FEBS Letters 587 (2013) 976–982

journal homepage: www.FEBSLetters.org

Systematic analysis of genomic organization and structure of long non-coding RNAs in the human genome

Jie Sun [a,1], Meng Zhou [a,*,1], Zhi-Tao Mao [a], Da-Peng Hao [a], Zhen-zhen Wang [a], Chuan-Xing Li [a,b,*]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China
[b] Institute for Systems Biology, Seattle 98109, USA

ARTICLE INFO

Article history:
Received 9 January 2013
Revised 17 February 2013
Accepted 19 February 2013
Available online 27 February 2013

Edited by Takashi Gojobori

Keywords:
Fragile site
Genomic organization
Long non-coding RNA
Repetitive element
Segmental duplication

ABSTRACT

The genomic architecture of several functional elements in animals and plants, such as microRNAs and tRNA, has been better characterized. As yet, there is very little known about genomic organization and structure of lncRNA in animals and plants. Here, we conducted a genome-wide systematic computational analysis of genomic architecture of lncRNAs, and further provided a more comprehensive comparative view of genomic organization between lncRNAs and several other functional elements in the human genome. Our study not only provides comprehensive knowledge for further studies into the correlations between the genomic architecture of lncRNAs and their important functional roles in diverse cellular processes and in disease, but also will be valuable for understanding the origin and evolution of lncRNAs.

1. Introduction

lncRNAs are a recently discovered class of ncRNAs which are in general defined as all non-coding transcripts longer than 200 nucleotides in length [1]. Since the discovery of the functional lncRNA *Xist* in the early 1990s, an increasing number of lncRNAs has been identified and predicted through high-throughput experimental technologies or computational approaches. However, only a very small percentage of these lncRNAs has been functionally and mechanistically characterized. The biogenesis, evolution and biological function of most lncRNAs still remain unknown. Accumulative evidences from the studies concentrated on only a very limited number of lncRNAs have demonstrated that lncRNAs play critical functional roles in diverse biological processes such as chromatin modification, transcriptional and post-transcriptional regulation, genomic imprinting, nuclear-cytoplasmic trafficking and so on [2–5]. In addition, large-scale lncRNA expression profiling analyses across human tissue and cancer types have revealed highly aberrant lncRNA expression in human cancers [6].

The genomic architecture of several functional elements in animals and plants, such as microRNAs (miRNAs) and tRNA, has been better characterized. As yet, there is very little known about genomic organization and structure of lncRNA in animals and plants. As an initial analysis toward exploring the genomic organization and structure of lncRNAs, we focus not only on characterizing the genomic distribution of lncRNAs in the human genome, but also on elucidating the association between lncRNAs and genomic fragile sites, and between lncRNAs and repetitive elements or segmental duplication events. To gain further insight into the genomic distribution features of lncRNAs, the parallel analysis was performed for the other four types of RNA. The systematic analysis of the genomic organization and structure of human lncRNAs not only provides comprehensive knowledge for further studies into the correlations between the genomic architecture of lncRNAs and their important functional roles in diverse cellular processes and in disease, but also will be valuable for understanding the origin and evolution of lncRNAs.

2. Materials and methods

2.1. Data sets

Human lncRNAs included in our analysis were manually curated from the existing literatures [7–9] and publicly available lncRNA database [10], and were mapped onto the human genome build GRCh37 (UCSC hg 19) using the BLAT program on the UCSC Genome Bioinformatics website [11]. The genomic coordinates of protein-coding genes with protein product, tRNAs and snoRNAs

* Corresponding authors. Address: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China. Fax: +86 045151802696 (M. Zhou).
E-mail addresses: biofomeng@hotmail.com (M. Zhou), lichuanxing@gmail.com (C.-X. Li).
[1] These authors contributed equally to this work.

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.02.036

were obtained from the "knownGene" table of the hg19 genome database in the UCSC genome browser [11] and HUGO Gene Nomenclature Committee (HGNC) [12]. All the human miRNAs and their genomic information were extracted from miRBase 18.0 [13]. Finally, all the data sets included in our study are comprised of 5964 lncRNAs, 1523 miRNAs, 402 snoRNAs, 622 tRNAs and 18654 protein-coding RNAs (Table S1).

2.2. Analysis of clustering patterns of lncRNA genes

To investigate the distance distribution of lncRNAs in the human genome, we computed the distances between two adjacent lncRNAs in the same chromosomes, and used the cumulative distribution function to analyze genomic distances of consecutive lncRNA pairs. The statistical significance was evaluated, using a combination of randomization analysis and Kolmogorov–Smirnov test (KS test), to compare the real genomic organization of lncRNAs with random sampling as follows: First, we selected random positions from a uniform distribution whose number is equal to the number of lncRNAs on each chromosome. Second, we computed the distances between two adjacent random points in one random sampling analysis. The empirical *P*-value was defined as:

$$P = \#\{i|P_{KS}(i) \geqslant 0.001\}/N$$

where P_{KS} (i) is *P*-value obtained from KS test in distance distribution between lncRNA pairs and adjacent random points in the *i*th randomized configurations. *N* is the number of random sampling. Here, we performed 1000 random sampling analysis.

Next, we further performed a randomization analysis to determine the statistical significance of clustering pattern of lncRNAs as follows: First, the average distance of lncRNA pairs was calculated across all chromosomes and defined as $\bar{x}$. Second, we selected random positions from a uniform distribution whose number is equal to the number of lncRNAs on each chromosome. Finally, we computed the distances between two adjacent random points and the average distance. The empirical *P*-value was defined as:

$$P = \#\{i|AVE(i) \geqslant \bar{x}\}/N$$

where *AVE*(i) is the average distance between two adjacent random points in the *i*th randomized configurations. *N* is the number of random sampling. Here, we performed 1000 random sampling analysis.

2.3. Analysis of incidence of lncRNA in fragile and non-fragile regions

The set of fragile sites was retrieved from HGNC database and are comprised of 117 fragile sites. We defined fragile regions and non-fragile regions by scanning genomic sequential positioning of fragile sites according to previous study [14]. The fragile region is defined as the sequential chromosomal bands associated with fragile sites, and non-fragile region is defined as sequential chromosomal bands, which are not known to be associated with fragile sites. Finally, we divided the 22 chromosomes (except for chromosome 21 and Y) into 87 fragile regions and 101 non-fragile regions.

Then we used Poisson Regression Models (PRM) to determine whether lncRNAs and other different types of RNA appear more frequently in fragile regions. This model is often used to model the number or rate of occurrences of an event of interest. In our analysis, "events" are defined as the number of lncRNAs or other RNAs in each fragile or non-fragile region, and the length of a region of interest is considered as exposure time. Because the distribution density of lncRNAs or other RNAs across different chromosomes varies considerably, this model also accounted for the effect of the differing densities of different RNA on each chromosome. Taking into account different distribution characteristics of lncRNAs and other RNAs in fragile and non-fragile regions, the standard PRM is performed for lncRNAs and protein-coding RNAs, and the zero-inflated Poisson (ZIP) model is carried out for miRNAs, snoRNAs and tRNAs, because count data with zero value is common in many fragile and non-fragile regions for miRNAs, snoRNAs or tRNAs. The incidence rate ratio (IRR), two-sided 95% confidence intervals of the IRR and two-sided *P* values are presented by regression analysis. An IRR significantly >1 reveals a marked increase in the number of lncRNAs or other RNAs in a region of interest. The regression analysis is performed by using STATA v12.0 software. The minimum free energy (MFE) was calculated using the RNAfold program [15].

2.4. Genome-wide identification of repeats-related lncRNA

The genomic positions of repeats and segmental duplications were obtained from UCSC Genome Browser [16]. We compared the genomic coordinates of lncRNA and repeats to identity all lncRNAs overlapping with repetitive elements. If lncRNAs tend to overlap fully repetitive elements, the lncRNA was considered to be a repeats-related lncRNA (RrlncRNA). The other lncRNAs with no or partially overlapping repetitive elements is defined as non-repeats-related lncRNA (NRrlncRNA).

3. Results and discussion

3.1. Genome-wide analysis of clustering patterns of lncRNA genes

Many studies have suggested that miRNAs are non-randomly distributed across chromosomes, and tend to be organized as clusters within several kilobases in the human genome [17–19]. The clustering propensity of human miRNAs and their target genes has been comprehensively characterized in several early studies [19–21]. Bermudez-Santana et al. examined genomic distribution of tRNAs in 74 eukaryotic genomes in a recent study, and found 17–36% of tRNAs are located in clusters in higher primates [22]. However, there is very little known about genomic arrangement and clustering pattern of lncRNA genes on a genome-wide scale. In order to gain novel insight into the genomic organization of human lncRNAs, we surveyed the genomic distribution of lncRNA genes in the human genome, and computed the neighbor distances between two adjacent lncRNAs on the same strand and the average distance. The statistical analysis revealed a statistically significant difference in distance distribution between lncRNAs with 1000 random sampling (*P*-value < 0.001). By comparing the real average distance of lncRNA pairs with 1000 random sampling, we found that the distances of lncRNA pairs are statistically smaller than expected at random (*P*-value < 0.001). These results suggested that lncRNAs are not uniform randomly distributed throughout the human genome. We further computed the neighbor distances for protein-coding RNAs and other non-coding RNAs as described above, and performed a comparative analysis between them. The distance distribution of lncRNAs and other four types of RNA is present in Fig. 1. Comparative analysis revealed significant differences in clustering pattern between lncRNAs and other four types of RNA (Kruskal–Wallis chi-squared = 8862.05, df = 4, *P*-value = 0). As shown in Fig. 1, a small percentage of miRNA pairs and protein-coding RNA pairs are separated by the pairwise distance of ≈10 kb, while more snoRNAs and tRNAs are distributed with a pairwise distance of ≈10 kb, and reveal a highly clustering propensity which is consistent with previous study [19]. However, only a very small percentage of lncRNA pairs are separated by the very small genomic distance (⩽1 kb). The fraction of lncRNA pairs still stays below 5%, even though the pairwise distance was extended to 10 kb. Further analysis showed that the fractions of lncRNA pairs are lower than those of the other four types of RNA at a pairwise distance of ⩽100 kb. Furthermore, most of lncRNA pairs are found