New Player of ncRNAs: Long Non-coding RNAs

Abstract
Long non-coding RNAs (lncRNAs) play important roles in a wide range of biological processes as regulatory factors at the epigenetic, transcriptional and post-transcriptional levels. In this review, we summarized the current knowledge of lncRNAs discoveries including their identification, classifications and functions.

Keywords: Long non-coding RNAs; Discovery of lncRNAs; Classification of lncRNAs; Post-transcriptional levels; Epigenetic; Regulatory factors

Introduction
The mechanism underlying the functions of non-protein coding RNAs (ncRNAs or npcRNAs) that have no or little protein-coding potential is a fascinating area of research [1]. Based on transcript length, ncRNAs are classified as short (<200 nt) and long ncRNAs (lncRNAs; >200 nt). The recent high-throughput analysis such as cDNA/EST in silico mining, whole-genome tiling array and RNA-sequencing (RNA-seq) has revealed that the transcription landscape in eukaryotes is much more complex than had been expected [2-4]. Transcriptome analysis estimates transcripts over 90% of eukaryotic genome [5]. These approaches have facilitated the identification of thousands of novel ncRNAs (or npcRNAs) in many organisms, such as humans, animals, and plants [6-9].

LncRNAs are arbitrarily defined as RNA transcripts that contain >200 nt but lack protein coding potential which are transcribed by RNA polymerase II or III, and additionally, by polymerase IV/V in plants [10-12]. They are processed by splicing or nonsplicing, polyadenylation or non-polyadenylation, and can be located in the nucleus or cytoplasm. The researches have revealed that lncRNAs may represent alternatively spliced forms of known genes [13], products of antisense RNAs [14-17], double stranded RNAs [18], retained introns [13,19], short open reading frame [1,20,21], RNA polymerase III-derived RNAs [22] and RNA decoys mimicking miRNA targets [23].

Discovery of lncRNAs
In 1990s, H19 and Xist (X-inactive specific transcript) lncRNAs were discovered by using traditional gene mapping approaches [24-16]. In the later years, HOTAIR (HOX antisense intergenic RNA) and HOTTIP (HOKA transcript at the distal tip) were discovered by using tiling arrays in the homeobox gene regions (HOX clusters) [27,28]. Using genome-wide approach, 1600 novel mouse lncRNAs have been identified by Guttman et al. [8]. Since then, thousands of lncRNAs have been determined using similar genome-wide approaches in human, mouse and plants [29-32].

Novel lncRNAs can be detected and discovered by both experimental (next generation sequencing, NGS, technologies) and computational screenings [33-35]. First, the fragments of transcripts are obtained by using NGS technologies or tilling microarrays. Then, the transcripts sequences are mapped to the reference genome and identified transcribed units of the RNAs. The criteria for discriminating between coding and non-coding sequences of RNAs are based on similarity to known coding sequences or statistics of codon frequencies for coding potential [36]. Typically, BLASTX is most commonly used tool for known sequence similarity detection [37]. Alternatively, HMMER3 help to determine homologous domains in protein data to eliminate transcripts with protein-coding potential [38]. However, there is much more alternative tools for evaluating coding potential. The most used tools are CPC (Cording-Potential Calculator) [39] and PORTRAIT [40] use pair wise comparisons; in contrast, PhyloCSF [41] and RNAcode [42] use multiple alignments. Another popular approach, Coding Potential Assessment Tool, also uses an alignment-free logistic regression model [43]. Except these computational approaches, experimental methods such as ribosomal profiling have been utilized to compute the protein coding capacity of lncRNAs based on the periodicity of ribosome occupancy along the short translated ORFs [44].

About 1600 novel mouse lncRNAs have been identified by genome-wide approach which used gene expression data and the presence of chromatin marks for promoter regions [8]. Combination of chromatin marks and RNA-seq data sets have been used to generate the human long intervening non-coding (lincRNA) catalog which comprise 8000 lincRNAs from 24 different human cell types and tissues [45]. More than 13,500 human lncRNAs have been annotated by GENCODE and also, datasets from the 1000 Genomes Project have been utilized to reveal the association between lncRNAs and prostate cancer [30,46]. Cunnington et al. have reported the association between 56 lncRNAs and disease related to traits ranging from diabetes to multiple sclerosis, Alzheimer’s disease, etc [47]. Both computational and experimental analyses have shown that 125
putative stress responsive IncRNAs in wheat were tissue-specific and can be induced by powdery mildew infection and heat stress [48]. In addition, Zhang et al. [15] systematically identified 2224 IncRNAs by performing strand-specific RNA sequencing of rice anthers, pistils, seeds, and shoots and combining with the analysis of other available rice RNA-seq datasets [32].

Classification of IncRNAs

IncRNAs are classified based on several properties such as transcript length, sequence and structure conservation, genomic location, functions exerted on DNA or RNA, functioning mechanisms, and targeting mechanisms, association with annotated protein coding genes or repeats or biochemical pathway or stability or subcellular structures [49,50]. Besides lots of criteria for IncRNA classification, the most commonly used attributes are their size, localization and function. Typically, the threshold value is 200 bases for length discrimination of ncRNAs. Fewer than 200 bases are considered as small ncRNAs and more than 200 bases are classified as long ncRNAs [51]. After length size discrimination, genomic locations of IncRNAs are also popular for classifying. According to GENCODE for their genomic locations, IncRNAs are classified into five groups:

I. Antisense IncRNAs, which are transcribed from the antisense strand, intersect any exon of a protein-coding locus on the opposite strand, or published evidence of antisense regulation of a coding gene. Their transcription was found to be overlap genes related to condition specific or the stress response. It is considered that antisense IncRNAs, which involve genome imprinting, regulation of alternative splicing and translation, exert their function as on-off switch for these genes [52-54].

II. Sense IncRNAs are transcribed from sense strand of protein-coding genes that overlapping transcripts contain a coding gene within an intron on the same strand.

III. Intronic transcripts reside within introns of a coding gene, which do not have exon-exon overlapping, is defined as sense intronic IncRNAs. Differential expression studies demonstrated that expression levels of intronic IncRNAs and their biological variation during a physiological time course, or among different individuals of the same strain are tightly correlated with their adjacent exons [55].

IV. Long intervening non-coding RNAs with a length >200 bp, are also called long “intergenic” non-coding RNAs, do not overlap exons of either protein coding and lies within the genomic interval between two genes. Approximately 20% of lincRNAs are found to be bound by polycromy repressive complex 2 (PRC2) or other chromatin-modifying complexes which indicated that they play role as enhancer-like functions by guiding chromatin-modifying complexes to specific genomic loci, transmitting information from higher order chromosomal looping into chromatin modifications to coordinate long-range gene activation [28,56].

V. Processed transcript which do not have any open reading frame (ORF) and also, cannot be placed in any type of categories [57]. In addition to GENCODE classification, extra two categories are also emerged as bidirectional and enhancer IncRNAs. Bidirectional IncRNAs, which are tending to be highly conservative, are expressed within 1 kb of promoters in the opposite direction from the neighboring protein-coding gene [58,59]. Several studies showed that bidirectional IncRNAs are associated with transcriptional regulatory genes implicated in cell differentiation and development [60]. Enhancer IncRNA (elncRNA or eRNA), which are generally <2 kb, is transcribed from enhancer regions of the genome and may contribute to enhancer function [59], eRNAs have been found to exert their functions in chromatin looping and long-range gene activation, playing an important role in system development and the formation of homeostasis [61,62].

Conclusion

LncRNAs play important roles in a numerous biological processes as regulatory factors. Functional analyses of IncRNAs have indicated that they are effective cis- and transregulators of gene transcription, and also act as scaffolds for chromatin-modifying complexes. Nowadays, IncRNAs are considered as major regulators involved in numerous cellular processes, including cell differentiation and development, chromosome dosage compensation, cell cycle control and adaptation to environmental changes [63-65]. Our group has been investigating the association between salinity stress metabolism and barley IncRNAs (unpublished data). Identification of novel IncRNAs is likely to provide new insight into the complicated gene regulatory network involving IncRNAs, provide novel diagnostic opportunities, and pinpoint novel therapeutically targets.

Acknowledgement

Dr. Nermin Gözükırmızı is professor at Molecular Biology and Genetics Department, Science Faculty of Istanbul University, Turkey. She received her Dr. rer. nat. Degree in Botany-Genetics at Istanbul University, Turkey in 1979. Her current research interest focuses on plant stress metabolism, transposons and epigenetics marks.

M. Sc. Elif Karlık is PhD Can didate at Biotechnology Department, Institute of Science of Istanbul University, Turkey. She received her master degree in Molecular Biology and Genetics at Istanbul University, Turkey. Her current research interest focuses on plant stress metabolism and regulatory network of IncRNAs.

References


