

# 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

Volume 3 Issue 4 - 2016

 Elif Karlik,<sup>1</sup> Nermin Gozukirmizi<sup>2</sup>
<sup>1</sup>Department of Biotechnology, Istanbul University, Turkey

<sup>2</sup>Department of Molecular Biology and Genetics, Istanbul University, Turkey

**Correspondence:** G Nermin Gözükırmızı, Department of Molecular Biology and Genetics, Istanbul University, Faculty of Science, Istanbul, 34118 Vezneciler, Turkey, Email nermin@istanbul.edu.tr

**Received:** November 04, 2016 | **Published:** November 15, 2016

## 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.<sup>1</sup> 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.<sup>2-4</sup> Transcriptome analysis estimates transcripts cover 90% of eukaryotic genome.<sup>5</sup> These approaches have facilitated the identification of thousands of novel ncRNAs (or npcRNAs) in many organisms, such as humans, animals, and plants.<sup>6-9</sup>

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.<sup>10-12</sup> 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,<sup>13</sup> products of antisense RNAs,<sup>14-17</sup> double stranded RNAs,<sup>18</sup> retained introns,<sup>13,19</sup> short open reading frame.<sup>1,20,21</sup> RNA polymerase III-derived RNAs<sup>22</sup> and RNA decoys mimicking miRNA targets.<sup>23</sup>

## Discovery of lncRNAs

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

Novel lncRNAs can be detected and discovered by both experimental (next generation sequencing, NGS, technologies) and computational screenings.<sup>33-35</sup> First, the fragments of transcripts are obtained by using NGS technologies or tiling 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.<sup>36</sup> Typically, BLASTX is most commonly used tool for known sequence similarity detection.<sup>37</sup> Alternatively, HMMER3 help to determine homologous domains in protein data to eliminate transcripts with protein-coding potential.<sup>38</sup> However, there is much more alternative tools for evaluating coding potential. The most used tools are CPC (Coding-Potential Calculator)<sup>39</sup> and PORTRAIT<sup>40</sup> use pair wise comparisons; in contrast, PhyloCSF<sup>41</sup> and RNAcode<sup>42</sup> use multiple alignments. Another popular approach, Coding Potential Assessment Tool, also uses an alignment-free logistic regression model.<sup>43</sup> 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.<sup>44</sup>

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.<sup>8</sup> 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.<sup>45</sup> 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.<sup>30,46</sup> Cunningham et al. have reported the association between 56 lncRNAs and disease related to traits ranging from diabetes to multiple sclerosis, Alzheimer's disease, etc.<sup>47</sup> Both computational and experimental analyses have shown that 125 putative stress responsive lncRNAs in wheat were tissue-specific and can be induced by powdery mildew infection and heat stress.<sup>48</sup> In addition, Zhang et al.<sup>15</sup> systematically identified 2224 lncRNAs 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.<sup>32</sup>

## Classification of lncRNAs

lncRNAs 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.<sup>49,50</sup> Besides lots of criteria for lncRNA 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.<sup>51</sup> After length size discrimination, genomic locations of lncRNAs are also popular for classifying. According to GENCODE for their genomic locations, lncRNAs are classified into five groups:

- I. Antisense lncRNAs, 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 lncRNAs, which involve genome imprinting, regulation of alternative splicing and translation, exert their function as on-off switch for these genes.<sup>52–54</sup>
- II. Sense lncRNAs 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 lncRNAs. Differential expression studies demonstrated that expression levels of intronic lncRNAs and their biological variation during a physiological time course, or among different individuals of the same strain are tightly correlated with their adjacent exons.<sup>55</sup>
- 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 lncRNAs are found to be bound by polycomb 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.<sup>28,56</sup>
- V. Processed transcript which do not have any open reading frame (ORF) and also, cannot be placed in any type of categories.<sup>57</sup> In addition to GENCODE classification, extra two categories are also emerged as bidirectional and enhancer lncRNAs. Bidirectional lncRNAs, which are tending to be highly conservative, are expressed within 1 kb of promoters in the opposite direction from the neighboring protein-coding gene.<sup>58,59</sup> Several studies showed that bidirectional lncRNAs are associated with transcriptional regulatory genes implicated in cell differentiation and development.<sup>60</sup> Enhancer lncRNA (elncRNA or eRNA), which are generally <2 kb, is transcribed from enhancer regions of the genome and may contribute to enhancer function.<sup>59</sup> 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.<sup>61,62</sup>

## Conclusion

lncRNAs play important roles in a numerous biological processes as regulatory factors. Functional analyses of lncRNAs have indicated that they are effective cis- and transregulators of gene transcription, and also act as scaffolds for chromatin-modifying complexes. Nowadays, lncRNAs 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.<sup>63–65</sup> Our group has been investigating the association between salinity stress metabolism and barley lncRNAs (unpublished data). Identification of novel lncRNAs is likely to provide new insight into the complicated gene regulatory network involving lncRNAs, provide novel diagnostic opportunities, and pinpoint novel therapeutically targets.

## Acknowledgements

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 candidate at Biotechnology Department, Institution 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 lncRNAs.

## Conflicts of interest

The authors declare no conflict of interest.

## Funding

None.

## References

1. Ben Amor B, Wirth S, Merchan F, et al. Novel long non-protein coding RNAs involved in *Arabidopsis* differentiation and stress responses. *Genome Res.* 2009;19(1):57–69.
2. Maeda N, Kasukawa T, Oyama R, et al. Transcript annotation in FANTOM3: Mouse gene catalog based on physical cDNAs. *PLoS Genet.* 2006;2(4):e62.
3. Jacquier A.. The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet.* 2009;10(12):833–844.
4. Khachane AN, Harrison PM. Mining mammalian transcript data for functional long non-coding RNAs. *PLoS One.* 2010;5(4):e10316.
5. Taft RJ, Pheasant M, Mattick JS. The relationship between nonprotein-coding DNA and eukaryotic complexity. *Bioessays.* 2007;29(3):288–299.
6. Ravasi T, Suzuki H, Pang KC, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* 2006;16(1):11–19.
7. Consortium EP, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007;447(7146):799–816.
8. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature.* 2009;458:223–227.
9. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell.* 2009;136(4):629–641.
10. Dinger ME, Pang KC, Mercer TR, et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol.* 2008;4(11):e1000176.

11. Wierzbicki AT, Haag JR, Pikaard CS. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*. 2008;135(4):635–648.
12. Dinger ME, Pang KC, Mercer TR, et al. NRED:a database of long noncoding RNA expression. *Nucleic Acids Res*. 2009;37:D122–D126.
13. Filichkin SA, Priest HD, Givan SA, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*. 2010;20(1):45–58.
14. Matsui A, Ishida J, Morosawa T, et al. *Arabidopsis* transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol*. 2008;49(8):1135–1149.
15. Zhang X, Lii Y, Wu Z, et al. Mechanisms of small RNA generation from cis-NATs in response to environmental and developmental cues. *Mol Plant*. 2013;6(3):704–715.
16. Swiezewski S, Liu F, Magusin A, et al. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature*. 2009;462(7274):799–802.
17. Luo C, Sidote DJ, Zhang Y, et al. Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. *Plant J*. 2013;73(1):77–90.
18. Rajeswaran R, Aregger M, Zvereva AS, et al. Sequencing of RDR6-dependent double-stranded RNAs reveals novel features of plant siRNA biogenesis. *Nucleic Acids Res*. 2012;40:6241–6254.
19. Ner-Gaon H, Halachmi R, Goldstein SS, et al. Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J*. 2004;39(6):877–885.
20. Hanada K, Takeuchi HM, Okamoto M, et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci USA*. 2013;110(6):2395–2400.
21. Moghe GD, Lehti-Shiu MD, Seddon AE, et al. Characteristics and significance of intergenic polyadenylated RNA transcription in *Arabidopsis*. *Plant Physiol*. 2013;161(1):1210–1224.
22. Wu J, Okada T, Fukushima T, et al. A novel hypoxic stress-responsive long non-coding RNA transcribed by RNA polymerase III in *Arabidopsis*. *RNA Biol*. 2012;9(3):302–313.
23. Zorrilla FJM, Valli A, Todesco M, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet*. 2007;39(8):1033–1037.
24. Brannan CI, Dees EC, Ingram RS, et al. The product of the H19 gene may function as an RNA. *Mol Cell Biol*. 1990;10(1):28–36.
25. Borsani G, Tonlorenzi R, Simmler MC, et al. Characterization of a murine gene expressed from the inactive X chromosome. *Nature*. 1991;351(6324):325–329.
26. Brockdorff N, Ashworth A, Kay GF, et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*. 1992;71(3):515–526.
27. Rinn JL, Kertesz M, Wang JK, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. 2007;129(7):1311–1323.
28. Wang KC, Yang YW, Liu B, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*. 2011;472(7341):120–124.
29. Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA*. 2009;106(28):11667–11672.
30. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22(9):1775–1789.
31. Liu J, Jung C, Xu J, et al. Genome wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell*. 2012;24(11):4333–4345.
32. Zhang YC, Liao JY, Li ZY, et al. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biology*. 2014;15:512.
33. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–515.
34. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 2013;14:R36.
35. Dobin A, Davis CA, Schlesinger F, et al. STAR:ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
36. Iwakiri J, Hamada M, Asai K, et al. Bioinformatics tools for lncRNA research. *Biochimica et Biophysica Acta – Gene Regulatory Mechanisms*. 2016;1859(1):23–30.
37. Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet*. 1993;3(3):266–272.
38. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011;7(10):e1002195.
39. Kong L, Zhang Y, Ye ZQ, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 2007;35:W345–W349.
40. Arrial RT, Togawa RC, Brigido MM. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics*. 2009;10:239.
41. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27(13):i275–i282.
42. Washietl S, Findeiss S, Muller SA, et al. NCcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*. 2011;17(4):578–594.
43. Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41:e74.
44. Bazzini AA, Johnstone TG, Christiano R, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*. 2014;33(9):981–993.
45. Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25(18):1915–1927.
46. Jin G, Sun J, Isaacs SD, et al. Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis*. 2011;32(11):1655–1659.
47. Cunningham MS, Santibanez-Koref M, Mayosi BM, et al. Chromosome 9p21 SNPs associated with multiple disease phenotypes correlate with ANRIL expression. *PLoS Genet*. 2010;6(4):e1000899.
48. Xin M, Wang Y, Yao Y, et al. Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biol*. 2011;11:61.
49. Ma L, Bajic VB, Zhang Z. On the classification of long noncoding RNAs. *RNA Biol*. 2013;10(6):925–933.

50. St Laurent G, Wahlestedt C, Kapranov P. The Landscape of long non-coding RNA classification. *Trends Genet.* 2015;5:239–251.
51. Kapranov P, Cheng J, Dike S, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science.* 2007;316(5830):1484–1488.
52. Berteaux N, Aptel N, Cathala G, et al. A novel H19 antisense RNA overexpressed in breast cancer contributes to paternal IGF2 expression. *Mol Cell Biol.* 2008;28(22):6731–6745.
53. Ni T, Tu K, Wang Z, et al. The prevalence and regulation of antisense transcripts in *Schizosaccharomyces pombe*. *PLoS One.* 2010;5(12):e15271.
54. Xu Z, Wei W, Gagneur J, et al. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol.* 2011;7:468.
55. St Laurent G, Shtokalo D, Tackett MR, et al. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics.* 2012;13:504.
56. Zhang K, Shi ZM, Chang YN, et al. The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene.* 2014;547(1):1–9.
57. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22(9):1760–1774.
58. Atkinson SR, Marguerat S, Bahler J. Exploring long non-coding RNAs through sequencing. *Semin Cell Dev Biol.* 2012;23(2):200–205.
59. Devaux Y, Zangrando J, Schroen B, et al. Long noncoding RNAs in cardiac development and ageing. *Nat Rev Cardiol.* 2015;12(7):415–425.
60. Lepoivre C, Belhocine M, Bergon A, et al. Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics.* 2013;14:914.
61. De Santa F, Barozzi I, Mietton F, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* 2010;8(5):e1000384.
62. Licastro D, Gennarino VA, Petrera F, et al. Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics.* 2010;11:151.
63. Wery M, Kwapisz M, Morillon A. Noncoding RNAs in generegulation. *Wires systems Biology and Medicine* 3(6):728–738.
64. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem.* 2011;81:145–166.
65. Sole C, Ribelles NM, de Nadal E, et al. A novel role for lncRNAs in cell cycle control during stress adaptation. *Curr Genet.* 2015;61(3):299–308.