Do enhancer-associated lincRNAs play a role in nuclear architecture ?

Cyril Matthey-Doret

December 5, 2016

Introduction

A large proportion of the mammalian transcriptome does not code for proteins and to date, the number of known noncoding genes is more than 3 times that of protein-coding genes [Iyer et al., 2015]. Among noncoding RNAs, long noncoding RNAs (> 200bp) that do not overlap protein-coding genes are the most abundant (long intergenic noncoding RNAs, lincRNAs). Functional and evolutionary analyses, together with extensive characterization of a handful of lincRNAs, support the general consensus that these transcripts can regulate gene product abundance transcriptionally and post-transcriptionally, and that they can contribute to organismal traits and diseases [Kornienko et al., 2013]. However, the mechanisms of function for the majority of lincRNAs remain unknown [Rinn and Chang, 2012].

It is thought that spatial organization of the genome is an important component of these mechanisms [Engreitz et al., 2016]. Indeed, contrary to traditional view, genomic DNA is not linear, but is folded into variably compact chromosomal structures that likely impact expression of the embedded genes [Gorkin et al., 2014]. On a global scale, regions with a high degree of compaction are classified as heterochromatin while uncondensed regions are called euchromatin [Passarge, 1979]. These are respectively associated with lower and higher expression levels [Tamaru, 2010]. Chromosomes can be further divided into smaller domains where frequent DNA-DNA interactions occur as a result of their close proximity within the cellular nuclear space. These are called topologically associated domains (TADs). These domains are largely conserved across cell lines and they frequently contain smaller loop structures that promote contacts between different genetic regulatory elements, such as enhancers and promoters [Rao et al., 2014]. Such chromatin loops are often found at TAD boundaries [Rao et al., 2014]. They are also enriched in architectural proteins, such as CTCF and cohesin [Pope et al., 2014]. Both proteins are thought to function in the delimitation between neighbouring TADs by acting as genomic insulators that prevent DNA-DNA interactions across multiple domains. In addition to modulating regulatory promoter-enhancer interactions, TAD boundaries are often gene-dense and are enriched in highly transcribed genes [Ong and Corces, 2014].

Chromosomal contacts within TADs, and particularly at TAD boundaries, are crucial for establishing correct regulatory interactions between enhancers and promoters. Deletion of TAD boundaries often disrupts those interactions, resulting in gene misexpression and disease phenotypes [Lupiáñez et al., 2016]. Recently, some lincRNAs such as Firre, were reported to promote intrachromosomal interactions, either by forming promoter-enhancer looping, or by regulating chromatin structural organization [Engreitz et al., 2016]. Furthermore, lincRNAs associated with human traits were shown to have enhancer-associated cis-regulatory roles and their loci are correlated with a higher density of DNA contact within TADs (Tan et al, 2016, under revision). This raises the question whether lincRNAs with enhancer-like activities (elincRNAs) contribute to gene regulation and the organization of the dynamic chromosomal structure of the nucleus. Those elincRNAs will therefore be the focus of my analysis.

Unlike most enhancer-associated RNAs, which are often transcribed bidirectionally and then rapidly degraded [Darrow and Chadwick, 2013], elincRNAs are transcribed preferentially in one direction [Marques et al., 2013]and are

elincRNA	Overlap enhancers only	348
plincRNA	Overlap promoters only	456
other lincRNA	Overlap neither enhancers nor promoters	1443

likely good candidates for studying the involvement of lincRNAs in the regulation of DNA-DNA contacts. There are already some characterized elincRNAs, such as HOTTIP that have been shown to act as a link between chromosomal interactions and transcription. My analysis shows elincRNAs are associated with high density of chromosomal contacts within TADs and are significantly enriched in cohesin and CTCF binding sites, supporting the idea that they may contribute to gene regulation by modulating chromosomal organization.

Using various bioinformatics tools to analyze publicly available multi-omics data from the ENCODE project, I investigated the molecular properties of these elincRNAs, their enrichment in different regulatory elements and their association with the amount of DNA-DNA interactions to examine their role in regulating TAD organization.

Results

LincRNAs were divided into 3 categories based on overlap with enhancer and promoter elements (Table 1). See material and methods for more details on the overlap procedure. The expression of elincRNAs should be driven by enhancers since neither the region upstream of the genes nor their body contains any promoter. In this study, I focus on the elincRNAs, as it was shown that trait-relevant lincRNAs tend to be associated with enhancer activity (Tan et al, 2016, under revision).

Expression levels

The expression of elincRNA is significantly lower than that of PCG and other lincRNAs (Figure 1). These results concur with previous observations in mice [Marques et al., 2013]and are consistent across different human cell lines when keeping the same set of genes.

Tissue specificity

Conservation

Enrichment at TAD boundaries

There was no direct signal for enrichment of elincRNAs at TAD boundaries, however the promoter regions of elincRNAs were found to be enriched at anchors, which are themselves known to be enriched at boundaries. The lack of direct signal for enrichment at TAD boundaries may be a consequence of the method used to defined them. Notably, I restricted boundaries to the inside of TADs, therefore genes that are close to a TAD, but outside the border will not be counted.

Architectural proteins binding sites

CTCF and cohesin are often called architectural or insulator proteins, as they are thought to prevent TADs from interacting with each other. Enrichment tests for the binding sites of these proteins in the different categories of lincRNAs revealed that, although both enplincRNAs and neplincRNAs are enriched for binding sites of CTCF, this enrichment is not as strong in enplincRNAs as in neplincRNAs. Cohesin binding sites, on the other hand, were strongly enriched in both nep and enplincRNAs. Interestingly, RAD21 and SMC3, which are 2 subunits of cohesin, followed different patterns. SMC3 showed a stronger enrichment in neplincRNAs, while RAD21 was more enriched in enplincRNAs.

DNA-DNA contacts

Figures and tables

Discussion

Material and methods

Unless stated otherwise, all statistical tests were performed using R 3.3.1 [**?**]. Overlapping of genomic elements were done using either bedtools 2.26 [**?**] or the "intervals" package [**?**]in R. Manipulations on Hi-C contact matrices were performed using the "Matrix" package [**?**].

Genes

LincRNAs and protein-coding genes used were retrieved from the ENCODE website. The list of genes used in all analyses corresponds to genes expressed in the GM12878 lymphoblastoid cell line. Subcategories of genes were defined based on overlap between their body and the 1kb region upstream of their transcription start site, and regulatory elements available on ENCODE. These regulatory elements have been predicted computationally from Chip-seq data by a hidden Markov-model. Only predicted active promoters where considered when using promoters, and all enhancers when considering enhancers.

TAD definition

The list of TADs used in the computations is based on that from Rao et al (2014). They called the TADs based on Hi-C data across different human cell lines normalized and processed with their own protocol. Here, all the large TADs that completely encompassed smaller ones were removed to preserve the signal from the boundaries of the small TADs.

TAD boundaries definition

Instead of defining TAD boundaries based on arbitrary thresholds, TADs were split into 10 bins of 10% their length. This threshold was chosen based on previous findings showing an increase in transcriptional activity at 10% from the TAD border (Histogram from summary 3).

Conservation and tissue specificity

The sequence conservation was previously calculated (Tan Yihong, J., unpublished) using phastCons scores [**?**]. Tissue specificity index (Tau) was computed following the described in [**?**], considering only genes with expression above a 0.1 RPKM cutoff.

Expression levels

Median expression levels were computed from ENCODE data.

DNA-DNA contacts

Contacts were calculated using Hi-C contact matrices from Rao et al (2014). All computations are performed on 5kb resolution matrices with a MAPQ score of at least 30. The matrices were normalized using the KR normalization vector provided by the authors, when analyzing only GM12878. When comparing between different cell lines, SQRTVC (square root vanilla coverage) was used for chromosome 9 of all cell lines, because the KR algorithm did not converge for chromosome 9 of K562. For each gene overlapping a TAD, the mean contact inside the respective TAD was used as a measure. For single genes that overlap several TADs, the contacts are computed for each TAD independently. The mean contact in a TAD is computed by taking the arithmetic mean in a square submatrix spanning from the beginning to the end of the TAD in the intrachromosomal matrix.

Chip-seq

Chip-seq data for CTCF, RAD21 and SMC3 in GM12878 were retrieved from the ENCODE website. All three datasets were produced in the Snyder lab, at Stanford University.

Enrichment of genetic elements

All enrichment tests were performed using the genome association tester (GAT) [**?**]version 1.2. All tests for enrichment of lincRNAs were performed using the intergenic space of the genome as a workspace. When testing for enrichment of protein-coding genes, the protein coding space of the genome was used as the workspace. Similarly, when looking for enrichment. For all tests, the number of samples was set to 10,000 and the number of buckets was consequently adjusted to 270,000.

Bibliography

[Darrow and Chadwick, 2013] Darrow, E. M. and Chadwick, B. P. (2013). Boosting transcription by transcription: enhancer-associated transcripts. *Chromosome Research*, 21(6):713–724.

[Engreitz et al., 2016] Engreitz, J. M., Ollikainen, N., and Guttman, M. (2016). Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nature Reviews Molecular Cell Biology*, 17(12):756–770.

[Gorkin et al., 2014] Gorkin, D. U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, 14(6):771–775.

[Iyer et al., 2015] Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Presner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F., Iyer, H. K., and Chinnnaiyan, A. M. (2015). The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nat Genet.*, 47(3):199–208.

[Kornienko et al., 2013] Kornienko, A. E., Guenzl, P. M., Barlow, D. P., and Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription. *BMC biology*, 11(1):59.

[Lupiáñez et al., 2016] Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*, 32(4):225–237.

[Marques et al., 2013] Marques, A. C., Hughes, J., Graham, B., Kowalczyk, M. S., Higgs, D. R., and Ponting, C. P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biology*, 14(11):R131.

[Ong and Corces, 2014] Ong, C.-t. and Corces, V. G. (2014). CTCF : an architectural protein bridging genome topology and function. *Nature Publishing Group*, 15(4):234–246.

[Passarge, 1979] Passarge, E. (1979). Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. *American journal of human genetics*, 31(2):106–15.

[Pope et al., 2014] Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos,

J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., and Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405.

[Rao et al., 2014] Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

[Rinn and Chang, 2012] Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81:145–166.

[Tamaru, 2010] Tamaru, H. (2010). Confining euchromatin/heterochromatin territory: Jumonji crosses the line. *Genes and Development*, 24(14):1465–1478.