OPEN ACCESS Freely available online

Journal of Proteomics & Bioinformatics

Review Article

Deciphering the Role of Long Non-coding RNAs in Cancer Using Modern Bioinformatics Approaches: A Comprehensive Review

Pankaj Jyoti Barua*

Bioinformatics consultant and Veterinary Surgeon at VETPJB (A Veterinary Diagnostics and Clinical Establishment), Assam, Índia

ABSTRACT

The structural and molecular characterization of long non-coding DNA molecules has created tremendous scope for gaining insights into the molecular genetics of different types of malignancies. Studies have clearly indicated that these long-non coding DNAs are excellent candidates as far as cancer interventions are concerned and recent developments in different bioinformatics tools and paradigms have completely changed our cancer therapeutic and diagnostic outlook. The inherent objective of this review article in to present an overview on the molecular characteristics of lncRNAs and their role in cancer progression and discuss the far reaching implications for cancer diagnostics and therapeutics through bioinformatics analysis of lncRNAs. This review elaborates on the bioinformatics analysis of RNA-seq data from expression profiling of lncRNAs and also emphasizes on the significance of different lncRNA databases.

Keywords: Cancer; Bioinformatics; Physiological analysis; Proteomics; iTRAQ LC-MS/MS

INTRODUCTION

When Francis Crick proposed the central dogma in 1958 stating that within the the nucleus of every cell there is transcription of DNA to RNA and translation of RNA to protein, scientists were unaware of the fact that less than 3% of the human DNA get expressed into protein [1]. This aspect came to light with the completion of the human genome project and it became clear that more than 97% of the human DNA is essentially "junk" and do not conform to the central dogma proposed by Francis Crick [1,2]. It was also assumed that these noncoding regions of the genome do not come under any form of selective pressure, allowing mutations to accumulate without any negative impact on the host organism. Considering the fact that more than 97% of the human genome is supposedly "noise" a key scientific inquisitiveness that needs to be addressed is wheather they are truly redundant in nature or if they have any significant regulatory role to play. Major functional annotation initiatives such as ENCODE and FANTOM have indicated that around 80% of the DNA in higher organisms such as mammals undergo transcription into noncoding RNA elements that are also subjected to extensive genetic regulation [3,4]. The volume of noncoding RNA in different mammalian species varies significantly and it has been observed that in higher species the abundance of the molecules increase significantly, thus indicating their molecular significance [5,6]. Majority of the non-coding RNAs present in the higher mammals are at least 200 base pairs long or more and play an active role in the epigenetic and transcriptional regulation of gene expression and enzymatic activity modulation [7]. Recent bioinformatics ribosomal profiling analyses have clearly indicated that large sections of non-coding RNAs contain highly conserved small ORFs that interact and bind with ribosome [8-10]. This observation is a strong indication of the coding capacity of the noncoding RNAs and it appears that until recently the molecular capabilities of these supposedly "redundant" genetic material have largely been overlooked [11,12]. The non-coding RNAs longer than 200 base pairs are referred to as long non-coding RNAs (lncRNAs) and bioinformatics analysis of targeted RNA sequencing data along with genome wide association studies have allowed annotation of large populations of multi-exonic non-coding RNAs (mencRNAs) associated to malignancies such as breast cancer [13]. At the moment scientific endeavour from across the world have been highly successful in cataloguing thousands of long non-coding RNAs (lncRNAs) and some key lncRNA that are implicated in cancer metastasis include MALAT1 which is over-expressed in a wide range of cancer types, HOTAIR associated with breast cancer and colon cancer and PCNCR1 associated with prostate cancer [14-18]. Considering the fact that majority of the annotated lncRNAs have still not been functionally characterized it is difficult to say how many more lncRNAs contribute towards cancer progression and metastasis [19,20]. Modern bioinformatics approaches that are available today can be very effective in identification and annotation of for lncRNAs at systemic and functional levels and could contribute immensely towards better understanding of cancer pathogenesis. LncRNAs have very low expression levels with majority of them devoid of the poly-A tail, making their identification and annotation a major challenge. However latest sequencing approaches such as RNA-seq, ChIRP-Seq and genome editing tools such as CRISPR have been very effective in high resolution identification and annotation of novel lncRNAs [21,22]. Furthermore, availability of lncRNA databases

Correspondence to: Pankaj Jyoti Barua, Bioinformatics consultant and Veterinary Surgeon at VETPJB (A Veterinary Diagnostics and Clinical Establishment), Assam, Índia, E-mail: pankajjbarua@gmail.com

Received: January 26, 2021; **Accepted:** February 09, 2021; **Published:** February 16, 2021

Citation: Barua PJ (2021) Deciphering the Role of Long Non-coding RNAs in Cancer Using Modern Bioinformatics Approaches: A Comprehensive Review. J Proteomics Bioinform. 14:532.

such as GENCODE, NONCODE, LncRNASNP2, LNCipedia, CHIPBase, DIANA-LncBase, Noncode v3.0 and lncRNome and bioinformatics computational analysis and prediction tools such as LncDisease, LncRscan-SVM and LncRNA-MFDL have been very resourceful in identifying and characterizing the properties and functions of lncRNAs [23].

This review will start by describing the molecular characteristics and functional mechanism of lncRNAs in cancer pathogenesis and discuss the promise of lncRNA biomarkers for early characterization of different cancer phenotypes. The review will then present a comprehensive discussion on the different bioinformatics analysis and prediction tools and annotation databases for lncRNAs and elaborate on how they can aid in better understanding of the molecular pathogenesis of different types of cancer.

THE MOLECULAR CHARACTERISTICS OF LNCRNAS AND THEIR ROLE IN CANCER PROGRESSION

The concentration of lncRNAs in the genome is very high and majority of them do not possess active open reading frames [24,25]. Furthermore the expression levels of most of the lncRNAs are very low and this makes is very challenging to properly characterize them. While modern sequencing approaches such as RNA-seq have clearly indicated that lncRNA conservation in the DNA is very insignificant, around 3% of the molecule is highly conserved in humans [26]. It is very likely that lncRNAs evolve very rapidly in certain organisms and they might not need a high level of sequence conservation to retain their inherent functional characteristics. Interestingly, the promoters regions of lncRNAs are found to be highly conserved thus indicating the fact that genetic regulation of lncRNA expression has lot of significance. A very large number of lncRNAs have very similar molecular characteristics with active protein coding genes such as more than exons, polyA+ tails and a 5′ cap and majority of the lncRNA genes are situated within a 10 kb region of the protein coding genes [27,28]. LncRNA can also exist as intronic RNA and antisense RNA as illustrated in Figure 1.

Thus, it is apparent that there is a good level of diversity as far as the functional profile of the lncRNAs is concerned and they are also present in different types of tissues. Studies have also shown that the diversity of lncRNA is highest in the CNS and they can be present both in the cytoplasm and the cell nucleus [29,30]. Considering the fact that lncRNAs have low levels of expression and also due to the discovery of promoter-associated RNA types such as promoter upstream transcripts that have very little molecular stability, these long non-coding molecules were also presumed to be highly unstable [31]. However recent studies have been able to establish that majority of lncRNAs are very stable with very healthy half-lives [32]. LncRNAs contribute very diversely towards cancer pathogenesis and it includes chromatin remodelling and looping, acting as natural antisense transcripts and formation of cancer networks. With regards to chromatin remodelling lncRNAs can function as signal lncRNAs or act as scaffolds and they also contribute towards stabilization of chromatin loops [33-35]. A key role of the chromatin loops is to facilitate interaction of distally-located enhancers with their respective target gene promoters [36,37].

BIOINFORMATICS ANALYSIS OF LNCRNA: FAR REACHING IMPLICATIONS FOR CANCER DIAGNOSTICS AND THERAPEUTICS

Given the extensive role of lncRNA in cancer growth, progression and proliferation as discussed in the previous section, comprehensive expression profiling of these RNA molecules can contribute immensely towards cancer diagnostics and therapeutics. Phenomenal development in high-throughput sequencing technologies and bioinformatics data analysis paradigms in the recent times has created possibilities for deeply insightful lncRNA expression profiling. Bioinformatics analysis of non-coding RNA data can also facilitate discovery as well as functional characterization of previously unknown lncRNA, profiling of inherent expression pattern and structural characterization of annotated lncRNAs. Expression profiling approaches such as microarray and high-

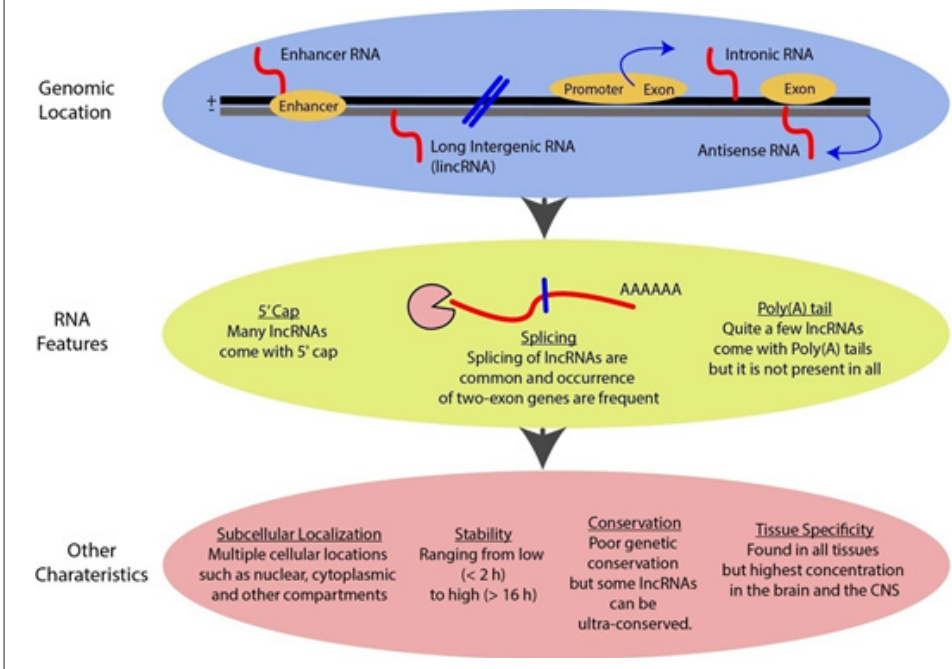

Figure 1: Molecular characteristics of lncRNAs.

throughput sequencing techniques such as Serial Analysis of Gene Expression (SAGE), Cap Analysis Gene Expression (CAGE) and RNA-seq are routinely used for characterization of lncRNA and bioinformatics analysis of the data generated involve steps such as 1. Detection of differential gene expression; 2. Gene expression based clustering; 3. Classification and, 4. Pathway analysis.

BIOINFORMATICS ANALYSIS OF RNA-SEQ DATA FROM EXPRESSION PROFILING OF LNCRNAS

Among different expression profiling techniques, RNA-seq is the method of choice for characterization of new lncRNA molecules because of high resolution at single nucleotide level and nominal sequencing costs. The working principle of RNA-seq is based on the conversion of RNA into cDNA, sequence fragmentation and finally high throughput sequencing of the fragments using techniques such as Illumina HiSeq. The high sequencing resolution of RNA-seq is very ideal for lncRNA expression profiling because in order to identify and characterize novel lncRNA molecules involved in cancer biology a very high sequencing depth of at least 150 million reads is necessary [38]. A study by Yu et al. successfully used the RNA-seq technique to detect nine different lncRNA markers to predict the different stages of oesophageal squamous cell carcinoma and present a very accurate prognosis for affected patients [39].

The first stage in the bioinformatics data analysis workflow for RNA-seq data involves pre-processing of the raw data to eliminate poor quality reads and reference mapping of the good reads with existing databases using tools such as Blat, SHRiMP and LastZ [40]. The next stage in the bioinformatics analysis workflow is assembly of the transcripts using tools such as BowTie [41] and mapping of the assembled transcripts with major lncRNA databases such as CPAT and Pfamscan. LncRNA data analysis pipelines such as lncRNAscan have been successfully used to detect novel lncRNAs involved in cancer and present highly accurate predictions of cancer therapeutics [36]. Functional characterization of detected lncRNAs can be performed using the existing lncRNA databases such as lnRNAdb that lead to the identification of lncRNA involved in malignancies such as gastric cancer [42]. Similarly, other lncRNA databases such as NONCODE lead to the identification and characterization of lncRNAs that are involved in hepatocellular carcinomas [43].

THE SIGNIFICANCE OF LNCRNA DATABASES

For efficient bioinformatics lncRNA analysis a number of database resources are available today. Notable mentions include databases such as ENCODE, FANTOM, and TCGA that are results of significant RNA-seq experiments on a wide range of tissue samples including samples from cancer patients [43-46]. While these databases contain a very good collection of lncRNAs from actual sequencing experiments, there are other database resources such as lncRNome and LncRNA Disease that consist of lncRNA lists derived from extensive data mining of scientific literature as well as scientific predictions [47,48]. While lncRNome may be considered as a general repository of characterized lncRNA molecules, the LncRNA disease database presents an excellent overview on the disease associations of different lncRNA molecules. Apart from the databases mentioned above, key bioinformatics tools have also been developed to access databases such as lncRNAtor which is a rich repository of information on gene-lncRNA co expression [49].

The bioinformatics tools can also be used to carry out molecular characterization of lncRNAs with regards to detection of similar functional motifs and structure prediction using resources such as LNCipedia and lncRNome.

CONCLUSION

While there is considerable therapeutic and research significance of lncRNAs, researchers are still not entirely aware of their underlying molecular functions. In order to gain further insights into the underlying molecular functions of long non-coding RNAs more studies are required that could lead to unravelling of novel functions of the molecule. Important characteristics of lncRNAs such as their ability to influence the process of alternative splicing through the regulation of processes such as phosphorylation and distribution of serine/arginine splicing factors have been characterized. It is very likely that through further bioinformatics approaches different lncRNAs molecules will be discovered that could cause the regulation of processes such as alternative splicing through different molecular mechanism. This review article did not focus on new mechanisms and types of long non-coding RNAs such as circular RNAs and a subject of future research could be discovery ad characterization of novel lncRNAs that could through significant insights into cancer pathways.

REFERENCES

1. Gibbs WW. The unseen genome: Gems among the junk. Sci Am. 2003;289(5):46-53.
2. Willingham AT, Gingeras TR. TUF love for junk DNA. Cell. 2006;125(7):1215-1220.
3. de Hoon M, Shin JW, Carninci P. Paradigm shifts in genomics through the FANTOM projects. Mamm Genome. 2015;26:391-402.
4. Khorkova O, Hsiao J, Wahlestedt C. Basic biology and therapeutic implications of lncRNA. Adv Drug Deliv Rev. 2015;87:15-24.
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. science. 2001;291:1304-1351.
6. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: Mechanisms and biological implications. Trends Genet. 2014;30(10):439-52.
7. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Molecular cell. 2011;43(6):904-914.
8. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, et al. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. Elife. 2014;3:e03528.
9. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014;3:e03523.
10. Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell. 2015;160(4):595-606.
11. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. Gene Biol. 2015;16(1):179.
12. Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G, et al. sORFs. org: A repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2016;44:324-329.

13. Marjaneh MM, Beesley J, O'Mara TA, Mukhopadhyay P, Koufariotis LT, Kazakoff S, et al. Non-coding RNAs underlie genetic predisposition to breast cancer. Genome Biol. 2020;21(1):1-4.

14. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L, et al. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. Rna. 2010;16(8):1478-1487.

15. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Gene Dev. 2011;25(18):1915-1927.

16. Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 2013;73(3):1180-1189.

17. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464:1071-1076.

18. Liu D, Xu B, Chen S, Yang Y, Zhang X, Liu J, et al. Long non-coding RNAs and prostate cancer. J Nanosci Nanotechnol. 2013;13(5):3186-3194.

19. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, et al. lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. Nucleic acids research. 2015 Jan 28;43(D1):D168-73.

20. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature Gene. 2015;47(3):199.

21. Ilott NE, Ponting CP. Predicting long non-coding RNAs using RNA sequencing. Methods. 2013;63(1):50-59.

22. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Rev Gene. 2009;10(1):57-63.

23. Fritah S, Niclou SP, Azuaje F. Databases for lncRNAs: A comparative evaluation of emerging tools. Rna. 2014;20(11):1655-1665.

24. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799.

25. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309(5740):1559-1563.

26. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635-640.

27. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005;308(5725):1149-1154.

28. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L, et al. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. Rna. 2010;16(8):1478-1487.

29. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Gene Res. 2006;16(1):11-19.

30. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Gene Res. 2012;22(9):1775-1789.

31. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008;322(5909):1851-1854.

32. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, et al. Genome-wide analysis of long noncoding RNA stability. Gene Res. 2012;22(5):885-898.

33. Lam MT, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci. 2014;39(4):170-182.

34. Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, et al. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol cell. 2014;53(6):1005-1019.

35. Ørom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. Cell. 2013;154(6):1190-1193.

36. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109-113.

37. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148:84-98.

38. Yu J, Wu X, Huang K, Zhu M, Zhang X, Zhang Y, et al. Bioinformatics identification of lncRNA biomarkers associated with the progression of esophageal squamous cell carcinoma. Mol Med Rep. 2019;19(6):5309-5320.

39. Liao P, Li S, Cui X, Zheng Y. A comprehensive review of web-based resources of non-coding RNAs for plant science research. Int J Biol Sci. 2018;14(8):819.

40. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Met. 2012;9(4):357.

41. Song H, Sun W, Ye G, Ding X, Liu Z, Zhang S, et al. Long non-coding RNA expression profile in human gastric cancer and its clinical significances. J Trans Med. 2013;11(1):1-10.

42. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016;26(3):304-319.

43. Yang L, Froberg JE, Lee JT. Long noncoding RNAs: Fresh perspectives into the RNA world. Trends Biochem Sci. 2014;39:35-43

44. Consortium F. A promoter-level mammalian expression atlas. Nature. 2014;507:462-470

45. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. Nature Gene. 2013;45(10):1113.

46. Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, et al. lncRNome: A comprehensive knowledgebase of human long noncoding RNAs. Database. 2013.

47. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: A database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2012;41:983-986.

48. Park C, Yu N, Choi I, Kim W, Lee S. lncRNAtor: A comprehensive resource for functional investigation of long non-coding RNAs. Bioinform. 2014;30(17):2480-2485.

49. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, et al. LNCipedia: A database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res. 2013 Jan 1;41(D1):D246-51.