.94E-06	0.001833053	0.007240134
Biological Process	GO:0031323~regulation of cellular metabolic process	7.06E-06	0.001932015	0.002531781
Biological Process	GO:0080090~regulation of primary metabolic process	8.71E-06	0.000746433	0.001635905
Biological Process	GO:0060255~regulation of macromolecule metabolic process	1.52E-05	0.001021052	0.015088588
Cellular Component	GO:0044464~cell part	2.64E-05	0.005138983	0.021192768
Cellular Component	GO:0005623~cell	2.75E-05	0.005138983	0.021192768
Biological Process	GO:0009889~regulation of biosynthetic process	4.64E-05	0.00153235	0.001998668
Biological Process	GO:0010556~regulation of macromolecule biosynthetic process	5.07E-05	0.001133669	0.004636373
Biological Process	GO:0031326~regulation of cellular biosynthetic process	5.93E-05	0.001770385	0.002769539
Biological Process	GO:0010468~regulation of gene expression	6.05E-05	0.001153647	0.019089475
Biological Process	GO:0019219~regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	7.45E-05	0.002835006	0.006403442
Biological Process	GO:0045449~regulation of transcription	9.02E-05	0.001133423	0.009147674
Biological Process	GO:0051171~regulation of nitrogen compound metabolic process	0.000115522	0.003953563	0.006560818
Molecular Function	GO:0003700~transcription factor activity	0.000701959	0.006403948	0.003113804
Biological Process	GO:0051252~regulation of RNA metabolic process	0.002751656	0.012593576	0.006423226
Biological Process	GO:0006355~regulation of transcription, DNA-dependent	0.002836401	0.008313995	0.007792617
Molecular Function	GO:0030528~transcription regulator activity	0.003105196	0.00782068	0.001014153
Biological Process	GO:0031328~positive regulation of cellular biosynthetic process	0.007428451	0.007226598	0.033533698
Biological Process	GO:0009891~positive regulation of biosynthetic process	0.007469104	0.008740921	0.033533698
Biological Process	GO:0010557~positive regulation of macromolecule biosynthetic process	0.009196945	0.003489005	0.028269774
Biological Process	GO:0010628~positive regulation of gene expression	0.010415711	0.009098997	0.021490484
Biological Process	GO:0045941~positive regulation of transcription	0.011143783	0.00569233	0.021490484
Molecular Function	GO:0005515~protein binding	0.017163574	0.000809527	1.60E-06
Biological Process	GO:0045893~positive regulation of transcription, DNA-dependent	0.02105859	0.004978895	0.012497621
Molecular Function	GO:0008270~zinc ion binding	0.022962024	0.003010259	0.036242576
Biological Process	GO:0048869~cellular developmental process	0.024154786	0.006314016	9.66E-07
Biological Process	GO:0051254~positive regulation of RNA metabolic process	0.024566919	0.005669422	0.014428949
Biological Process	GO:0030154~cell differentiation	0.02953709	0.007655265	1.65E-06
Biological Process	GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.03326329	0.011738803	0.039427105
Biological Process	GO:0048468~cell development	0.033319932	0.007737614	0.003006631
Biological Process	GO:0051173~positive regulation of nitrogen compound metabolic process	0.033319932	0.012196797	0.04261773
Biological Process	GO:0044267~cellular protein metabolic process	0.042639534	0.003735008	0.011732507
Biological Process	GO:0001655~urogenital system development	0.048304941	0.012438853	0.04591464

*The GO terms were ordered by p-value in human.

doi:10.1371/journal.pone.0052275.t006

genomic fragments as ncRNAs, which ranged in size from 500 bp to 15,000 bp, were randomly selected from un-transcribed genomic regions (no ESTs mapped) as the control datasets for each species respectively. The cumulative frequency for each dataset was calculated and plotted using the R package.

5 Functional Classifications of Neighbor Genes of Gene-proximate Intergenic ncRNAs

Gene-proximate intergenic ncRNAs were selected from stringent intergenic ncRNAs located within 5 kb of the 5′ and 3′ ends of protein-coding genes. GO classification of neighbor genes was performed on the DAVID (Database for Annotation, Visualization and Integrated Discovery) web server [55]. The thresholds for over-represented GO terms were set as gene count >5 and p-value

Table 7. Previously annotated long ncRNA datasets used for comparison.

Dataset	Number of ncRNAs	Source	Method	Reference
Chromatin-based lincRNAs (Human)	4,860*	10 cell types	Chromatin signature identification (K4–K36 domain)	Khalil AM, 2009 [11]
Enhancer-like long ncRNAs (Human)	3,011	Multiple	Screening from GENCODE annotation	Orom UA, 2010 [25]
RNA-seq-based lincRNAs (Human)	8,195	24 tissues and cell types	Screening from assembled RNA-seq data	Cabili MN, 2011 [24]
Chromatin-based lincRNAs (Mouse)	2,127*	4 cell types	Chromatin signature identification (K4–K36 domain)	Guttman M, 2009 [10]
RNA-seq-based lincRNAs (Mouse)	1,140	3 cell types	Screening from assembled RNA-seq data	Guttman M, 2010 [12]
RNA-seq-based long ncRNAs (Zebrafish)	1,133	8 embryonic stages	Screening from assembled RNA-seq data	Pauli A, 2011 [14]

*These are the exons identified by microarray from non-coding k4-k36 domains.
doi:10.1371/journal.pone.0052275.t007

(EASE score) <0.05. The web server REViGO was used to reduce the redundancy and visualize the overrepresented GO terms based on semantic similarity [56].

The gene symbols of neighbor genes with annotations in GO were compared across species to find common genes. BLAST was used to carry out sequence similarity searches for conserved neighbor genes across all three species.

All protein-coding genes with neighbor genes located in their 5 kb flanking regions were analysed in the same fashion as neighbor genes of intergenic ncRNAs.

Supporting Information

Figure S1 PhastCons scores of ncRNAs identified from human (A, B), mouse (C, D) and zebrafish (E, F).
(TIF)

Figure S2 Phylop Scores of identified ncRNAs from human (A, B), mouse (C, D) and zebrafish (E, F).
(TIF)

Figure S3 Comparison of GERP++ scores of our ncRNAs with previously published lincRNA datsets in human.
(TIF)

Figure S4 Comparison of phastCons scores of our ncRNAs with previously published human lincRNA datasets.
(TIF)

Figure S5 Comparison of phyloP scores of our ncRNAs with previously published human lincRNA datasets.
(TIF)

Figure S6 The ''Treemap'' view of over-represented GO terms of neighbor genes with 5′ end gene-proximate intergenic ncRNAs in human (A), mouse (B) and zebrafish (C). Each rectangle represents a single cluster. The clusters are joined into 'superclusters' of loosely related terms, visualized with different colors. The size of the rectangles was adjusted to reflect the P-value (EASE score in DAVID) of the GO term, with a larger rectangle corresponding to a smaller p-value.
(TIF)

Figure S7 The ''Treemap'' view of over-represented GO terms of neighbor genes with 3′ end gene-proximate intergenic ncRNAs in human (A), mouse (B) and zebrafish (C). Each rectangle represents a single cluster. The clusters are joined into 'superclusters' of loosely related terms, visualized with different colors. The size of the rectangles was adjusted to reflect the P-value (EASE score in DAVID) of the GO term, with a larger rectangle corresponding to a smaller p-value.
(TIF)

Figure S8 Over-represented GO terms for all protein-coding genes with neighbor genes within 5 kb in human.
(TIF)

Table S1 Genomic coordinates of predicted ncRNAs in human, mouse and zebrafish. This excel file contains genomic coordinates of predicted ncRNAs identified by our pipeline in human (sheet 1), mouse (sheet 2) and zebrafish (sheet 3).
(XLSX)

Table S2 Summary of human and mouse known long ncRNAs that align to ESTs. This table contains a summary of human known long ncRNAs (chromatin-based, enhancer-like and RNA-seq based) and mouse long ncRNAs (chromatin-based, RNA-seq based) mapped against ESTs.
(DOCX)

Table S3 Annotation of common protein-coding genes with proximate intergenic ncRNAs (<5 kb) in human, mouse and zebrafish. Sheet 1 in this excel table shows 12 conserved genes with ncRNAs at the 5′ end and sheet 2 shows 96 conserved genes with ncRNAs at the 3′ end.
(XLSX)

Material S1 Supporting results.
(DOCX)

Acknowledgments

The authors thank Dan Kortschak, Sim Lim, Ali Walsh and Reuben Buckley for valuable discussions.

Author Contributions

Conceived and designed the experiments: ZQ DLA. Performed the experiments: ZQ. Analyzed the data: ZQ DLA. Contributed reagents/materials/analysis tools: ZQ. Wrote the paper: ZQ DLA.

References

1. Frith MC, Pheasant M, Mattick JS (2005) The amazing complexity of the human transcriptome. Eur J Hum Genet 13: 894–897.
2. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.
3. Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21: 93–102.
4. Carninci P (2006) Tagging mammalian transcription complexity. Trends Genet 22: 501–510.
5. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, et al. (2006) The complexity of the mammalian transcriptome. J Physiol 575: 321–332.
6. Numata K, Kanai A, Saito R, Kondo S, Adachi J, et al. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res 13: 1301–1306.
7. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, et al. (2007) Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 17: 852–864.
8. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
9. Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15 Spec No 1: R17–29.
10. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.
11. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106: 11667–11672.
12. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503–510.
13. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, et al. (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142: 409–419.
14. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, et al. (2011) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22: 577–591.
15. Braidotti G, Baubec T, Pauler F, Seidl C, Smrzka O, et al. (2004) The Air noncoding RNA: an imprinted cis-silencing transcript. Cold Spring Harb Symp Quant Biol 69: 55–66.
16. Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM (2006) Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev 20: 1268–1282.
17. Martens JA, Laprade L, Winston F (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature 429: 571–574.
18. Uhler JP, Hertel C, Svejstrup JQ (2007) A role for noncoding transcription in activation of the yeast PHO5 gene. Proc Natl Acad Sci U S A 104: 8011–8016.
19. Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, et al. (2010) Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. BMC Neurosci 11: 14.
20. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, et al. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477: 295–300.
21. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, et al. (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18: 1433–1445.
22. Fu X, Ravindranath L, Tran N, Petrovics G, Srivastava S (2006) Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. DNA Cell Biol 25: 135–141.
23. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, et al. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14: 723–730.
24. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915–1927.
25. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46–58.
26. Qu Z, Adelson DL (2012) Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes. PLoS One 7: e42638.
27. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
28. Shabalina SA, Spiridonov NA (2004) The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol 5: 105.
29. Marques AC, Ponting CP (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol 10: R124.
30. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
31. Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, et al. (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature 457: 1028–1032.
32. Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, et al. (2010) Expression of distinct RNAs from 3′ untranslated regions. Nucleic Acids Res 39: 2393–2403.
33. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6: e1001025.
34. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22: 1–5.
35. Lee JT (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev 23: 1831–1842.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
37. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, et al. (2009) Tiny RNAs associated with transcription start sites in animals. Nat Genet 41: 572–578.
38. Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet 5: e1000617.
39. Petruk S, Sedkov Y, Riley KM, Hodgson J, Schweisguth F, et al. (2006) Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. Cell 127: 1209–1221.
40. Schmitt S, Prestel M, Paro R (2005) Intergenic transcription through a polycomb group response element counteracts silencing. Genes Dev 19: 697–708.
41. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, et al. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32: 232–246.
42. Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, et al. (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456: 130–134.
43. Kurokawa R (2011) Promoter-associated long noncoding RNAs repress transcription through a RNA binding protein TLS. Adv Exp Med Biol 722: 196–208.
44. Song X, Wang X, Arai S, Kurokawa R (2011) Promoter-associated noncoding RNA from the CCND1 promoter. Methods Mol Biol 809: 609–622.
45. Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? Hum Mol Genet 19: R162–168.
46. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, et al. (2011) The reality of pervasive transcription. PLoS Biol 9: e1000625; discussion e1001102.
47. Dinger ME, Amaral PP, Mercer TR, Mattick JS (2009) Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic 8: 407–423.
48. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, et al. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15: 987–997.
49. Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV (2011) Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol Evol 3: 1390–1404.
50. Sasaki YT, Sano M, Ideue T, Kin T, Asai K, et al. (2007) Identification and characterization of human non-coding RNAs with tissue-specific expression. Biochem Biophys Res Commun 357: 991–996.
51. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A 105: 716–721.
52. Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, et al. (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biology 11: R72.
53. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.
54. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.
55. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
56. Supek F, Bosnjak M, Skunca N, Smuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6: e21800.
57. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, et al. (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010: baq020.

Chapter 4 Re-construction and Annotation of Human Protein-coding and Non-coding RNA Co-expression Networks

Introduction

Decoding the complexity of vertebrate genomes and transcriptomes has revealed pervasively transcribed ncRNAs, particularly long ncRNAs, in addition to well-annotated protein-coding genes (Frith et al., 2005;Carninci, 2006). Although the real number of transcribed functional ncRNAs is not yet accurately known, many thousands of ncRNAs, mainly long intergenic ncRNAs, have been predicted and characterized from human, mouse, and other organisms (Guttman et al., 2009;Khalil et al., 2009;Amaral et al., 2010;Guttman et al., 2010;Orom et al., 2010;Pauli et al., 2011). These studies strongly support the view that except for protein-coding genes, RNAs, especially long ncRNAs, are also involved in the regulatory networks of different biological pathways (Huarte et al., 2010;Orom et al., 2010;Guttman et al., 2011;Prensner et al., 2011). An increasing number of experimental results have also confirmed that long ncRNAs regulate broad biological functions. These regulatory roles include the activation or repression of genes by chromatin remodeling, transcriptional regulation and post-transcriptional regulation (Mercer et al., 2009;Qu and Adelson, 2012b). Long ncRNAs can regulate gene transcription in both *cis* or *trans* fashion (Mercer et al., 2009;Orom et al., 2010;Guttman et al., 2011).

Previous studies have confirmed that intergenic ncRNAs are primarily transcribed from regions proximate to protein-coding genes (Guttman et al., 2010;Cabili et al., 2011;Qu and Adelson, 2012a). One explanation for this observation is that ncRNAs are just transcriptional by-products of protein-coding genes (van Bakel et al., 2010). However, the evidence shows that this kind of transcriptional co-localization of protein-coding and non-coding RNAs is also the result of functional association (Ponjavic et al., 2009). The co-expression of long intergenic ncRNAs and protein-coding genes, in particular for nearby gene partners, has been observed in brain (Mercer et al., 2008;Ponjavic et al., 2009). This co-expression is also regarded as support for long ncRNAs function as *cis*-regulators of target genes (Orom et al., 2010). In fission yeast, the transcription of ncRNAs was required for the transcriptional activation of genes in nearby upstream regions by remodeling chromatin status (Hirota et al., 2008).

In addition to transcriptional co-localization of long ncRNAs with protein-coding genes, long ncRNAs were also identified as showing region, tissue or cell-type specific expression patterns. In mouse brain, the specific expression profiles of long ncRNAs were found at both regional and subcellular levels (Mercer et al., 2008). Brain-specific expressed long ncRNAs were also confirmed in mouse developing brain and found adjacent to protein-coding genes involved in nervous system developmental associated transcriptional regulation (Ponjavic et al., 2009). A model was proposed by Guttman *et al*. postulating that lincRNAs are transcribed in cell-type-specific fashion and interact with cell-type-specific RNA-binding proteins to affect global gene expression, to maintain the pluripotent state of ES cell (Guttman et al., 2011). In human, the expression of many long ncRNAs was detected within restricted subsets or specific tissues by RT-PCR and northern blot hybridisation analyses (Sasaki et al., 2007). In the human cancer transcriptome, long ncRNAs showed extensive tissue-specific expression in cancer samples. Some of them exhibited highly aberrant expression in cancer tissues compared to normal ones (Gibb et al., 2011).

To interpret the global expression profiles of long ncRNAs and protein-coding genes and their potential connections in human, we re-constructed co-expression networks for protein-coding and non-coding RNAs by leveraging the transcriptome data from 16 different human tissues. More than 30,000 transcripts, including 21,725 reference genes and 10,708 previously predicted intergenic ncRNAs, were clustered into 43 co-expression modules, most of which represented tissue-specific co-expressed protein-coding and non-coding RNAs. The functional classification of protein-coding genes in each module showed that tissue-specific regulatory or developmental genes were over-represented, indicating that co-expressed intergenic ncRNAs may be also involved in, and have significant contribution to, tissue specific developmental regulatory networks.

Results

1 Re-construction of co-expression networks combining ncRNAs and protein-coding genes

We previously identified tens thousands of human intergenic ncRNAs based on all publicly available ESTs (Chapter 3). We used RNA-seq data from different human tissues to generate the expression profiles for predicted intergenic ncRNAs and all

annotated protein-coding genes. We observed that more reference genes were widely expressed in multiple tissues (Figure 1). We did also see that quite a number of intergenic ncRNAs were detected in multiple tissues. Globally, more intergenic ncRNAs showed greater tissue-specific expression based on normalized FPKM (Fragments Per Kilobases per Million reads) values (Figure 1).

Table 1. Summary of 43 eigengene modules and tissue-specific branches

Module ID	Number of transcripts in module	Number of ncRNAs in module	Number of genes in module	Intersection between module genes with neighbour genes of ncRNAs	Tissue specific expression in this module	Tissue specific branches
module1	648	136	512	30	heart+skeletal muscle	A
module2	1289	312	977	61	skeletal muscle	A
module3	206	18	188	2	skeletal muscle+testes	A
module4	307	80	227	8	kidney+liver	B
module5	1506	288	1218	73	liver	B
module6	49	33	16	0	breast+heart	C
module7	53	45	8	0	adrenal+heart	C
module8	123	57	66	2	brain+heart	C
module9	898	339	559	51	heart	C
module10	502	149	353	20	adipose	D
module11	870	455	415	39	breast	E
module12	323	90	233	9	adipose+breast	E
module13	832	413	419	11	adrenal+lymph node	F
module14	826	539	287	13	lymph node	F
module15	107	48	59	1	adrenal+lung	G
module16	127	92	35	0	adrenal+kidney	G
module17	299	144	155	2	adrenal+thyroid	G
module18	1026	738	288	12	adrenal	G
module19	1114	326	788	42	lung	H
module20	83	21	62	1	lung+testes	H
module21	173	64	109	5	lung+ovary	H
module22	280	23	257	1	prostate+white blood cells	I
module23	165	17	148	2	lung+white blood cells	I
module24	2683	508	2175	146	white blood cells	I
module25	799	212	587	23	adrenal+lymph node+white blood cells	I
module26	406	111	295	10	colon+prostate	J
module27	516	257	259	8	colon	J
module28	5283	855	4428	211	testes	K

module29	105	36	69	0	brain+white blood cells	L
module30	3735	1458	2277	246	brain	L
module31	307	93	214	3	brain+testes	L
module32	701	283	418	11	prostate	M
module33	172	85	87	4	brain+prostate	M
module34	108	44	64	2	prostate+thyroid	M
module35	322	105	217	7	prostate+ovary	M
module36	1788	730	1058	81	ovary	N
module37	271	139	132	6	brain+ovary	N
module38	168	106	62	1	brain+kidney	O
Module39	1402	593	809	50	kidney	O
Module40	121	32	89	0	kidney+testes	O
module41	286	81	205	7	kidney+thyroid	P
module42	1200	445	755	43	thyroid	P
module43	254	108	146	5	brain+thyroid	P

Based on the pairwise correlations of expression profiles of 10,708 intergenic ncRNAs and 21,725 reference genes across 16 different tissues in human, these transcripts were clustered into 43 eigengene modules (or sub-networks, an eigengene can be defined as the most representative gene expression profile of the module) (Langfelder and Horvath, 2008), representing transcripts with similar expression patterns across different tissues (Figure 2 and Table 1). The co-expression patterns of these 43 modules showed that most of them represent clusters of transcripts with tissue-specific expression (Figure 3 and Figure 4). For example, in module30, most of the involved transcripts, including 1,458 intergenic ncRNAs and 2,277 genes, showed relatively high expression in brain, compared to the other 15 tissues (Figure 3). In module 28, we observed 4,428 genes as well as another 855 intergenic ncRNAs showing similar high expression in testes (Figure 4). We summarized these tissue-specific expression patterns for all modules. 16 of them were corresponding clusters of transcripts showing single tissue-specific high expression in 16 different tissues respectively (Table 1). Transcripts in 26 modules showed high co-expression in two tissues, and only 1 module (module25) showed high co-expression in three tissues, which are adrenal, lymph node and white blood cells.

We observed that both reference genes as well as intergenic ncRNAs are involved in all of these 43 co-expression modules, but the numbers of transcripts involved in each module are diverse (Table 1 and Figure 5). The mean number of transcripts in all 43 modules was approximately 754. Module28 (testes) had the largest number of

transcripts, including 855 intergenic ncRNAs and 4,428 genes. We also observed that module30 (brain) had the largest number of intergenic ncRNAs (1,458) compared to all other modules. Transcripts in 5 of the 26 two-tissues-specific modules also showed high expression in brain. The correlation of numbers of intergenic ncRNAs and genes in all 43 sub-networks was ~0.81 (Spearman's rank correlation, P-value = 7.045e-11). 35 modules had more genes than intergenic ncRNAs (Table 1). This was not surprising because we had more genes (twice as many compared to ncRNAs) to build the co-expression networks. However, we also found 8 co-expression modules that had more intergenic ncRNAs than genes (Table 1).

We identified the neighbour genes of intergenic ncRNAs for all of the above eigengene modules, and found that only limited numbers of these were also shown in the corresponding co-expressed module gene list (Table 1).

2 Functional annotations of reference genes in co-expression modules

To demonstrate the functional significance of transcripts in co-expression modules, we used DAVID to do the GO (Gene Ontology) classification and some other functional analyses (Ashburner et al., 2000;Huang da et al., 2009). Firstly, we further classified the 43 eigengene modules into 16 general tissue-specific co-expression branches based on the correlation of their expression patterns. Figure 6 shows the dendrogram of 43 modules and 16 tissue-specific co-expression branches. The functional classifications of well-annotated reference genes inside tissue-specific branches showed that functional terms highly associated with tissue types were over-represented (Table S1). For example, the top three most significant over-represented biological function GO terms for "brain" co-expression branch (branch "L" in dendrogram, Figure 6) were "cell-cell signalling", "transport" and "neuron projection development". In "testes" co-expression branch (branch "K" in dendrogram, Figure 6), the top three significant over-represented biological function GO terms were "cell cycle phase", "male gamete generation" and "cell cycle process". The tissue expression analyses from DAVID also confirmed the tissue-specific expression of module genes (Table S1).

It's interesting to note that in most of 16 co-expression branches, disease-associated genes were also over represented with statistical significance (P-value < 0.05) (Table S1). Because these networks contain co-expressed intergenic ncRNAs, we believe that

intergenic ncRNAs should also be involved in these disease-associated regulatory networks.

Figure 6 describes the clustering of co-expression modules and reveals that the transcriptional sub-networks are tissue specific, but are not necessarily clustered with respect to their developmental tissue layer of origin. Whilst the transcriptional co-expression modules in human tissues are a function of their developmental origin, they can be more similar to modules from other tissues/germ layers. This suggests that these modules act as tissue specific functional sub-networks.

Discussion

Genome and transcriptome complexity of eukaryotic organisms has been confirmed by many previous studies (Frith et al., 2005;Carninci, 2006;Birney et al., 2007). An increasing number of ncRNAs, particularly long ncRNAs, have been identified from a number of organisms (Maeda et al., 2006;Guttman et al., 2009;Khalil et al., 2009;Orom et al., 2010;Cabili et al., 2011;Pauli et al., 2011). However, only a small number of these ncRNAs had annotation based on experimental data (Amaral et al., 2010;Huarte et al., 2010;Guttman et al., 2011;Hung et al., 2011;Prensner et al., 2011). The progress of functional elucidation of ncRNAs is far behind of the speed of discovery and identification of ncRNAs. Therefore, new high-throughput analysis methods are required to further interpret the functional significance of identified ncRNAs, particularly long ncRNAs.

We used human transcriptome data from 16 different human tissues to generate the expression profiles of all reference genes (mostly protein-coding genes) and more than 10,000 predicted human intergenic ncRNAs (See chapter 3). About 68% of our intergenic ncRNAs could be detected with expression (read counts) in at least 1 tissue. More intergenic ncRNAs showed tissue-specific expression compared to protein-coding genes. The tissue-specific expression of long ncRNAs has been observed by many previously published studies (Mercer et al., 2008;Ponjavic et al., 2009;Gibb et al., 2011;Guttman et al., 2011). As most previously annotated long ncRNAs were show to have regulatory roles, the tissue-specific expression of intergenic ncRNAs may indicate their regulatory roles in tissue development or other regulatory networks.

The co-expression networks built using both protein-coding genes and intergenic ncRNAs are comprehensive tools to help us understand the potential regulatory connections between coding and non-coding transcripts. Previous studies have demonstrated that many long ncRNAs exhibit highly tissue-specific, even cell-specific expression (Mercer et al., 2008;Gibb et al., 2011), so we expect that our tissue-specific co-expression modules provide more global view of this scenario. The large number of intergenic ncRNAs as part of each different tissue-specific co-expression module indicated that they might also play important regulatory roles in tissue-specific developmental pathways. The strong correlation of values of intergenic ncRNAs and protein-coding genes in tissue-specific co-expression modules indicated that ncRNAs might also contribute to the complexity of gene regulation networks in different tissues (Mattick, 2003). However, the co-expression modules in some tissues include a greater proportion of intergenic ncRNAs, particularly in module14 (lymph node) and module18 (adrenal) (Table 1). This may indicate that expression of intergenic ncRNAs might have major roles for tissue-specific regulation in some specific tissues.

For intergenic ncRNAs that were part of co-expression modules, we observed that only some of their corresponding neighbour protein-coding genes were also co-expressed in the same modules. The co-expression of intergenic ncRNAs and their neighbour gene partners was originally regarded as support for intergenic ncRNAs as transcription by-products (Struhl, 2007). It is more generally regarded now that these co-expressed intergenic ncRNAs might function as chromatin modulators (Hirota et al., 2008;Mercer et al., 2009). Their transcription may alter chromatin status to activate the transcription of neighbour genes, resulting in co-expression with intergenic ncRNAs. However, this now seems like a less general mechanism for lincRNAs, as many intergenic ncRNAs are not co-expressed with their neighbour protein-coding genes, suggesting that *trans* regulatory roles with respect to other target genes are more prevalent.

The functional classification of protein-coding genes in these modules has provided evidence for the effectiveness of clustering co-expressed and tissue-specific transcripts. All of these tissue-specific modules showed over-represented tissue-associated functional terms, representing tissue-specific developmental or regulatory processes. We also provided a classification for all human reference genes based on their co-expression in 16 different tissues. The involvement of intergenic ncRNAs enhanced this

classification by adding another RNA-layer of regulation. Consistent with the emerging regulatory roles of long ncRNAs, these tissue-specific modules will be useful candidate sources to unravel the functional significance of ncRNAs. We believe that with transcriptome data from more specific tissue/cell samples, we can sub-divide these modules into even more specific sub-networks.

Materials and methods

1 Generation of expression profiles for co-expression network re-construction

Human RNA-seq data were Illumina bodyMap2 transcriptome (http://www.ebi.ac.uk/ena/data/view/ERP000546). These RNA-seq data were sequenced from 16 different human tissues. Information about these data is shown in Table S2. These short reads were mapped to the human genome (hg19) using GSNAP with no more than 3 mismatches and splicing mapping allowed (Wu and Nacu, 2010). All mapped reads with mapping quality greater than 1 were kept for further expression normalization (Li et al., 2009). The normalized expression for each transcript, including all reference genes and intergenic ncRNAs, was calculated using Cufflinks (Trapnell et al., 2012).

2 Re-construction of co-expression networks

The expression profiles of reference genes and intergenic ncRNAs were merged. Only transcripts with expression (normalized read counts) in at least 3 tissues were kept to build the transcriptome co-expression networks. The re-construction of co-expression networks was performed using the "WGCNA" library of R (Langfelder and Horvath, 2008). In the process of network construction, the soft power used to fit the scale free topology model for expression profiles was "10". For classified eigengene modules, tissue-specific expression was annotated by comparing the mean expression of all transcripts across tissues.

3 Functional classifications of genes involved in modules

The 43 eigengene modules were clustered into general tissue-specific co-expression branches based on the correlations among modules (Figure 6). The reference genes for each branch were merged for functional annotation. The functional annotation of

reference genes was conducted by using the DAVID web server (Huang da et al., 2009). Terms, including GO terms, KEGG terms, OMIM disease terms and UP_expression terms, with P value < 0.05 and at least 5 genes were selected as the statistically significant over-represented terms.

Legends for supplemental tables

Table S1. Over-represented functional or expression-related terms for reference genes in 16 tissue-specific branches from DAVID. This excel table contains 17 sheets. The first sheet describes the summary of numbers of genes from DAVID annotation, and the rest 16 sheets are over-represented terms for genes in 16 different tissue-specific branches as shown in Table 1.

Table S2. Sample description for RNA-seq data from 16 human tissues.

References

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., and Mattick, J.S. (2010). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* 39**,** D146-151.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25**,** 25-29.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447**,** 799-816.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25**,** 1915-1927.

Carninci, P. (2006). Tagging mammalian transcription complexity. *Trends Genet* 22**,** 501-510.

Frith, M.C., Pheasant, M., and Mattick, J.S. (2005). The amazing complexity of the human transcriptome. *Eur J Hum Genet* 13**,** 894-897.

Gibb, E.A., Vucic, E.A., Enfield, K.S., Stewart, G.L., Lonergan, K.M., Kennett, J.Y., Becker-Santos, D.D., Macaulay, C.E., Lam, S., Brown, C.J., and Lam, W.L. (2011). Human cancer long non-coding RNA transcriptomes. *PLoS One* 6**,** e25915.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., and Lander, E.S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458**,** 223-227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., and Lander, E.S. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477**,** 295-300.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28**,** 503-510.

Hirota, K., Miyoshi, T., Kugou, K., Hoffman, C.S., Shibata, T., and Ohta, K. (2008). Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* 456**,** 130-134.

Huang Da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4**,** 44-57.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T., and Rinn, J.L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142**,** 409-419.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., Kong, B., Langerod, A., Borresen-Dale, A.L.,

Kim, S.K., Van De Vijver, M., Sukumar, S., Whitfield, M.L., Kellis, M., Xiong, Y., Wong, D.J., and Chang, H.Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* 43**,** 621-629.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., Van Oudenaarden, A., Regev, A., Lander, E.S., and Rinn, J.L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 106**,** 11667-11672.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9**,** 559.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**,** 2078-2079.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engstrom, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., Bult, C.J., Fletcher, C.F., Forrest, A.R., Furuno, M., Hill, D., Itoh, M., Kanamori-Katayama, M., Katayama, S., Katoh, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B.Z., Shibata, K., Sugiura, K., Takenaka, Y., Teasdale, R.D., Wells, C.A., Zhu, Y., Kai, C., Kawai, J., Hume, D.A., Carninci, P., and Hayashizaki, Y. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2**,** e62.

Mattick, J.S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25**,** 930-939.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. *Nat Rev Genet* 10**,** 155-159.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* 105**,** 716-721.

Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., and Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143**,** 46-58.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2011). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* 22**,** 577-591.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 5**,** e1000617.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J.T., Robinson, D., Iyer, H.K., Palanisamy, N., Maher, C.A., and Chinnaiyan, A.M. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 29**,** 742-749.

Qu, Z., and Adelson, D.L. (2012a). Bovine ncRNAs Are Abundant, Primarily Intergenic, Conserved and Associated with Regulatory Genes. *PLoS One* 7**,** e42638.

Qu, Z., and Adelson, D.L. (2012b). Evolutionary Conservation and Functional Roles of ncRNA. *Frontiers in Genetics* 3.

Sasaki, Y.T., Sano, M., Ideue, T., Kin, T., Asai, K., and Hirose, T. (2007). Identification and characterization of human non-coding RNAs with tissue-specific expression. *Biochem Biophys Res Commun* 357**,** 991-996.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 14**,** 103-105.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7**,** 562-578.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "Dark Matter" Transcripts Are Associated With Known Genes. *Plos Biology* 8**,** -.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26**,** 873-881.

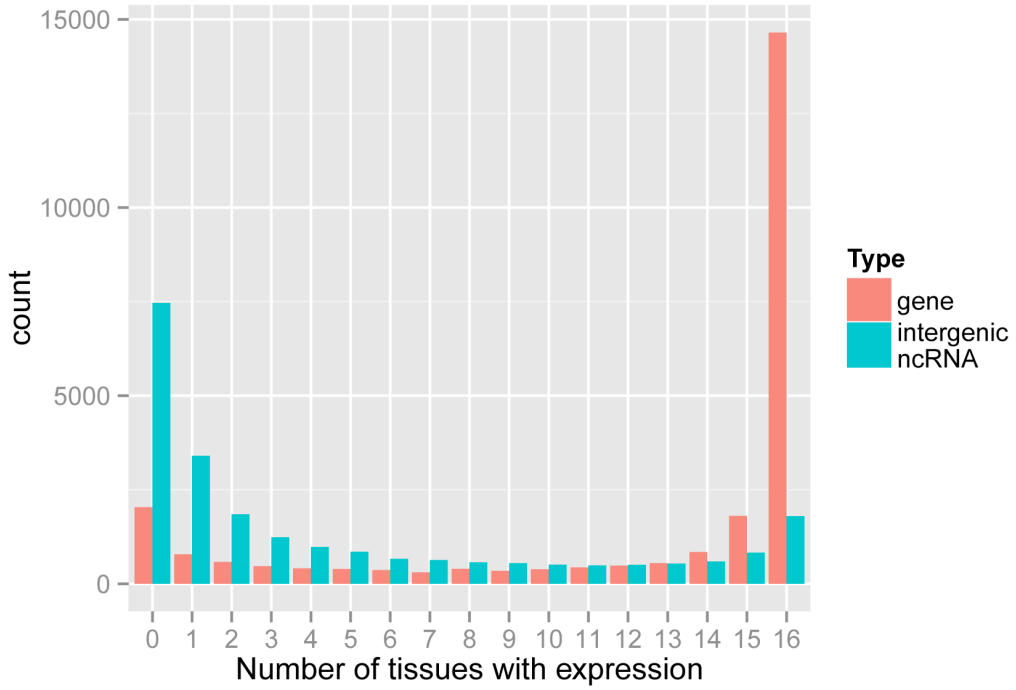

Figure 1. Frequency of expressed transcripts in 16 different tissues. The expression for each transcript was determined by the normalized read counts mapped to the transcript. The y-axis shows transcript count as a function of specific tissues (x-axis).

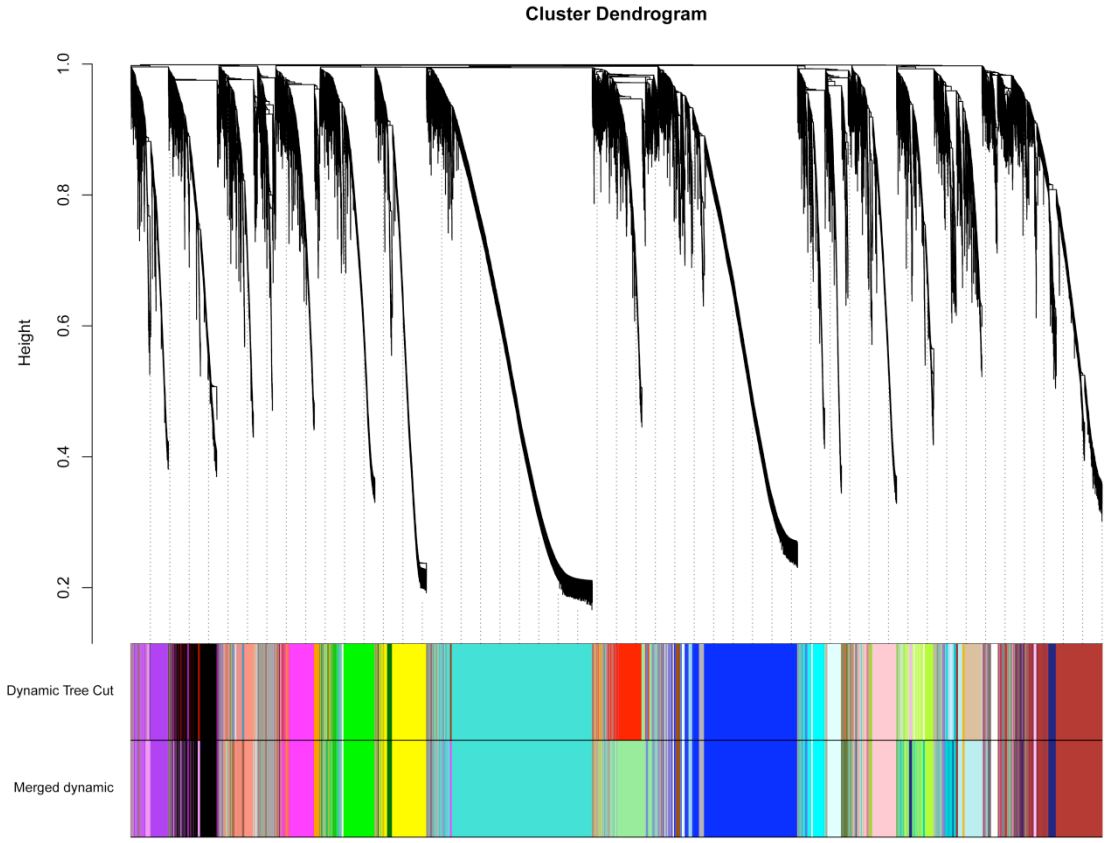

Figure 2. Dendrogram of all transcripts, including intergenic ncRNAs and reference genes, obtained by average linkage hierarchical clustering. The two color rows underneath the dendrogram show the module assignment determined by the original Dynamic Tree Cut and merged Dynamic Tree Cut respectively. Different colors represent different modules.

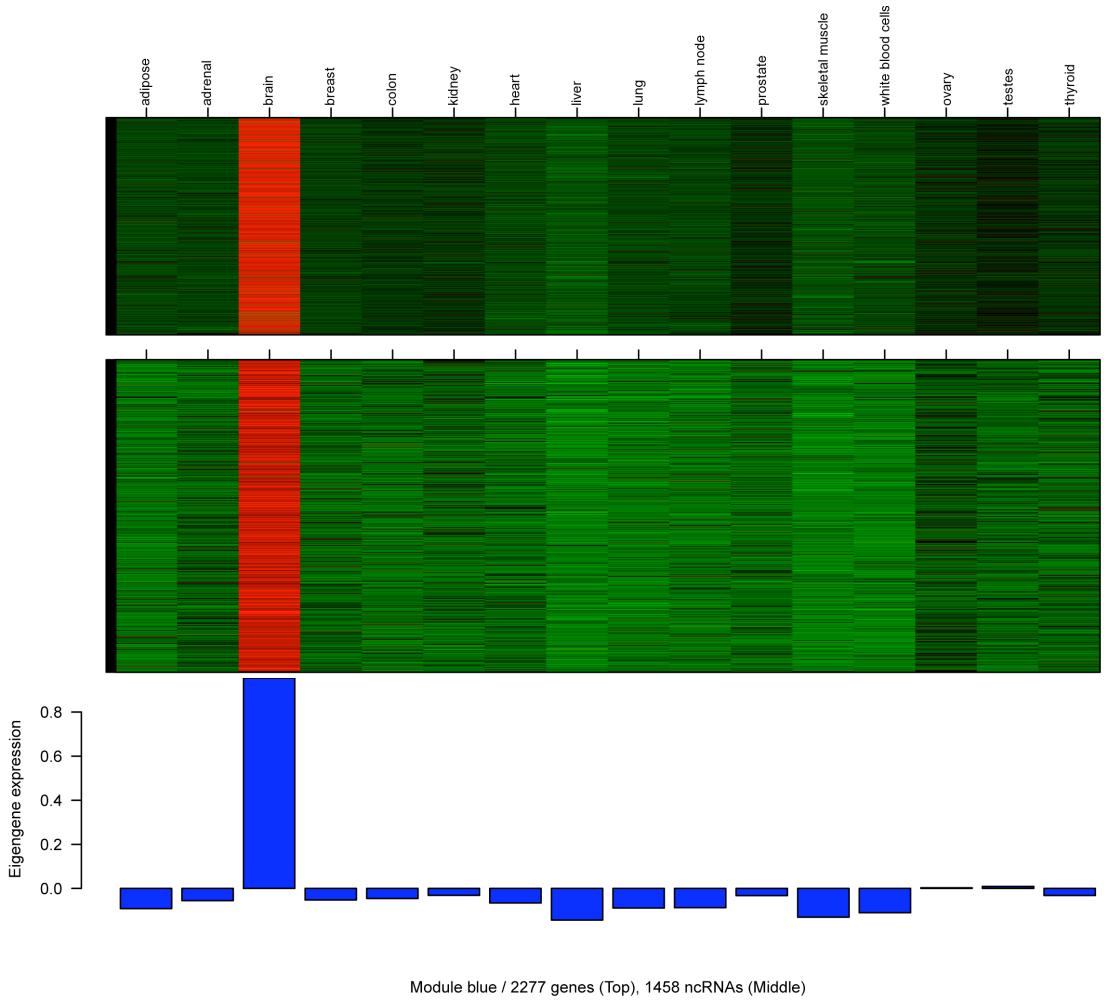

Figure 3. Expression profile of transcripts in module30 (brain). The top and middle rows show the heatmaps of genes and intergenic ncRNAs in this module respectively. The bottom row shows the corresponding module eigengene (first principal component) expression values (y-axis) versus tissues. These module eigengenes can be considered the most representative gene expression profile of the module.

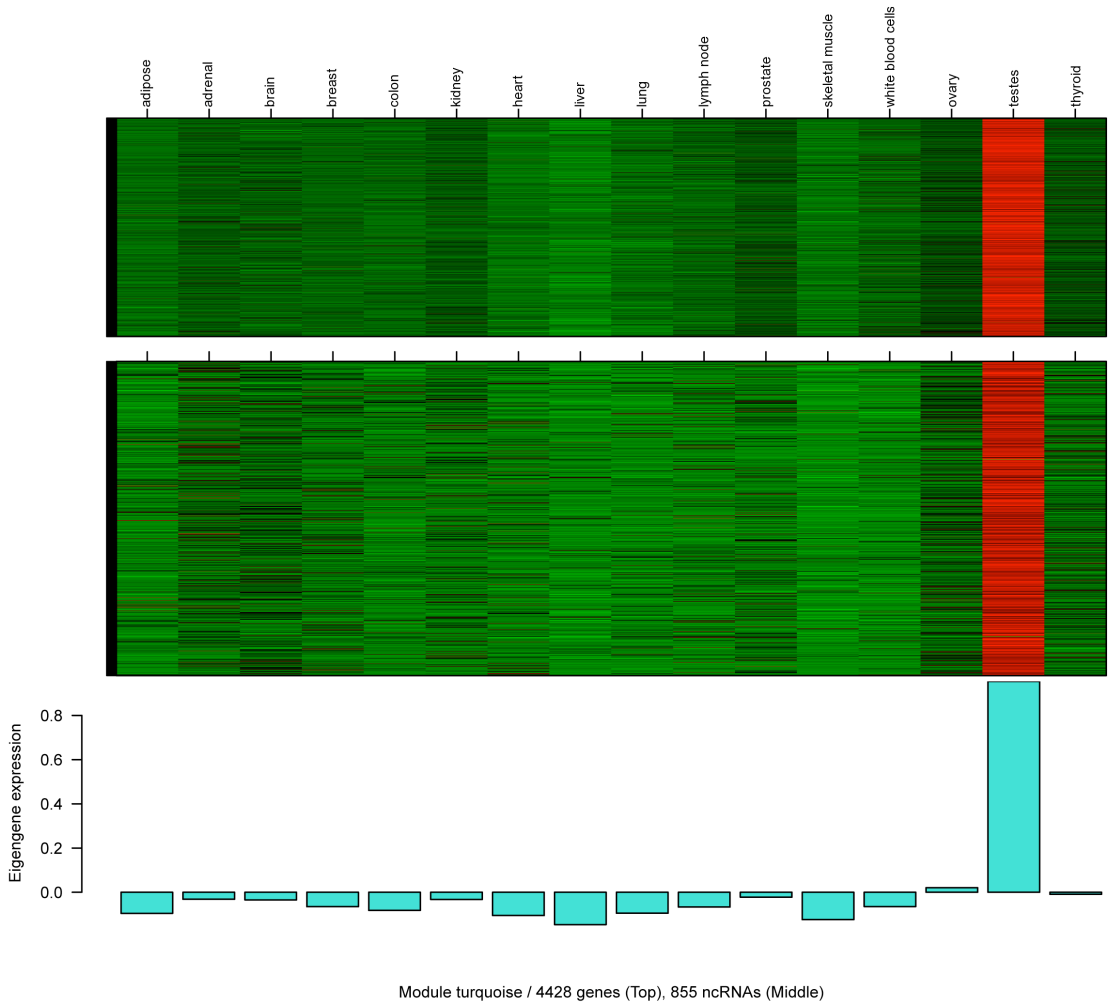

Figure 4. Expression profile of transcripts in module28 (testes). The top and middle rows show the heatmaps of genes and intergenic ncRNAs in this module respectively. The bottom row shows the corresponding module eigengenes (first principal component) expression values (y-axis) versus tissues. These module eigengenes can be considered the most representative gene expression profile of the module.

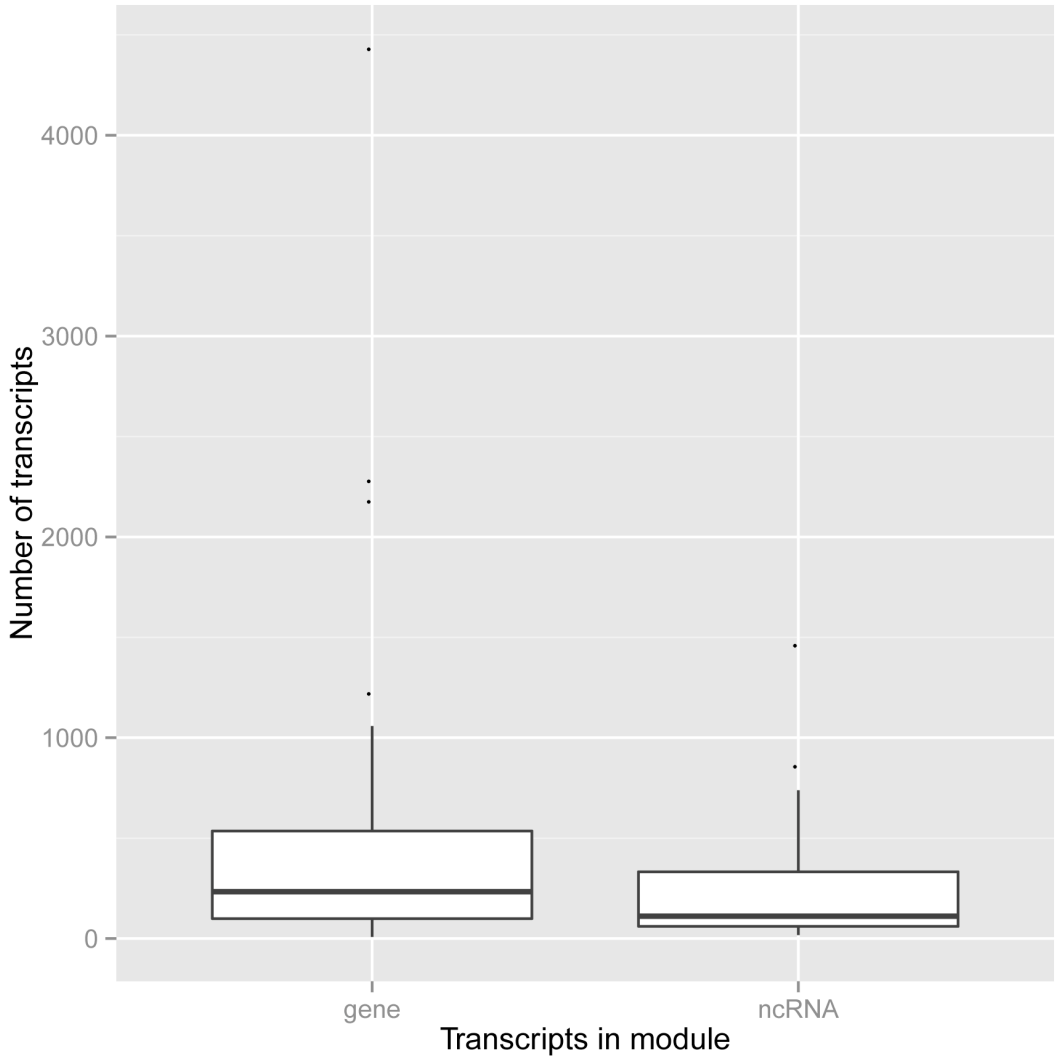

Figure 5. The numbers of transcripts, including genes and intergenic ncRNAs, in 43 co-expression modules vary significantly.

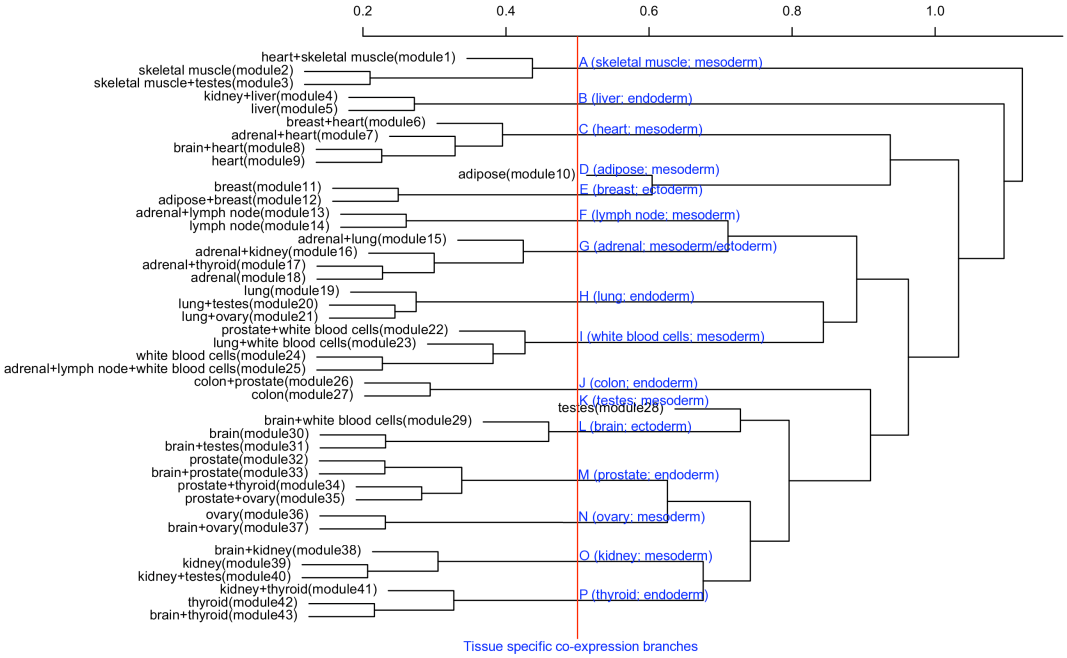

Figure 6. Hierarchical clustering of module eigengenes that summarize the modules found in the clustering analysis. End branches of the dendrogram group together eigengenes that are positively correlated. The red vertical line shows modules clustered into general “tissue specific co-expression branches”.

Chapter 5 Conclusions and Future Directions

Advances in 21st century microfluidics and automation for laboratory instrumentation have driven a brand-new "omics" view to understand the biological complexity of living organisms, particularly vertebrate animals. The sequencing and annotation of genomes and transcriptomes from human, mouse, and other organisms unveiled much more complicated regulatory networks than expected. These networks included not only protein-coding genes, but also a wide range of non-protein-coding RNAs, including small ncRNAs as well as long ncRNAs, demonstrated to be key modulators of regulatory networks in diverse biological pathways. As with protein-coding genes, a step-by-step process will be required to understand the functions and evolution of ncRNAs. The first step has been taken and encompassed the identification and functional characterisation of single or a few ncRNAs. The second step is the high throughput prediction and annotation of ncRNAs in several important model organisms, for example human, mouse and zebrafish. A significant advance in this respect has been reported in this thesis. Functional validation of these ncRNAs will be the final step and there is still a long way to go before that goal is achieved; because we know that the final step full functional annotation for protein-coding genes is still on going.

Prior to this thesis, a substantial amount of research had been done to confirm the pervasive transcription of ncRNAs and their annotation as discussed in chapter 1. However, the annotated ncRNAs were still far fewer than expected based on genomic scale analyses, particularly for cow, which is an important economic and model organism.

ESTs are rich transcription archives originally used to identify protein-coding genes. They also have the advantage of longer sequences and are sourced from a multitude of tissue types, developmental stages and treatments. We built a computational pipeline to screen potential non-protein-coding RNAs from all publicly available ESTs in cow as described in chapter 2. More than twenty thousands ncRNAs were identified from more than 1 million cow ESTs. This ncRNA dataset provided a comprehensive source for the studies of cow regulatory networks to the research community. The demonstration of sequence conservation for predicted cow ncRNAs in chapter 2 also supported the view that they should be functional. The positionally biased distribution and expression

correlation with protein-coding genes also indicated that ncRNAs might play *cis*-regulatory roles.

Further application of this pipeline allowed us to mine the non-protein-coding transcriptomes using millions of human, mouse and zebrafish ESTs as described in chapter 3. Tens of thousands of ncRNAs were identified from these 3 organisms, greatly extending the catalog of known ncRNAs. Some common features across organisms were also observed based on comparative analyses of these ncRNAs, in combination with the results in chapter 2. These features included clear sequence conservation and positionally biased location proximate to regulatory and developmental protein-coding genes. Although previous studies have reported these common features (see discussion in chapter 3), our results were based on the same type of transcriptome data and identification method, providing more accurate and comprehensive support for these properties. The most interesting result was that although intergenic ncRNAs were preferentially located proximate to regulatory and developmental genes in all 4 organisms, these neighbour genes were not the same across species. This indicated possible species-specific regulatory roles for these intergenic ncRNAs.

After identifying and reconstructing these non-protein-coding transcripts, the next step was to try and interpret the functional significance of these ncRNAs. In chapter 4, we focused on only the intergenic ncRNAs and re-constructed co-expression networks combining protein-coding genes with these ncRNAs. These co-expression networks classified human transcripts, including protein-coding and non-coding, into tissue-specific co-expression modules. The functional annotation of protein-coding genes in these co-expression modules confirmed the tissue-specific expression patterns and revealed the extent of possible tissue-specific regulatory connections between protein-coding genes and ncRNAs in human.

Our work has comprehensively identified and annotated ncRNAs at genome scale. The results from the comparative analyses and co-expression networks have provided valuable information and can be used as a starting point for the final step of functional validation analysis. The expression profiles used to reconstruct co-expression networks are based on recently sequenced RNA-seq data. Large numbers of our ncRNAs,

particularly intergenic ncRNAs, are expressed in single or multiple samples of these transcriptome data. This also provides strong support to the hypothesis that these ncRNAs are real transcription instead of artifacts. Due to technical aspects, there were also some limitations in our pipeline and analyses: First, we removed all repetitive elements from raw ESTs to avoid possible mis-assembly problems. Therefore, repetitive element derived ncRNAs were not in our ncRNA datasets, even though we know these exist. Second, although we clustered and assembled ESTs into longer consensus transcripts, many of our ncRNAs are probably not full-length transcripts. Third, many ESTs were sequenced without knowing their transcriptional orientation, affecting the annotation of transcription orientation of some predicted ncRNAs. Therefore, subsequent analyses should focus on rectifying the limitations of these current analyses. For example, combining additional RNA-seq data with our dataset to deduce the full-length transcripts of ncRNAs and get full transcription orientation information. However, the biggest challenge remains how to design and perform loss of function and rescue experiments to validate the regulatory roles of ncRNAs. Our protein-coding and non-coding co-expression networks also revealed tissue-specific and potential disease-associated regulatory roles of ncRNAs. If we can elucidate the regulatory mechanisms of ncRNAs in these regulatory networks, it will provide a clearer understanding of disease pathogenesis and an entry point for drug design to target ncRNA regulatory networks.

Supplemental Materials

The attached CD-ROM/DVD-ROM contains supplemental documents, figures and tables for chapter 2, chapter 3 and chapter 4.

For chapter 2, there are one supplemental document, 8 supplemental figures (Figure S1 to Figure S8) and 9 supplemental tables (Table S1 to Table S9). See chapter 2 for legends.

For chapter 3, there are one document describing the supplemental results, 8 supplemental figures (Figure S1 to Figure S8) and 3 supplemental tables (Table S1 to Table S3). See chapter 3 for legends.

For chapter 4, there are 2 supplemental tables (Table S1 and Table S2). See chapter 4 for legends.