UNIL | Université de Lausanne

Do enhancer-associated lincRNAs contribute to chromosomal organization ?

First Step Project
Molecular Life Sciences, Bioinformatics

Cyril Matthey-Doret
Supervised by: Jennifer Yihong Tan
Directed by: Ana Claudia Marques

Department of Computational Biology
Department of Physiology
University of Lausanne - Switzerland

December 16, 2016

Abstract

Introduction

It was only recently discovered that a surprisingly large proportion of the mammalian transcriptome does not code for proteins. To date, the number of annotated noncoding genes longer than 200 nucleotides (long noncoding RNA, lncRNA) excess by at least 3 times that of protein-coding genes [Iyer et al., 2015]. Among lncRNAs, those that do not overlap with 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, demonstrate that these transcripts are involved in gene regulation processes transcriptionally and post-transcriptionally, and that they can contribute to organismal traits and diseases [Kornienko et al., 2013]. However, the mechanisms and functions, if any, for the majority of lincRNAs remain unknown [Rinn and Chang, 2012].

LincRNAs associated with human traits have been shown to have enhancer-associated cis-regulatory roles and their loci are correlated with more compact chromatin relative to other lincRNAs in a human lymphoblastoid cell line (LCL) (Tan et al, under revision). Most active enhancers are transcribed, generating noncoding products, including lincRNAs [Guil and Esteller, 2012]. This raises the question whether lincRNAs with enhancer-like activities (elincRNAs) contribute to gene regulation and the organization of chromosomal contacts. Unlike most enhancer-associated noncoding RNAs, which are often transcribed bidirectionally and then rapidly degraded [Darrow and Chadwick, 2013], elincRNAs are transcribed preferentially in one direction and are more stable [Marques et al., 2013]. These distinct features make them less likely to be a product of pervasive transcription and thereby, good candidates to study the involvement of lincRNAs in the regulation of gene-enhancer interactions within chromatin domains. Those elincRNAs will therefore be the focus of my analysis.

Recently, there have been reports of elincRNAs involved in the spatial organization of the genome, such as Haunt [Yin et al., 2015], which can regulate intrachromosomal interactions by mediating promoter-enhancer looping. It is thought that the architecture of the genome is an major factor in gene regulation [Engreitz et al., 2016]. Indeed, genomic DNA 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 relatively uncondensed regions are called euchromatin [Passarge, 1979]. These are respectively associated with lower and higher levels of active transcription [Tamaru, 2010]. Chromosomes are further compartmentalized into smaller domains, called topologically associated domains (TADs). The amount of DNA-DNA interactions is high within TADs as a result of their close spatial proximity, and low across different TADs. TAD boundaries are the regions lying at the borders of TADs (Figure 9) and are essential for gene regulation. They are often gene-dense and are enriched in highly transcribed genes [Ong and Corces, 2014].

Chromosomal contacts within TADs, often seen as looping structures, occur particularly at TAD boundaries and are crucial for establishing correct interactions between regulatory elements, such as enhancers and promoters [Gorkin et al., 2014]. Deletion of TAD boundaries often disrupts those interactions, resulting in gene misexpression and disease phenotypes [Lupiáñez et al., 2016]. TAD boundaries are also enriched in architectural proteins, including CTCF [Pope et al., 2014], which functions to delimit TAD borders by acting as genomic insulators that prevent DNA-DNA

interactions across multiple TADs. Cohesin, another architectural protein, is also enriched at TAD boundaries. It is a multi-protein complex that is thought to be involved in establishing enhancer-promoter interactions [Ji et al., 2016]. While most CTCF sites are shared between different cell types and species [Ji et al., 2016], cohesin binding at gene regulatory elements is often cell-type specific [Hadjur et al., 2009].

Using various bioinformatics tools to analyze publicly available multi-omics data from the ENCODE project [ENCODE Project et al., 2012] and data from whole-genome chromosome conformation capture (Hi-C) experiments [Rao et al., 2014], I investigated the molecular properties of elincRNAs, their enrichment in different regulatory regions and their association with the amount of DNA-DNA interactions to gain insight into their roles in gene regulation within topological domains in human LCLs. My analysis shows that elincRNAs are associated with high density of chromosomal contacts within TADs and are significantly enriched in loop anchors where promoter-enhancer interactions occur. I also find that they are strongly enriched in cohesin binding, supporting the idea that they may contribute to gene regulation by establishing contacts between regulatory elements and modulating chromosomal organization.

Results

LincRNAs were divided into categories based on overlap at their putative promoter regions (estimated as the region 1kb upstream of their transcriptional start site) with GM12878 LCLs enhancers and promoters predicted by the ENCODE consortium [ENCODE Project et al., 2012]. As I am interested in elincRNAs, I compared lincRNAs overlapping only enhancers in their putative promoter regions ($n = 236$) with other lincRNAs, whose promoter region overlaps neither promoters nor enhancers ($n = 1756$).

elincRNAs show similar expression levels to other lincRNAs

Enhancer-associated RNAs (eRNAs) are often found at relatively low transcript abundances as they tend to be rapidly degraded by the nuclear exosome [Lam et al., 2014]. To investigate if that is the case for LCL-expressed elincRNAs, I compared their expression levels in LCLs to that of other lincRNAs and protein-coding genes **EXPR**. I found a lower yet non-statistically significant difference between elincRNAs and other lincRNAs median expression levels (two-tailed Mann-Whitney U test, $p = 0.258$) in GM12878. This similarity of expression between elincRNAs and other lincRNAs may be linked to the distinct features of lincRNAs compared to eRNAs. Notably, they are unidirectionally transcribed and often polyadenylated, whereas eRNAs comprise many short, non-polyadenylated unstable transcripts that are transcribed bidirectionally [Darrow and Chadwick, 2013].

elincRNAs show lower conservation than other lincRNAs

To gain insights into elincRNAs evolution, I investigated the nucleotide conservation of their exons in primates and placental mammals. I found that elincRNAs are less conserved than other LCL-expressed lincRNAs as well as protein coding genes **PHASTCONS**. These differences are observed both when looking at conservation in mammals and in primates.

elincRNAs promoter regions are enriched at loop anchors but not TAD boundaries

ElincRNAs being associated with increased expression of neighboring protein-coding genes (Tan et al., under revision) and given the increasing number of lincRNAs associated with roles in chromosomal organization, it would be relevant to investigate if elincRNAs are often found at loop anchors and TAD boundaries, where important regulatory interactions take places. Therefore, I tested whether elincRNAs promoter regions are found significantly more often at loop anchors and TAD boundaries, than what would be expected if they were distributed randomly in the intergenic spaces of the genome.

Although elincRNAs promoter regions are enriched at loop anchors, where interactions between enhancer and promoter elements are known to occur [Ji et al., 2016], relative to other LCL-expressed lincRNAs , they are not significantly enriched at TAD boundaries **ENRICH-BOUNDARIES**. Despite the absence of elincRNA enrichment at TAD boundaries, dividing TADs into 10 equally sized bins **10BINS** reveals that elincRNAs tend to be more frequently found near the end of the TADs and are depleted at the center of the TADs (bin 5, 0.37 fold, $q = 0.06$) relative to other LCL-expressed lincRNAs. The trend is consistent with their enrichment at loop anchors, which are enriched at TAD boundaries (1.74 fold, $q < 0.001$, supplementary files).

The lack of significant enrichment of elincRNAs at TAD boundaries may be a consequence of the method used to define boundaries and the poor resolution of the current Hi-C technology. Notably, boundaries are extended from the TAD borders to the inside of TADs (see methods for details), therefore genes that are close to a TAD, but outside the border are not detected.

elincRNAs are enriched in cohesin

CTCF and cohesin are often called architectural or insulator proteins, as they are thought to prevent TADs from interacting with each other while increasing interactions within TADs. Enrichment tests for binding of these proteins in elincRNAs revealed that CTCF, SMC3 and RAD21 binding peaks were all highly enriched in elincRNA promoter regions (5.2-7.1 fold, $q < 0.001$) but only slighly enriched in other lincRNAs (1.1-1.3 fold, $q = 0.04 - 0.30$) **PEAK-ENRICH**.

Most binding peaks for CTCF and cohesin overlap in the genome **VENNOVER** and loops that are mediated by CTCF, cohesin or both CTCF and cohesin are thought to have different roles [Ji et al., 2016]. To determine if the enrichment of CTCF peaks is a consequence of the overlap with cohesin peaks, or if elincRNAs could be specifically involved in cohesin loops, I performed enrichment tests for CTCF- and cohesin-exclusive binding **ENRICH-EXCLU**. The difference in fold enrichment for cohesin peaks in elincRNAs compared to other lincRNAs much stronger when looking at exclusive binding sites, while it decreased for CTCF. This suggests elincRNAs specifically may be involved in the formation of cohesin-only loops. According to a recent model [Ji et al., 2016], loops mediated by CTCF or CTCF and cohesin have insulator properties and are important for the structural maintenance of boundaries and the formation of insulated neighbourhood within TADs, while cohesin only loops mediate promoter-enhancers interactions. My results may point towards a role of elincRNAs in the formation of promoter-enhancer loops.

elincRNAs are associated with high DNA-DNA contacts in TADs

As I found elincRNAs to be enriched at loop anchors and in cohesin binding, to further support their role in promoting promoter-enhancer contacts, I investigated whether elincRNAs are associated

with regions of higher DNA-DNA contact. To measure this, I used the average amount of contact in their respective TAD as a proxy (see material and methods for details). I find that elincRNAs are associated with TADs presenting higher amounts of contacts than other lincRNAs **HIGH-CONTACT** in GM12878 (two-tailed Mann-Whitney U test, $p < 0.001$), with a 1.24 fold increase in median contacts. When comparing the contacts for the same sets of genes in 3 other cell lines, these results are consistent, but less significant in HUVEC and K562 (two-tailed Mann-Whitney U test, $p < 0.05$) with respectives fold increases in median contacts of 1.05 and 1.07 for elincRNAs compared to other lincRNAs. The amount of contacts was not significantly higher for elincRNAs in NHEK (two-tailed Mann-Whitney U test, $p = 0.472$) although the trend was still going in the same direction with a fold increase in median contacts of 1.04. The strength of this association seems to be very cell line-dependent, but always pointing towards a higher amount of contacts for elincRNAs. Although these results do not give any insights into the mechanisms through which elincRNAs may promote contact, high DNA-DNA contacts, together with the enrichment of cohesin suggest a role for elincRNAs in the establishment of promoter-enhancer looping inside TADs.

Figures, tables and legends

Discussion

Materials and methods

Unless stated otherwise, all statistical tests were performed using R 3.3.1 [R Core Team, 2016]. Overlapping of genomic elements were done using either bedtools 2.26 [Quinlan and Hall, 2010]or the intervals package [Bourgon, 2015] in R. Manipulations on Hi-C contact matrices were performed using the Matrix package [Bates and Maechler, 2016].

Genes

LincRNAs and protein-coding genes sets 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 promoter region, defined as the 1kb region upstream of the transcription start site and regulatory elements available on ENCODE [ENCODE Project et al., 2012]. These regulatory elements are 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. The 2 categories of lincRNAs that are used throughout this report are elincRNAs, defined as overlapping enhancers but no promoters in their promoter region, and other lincRNAs defined as overlapping neither enhancer nor promoters in their promoter region.

TAD definition

The list of TADs used in the computations is based on that from Rao et al [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 encompass smaller ones were

removed to preserve the signal from the boundaries of the small TADs. Boundaries from very large TADs would otherwise contain the signal from smaller TADs inside, generating noise.

Hi-C data and normalization

Contacts were calculated using Hi-C contact matrices from Rao et al [Rao et al., 2014]. All computations are performed on 5kb resolution matrices constructed from all read pairs mapping to the genome with a MAPQ score of at least 30. The matrices were normalized using the KR normalization vector provided by the authors whenever possible. 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 probably as a result of the high sparsity of the matrix. I chose SQRTVC as a substition for KR as the authors reported this method to yield very close results to KR. The normalization procedure consists in dividing each entry in the contact matrix M by a corresponding value in the normalization vector V:

$M^*_{i,j} = \frac{M_{i,j}}{V_{KR}[\frac{i}{res}] * V_{KR}[\frac{j}{res}]}$

Where $M_{i,j}$ is an entry from the raw matrix and $M^*_{i,j}$ corresponding normalized entry.

TAD boundaries definition

Boundaries are extended from TAD borders towards the interior of TADs using a custom algorithm (Figure 10). The method used to define boundaries relies on the assumption that boundaries are insulated regions. In other words, there are few interactions between elements before and after the boundaries. The insulation is measured by sliding a diamond on every position along the matrix diagonal and computing the sum in the diamond at each position. Lower values represent more insulated regions. The size of the diamond has been set to an arbitrary threshold of 100kb, considered reasonable as the median length of filtered TADs is 140kb. More formally, the algorithm can be described as sliding a diamond of width w along the diagonal of a square matrix M of n dimensions on all positions d between w and n-(w-1). Those latter limits are set to prevent the diamond from getting out of the matrix. At each position, the sum of all values in the diamond is stored in a vector V. This can be rewritten as:

$\begin{cases} 1 \leq w \leq \frac{n}{2} + 1 \\ \forall d \in \{w, ..., n - (w-1)\} \end{cases} V_d = \sum_{i=d-(w-1)}^{d} \sum_{j=d}^{d+(w-1)} M_{i,j}$

The sums from the diamond are then used to compute boundaries. For all TADs, boundaries are extended inwards from the borders as long as the value of V does not exceed a threshold defined as the starting value (at the border) plus 10% of the maximum value in the TAD (Figure 11).

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 in exons (Tan et al, under revision) through mammalian and primate evolution using phastCons scores [Siepel et al., 2005] and averaged phastCons score were used as a measure of exonic sequence conservation. Tissue specificity index (Tau) was computed following the procedure described in [Kryuchkova and Robinson-Rechavi, 2015], considering only genes with expression above a cutoff of 0.1 RPKM.

Expression levels

Processed median expression data for elincRNAs and protein-coding genes in 4 different cell lines were calculated by Tan et al, under revision. The original data comes from ENCODE (ENCODE Project et al., 2012).

DNA-DNA contacts

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 of all values in a square submatrix spanning from the beginning to the end of the TAD in the intrachromosomal matrix (Figure 10, left).

Chip-seq

Chip-seq peaks for CTCF, RAD21 and SMC3 in GM12878 were retrieved from the ENCODE website (ENCODE Project et al., 2012). The CTCF and cohesin exclusive peaks were obtained by using the intersect and subtract tools from the bedtools suite.

Enrichment of genetic elements

All enrichment tests were performed using the genome association tester (GAT) (Heger et al., 2013) version 1.2. This program allows to test if genomic segments of interest are found in a desired set of annotations more often than expected if they were distributed randomly in a workspace. All tests using lincRNAs as annotations or segments were performed using the intergenic space of the genome as a workspace. When testing for enrichment of anchors at boundaries, the whole genome was used as the workspace. For all tests, the number of samples was set to 10,000, the number of buckets was consequently adjusted to 270,000 and segments overlap was used as the measure.

For all enrichment tests, two values are reported in the results section: Fold enrichment and q-value. The fold enrichment corresponds to the ratio of observed/expected number of segments in the given annotation, it is higher than 1 if the segment is enriched compared to random expectation, and lower than 1 if the segment is depleted. The q-value corresponds to the p-value once it has been corrected for multiple testing using false discovery rate calculated with the Benjamin-Hochberg procedure.

Acknowledgements

I wish to thank Jennifer Yihong Tan for her precious advices and help throughout the project. I also wish to thank Ana Claudia Marques for her suggestions, general guidance and critical reading of the report. Finally, I want to thank Adam Alexander Thil Smith for his help with technical issues and code optimization and Maria Ferreira da Silva for her suggestions. Computationally demanding analysis were performed at the Vital-IT Center for High Performance Computing of the Swiss Institute of Bioinformatics (www.vital-it.ch).

Bibliography

[Bates and Maechler, 2016] Bates, D. and Maechler, M. (2016). *Matrix: Sparse and Dense Matrix Classes and Methods*.

[Bourgon, 2015] Bourgon, R. (2015). *intervals: Tools for Working with Points and Intervals*.

[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.

[ENCODE Project et al., 2012] ENCODE Project, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

[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.

[Guil and Esteller, 2012] Guil, S. and Esteller, M. (2012). Cis-acting noncoding RNAs: friends and foes. *Nature Structural & Molecular Biology*, 19(11):1068–1075.

[Hadjur et al., 2009] Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–413.

[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.

[Ji et al., 2016] Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., Weintraub, A. S., Hnisz, D., Pegoraro, G., Lee, T. I., Misteli, T., Jaenisch, R., and Young, R. A. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell*, 18(2):262–275.

[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.

[Kryuchkova and Robinson-Rechavi, 2015] Kryuchkova, N. and Robinson-Rechavi, M. (2015). A benchmark of gene expression tissue-specificity metrics. *bioRxiv*, (January):027755.

[Lam et al., 2014] Lam, M. T. Y., Li, W., Rosenfeld, M. G., and Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends in Biochemical Sciences*, 39(4):170–182.

[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.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[R Core Team, 2016] R Core Team (2016). R: A Language and Environment for Statistical Computing.

[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.

[Siepel et al., 2005] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050.

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

[Yin et al., 2015] Yin, Y., Yan, P., Lu, J., Song, G., Zhu, Y., Li, Z., Zhao, Y., Shen, B., Huang, X., Zhu, H., Orkin, S. H., and Shen, X. (2015). Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. *Cell Stem Cell*, 16(5):504–516.