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 21, 2016

Abstract

A large proportion of the human genome generates RNA that is not translated into proteins, denoted as noncoding RNA. Long noncoding RNAs (lncRNAs), are longer than 200 nucleotides and represent the largest part of the noncoding transcriptome in humans, outnumbering protein-coding genes by almost 3 times. Although less than 1% of these transcripts have been functionally annotated, a wide array of regulatory functions have already been attributed to lncRNAs. In particular, it has recently been found that intergenic lncRNAs (lincRNAs) associated with human trait variants often arise from enhancer transcription and coindide with regions of high frequencies of chromosomal contacts. There are already some characterized enhancer-associated lincRNAs (elincRNAs) known to regulate these chromosomal contacts, but whether this is a general function of elincRNAs or these are just isolated cases is unknown. Modulation of chromosomal architecture, for example by establishing loops or highly compact regions called topologically associated domains (TADs), is crucial for regulation of gene expression, allowing correct interactions between enhancers and promoters. Here, I investigate the role of these elincRNAs in spatial organization of the genome in a human lymphoblastoid cell line, making use of functional genomics and data from recent advances in chromosome conformation capture techniques that allow to quantify DNA:DNA contacts genome wide (Hi-C).

Introduction

It was only discovered in the past decade 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) exceeds that of protein-coding genes by at least 3 times [Iyer et al., 2015]. Among lncRNAs, those that do not overlap protein-coding genes, referred to as long intergenic noncoding RNAs (lincRNAs), are the most abundant. Functional and evolutionary analyses, together with extensive characterization of a handful of lincRNAs, have demonstrated that some of these transcripts are involved in the regulation of gene expression programs, both 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].

A large proportion of lincRNAs originate from active enhancers, which are widely transcribed and generate noncoding products, broadly classified as enhancer-associated noncoding RNAs (eRNAs). These noncoding transcripts include lincRNAs which are referred to as enhancer-associated lincRNAs (elincRNAs) [Guil and Esteller, 2012].

Recently, a set of such lincRNAs, identified in human lymphoblastoid cell lines (LCLs), were associated with human trait variants have been shown to have enhancer-associated roles in *cis*-regulation of local gene expression. Importantly, higher frequencies of chromosomal interactions are often observed at these loci relative to other lincRNAs in a human lymphoblastoid cell line (LCL), suggesting that elincRNAs may be involved in gene regulation through modulating chromatin architecture (Tan et al, 2016, under revision).

Spatial organization of the genome impacts gene regulation [Engreitz et al., 2016]. Specifically, the folding of genomic DNA into variably compact chromosomal structures can strongly influence expression of the embedded genes [Gorkin et al., 2014]. Globally, regions with low degree of compaction, referred to as euchromatin, are associated with high levels of active transcription.

On the other hand, relatively uncondensed and less transcriptionally active regions are called heterochromatin [Passarge, 1979] [Tamaru, 2010]. Chromosomes are further compartmentalized into smaller domains, called topologically associating domains (TADs), where frequent DNA:DNA interactions occur as a result of their close spatial proximity. These domains are key in modulating gene transcriptional programs. Recent findings show that regions near TAD borders, called TAD boundaries, are essential for modulating gene regulatory interactions within TADs and preventing chromatin contacts across TADs. Deletion of TAD boundaries can disrupt those interactions, resulting in gene misexpression and disease phenotypes [Lupiáñez et al., 2015].

Detailed functional characterizations of a few elincRNAs have established the molecular mechanisms underlying their roles in the spatial organization of the genome. For example, Haunt is one such elincRNA [Yin et al., 2015], which regulates the expression of the HOXA gene cluster by modulating intrachromosomal interactions, specifically through establishing promoter-enhancer looping. These recent findings raise the question on what is the prevalence of elincRNAs contributing to gene regulation through modulation of chromosomal conformation.

Using various bioinformatics tools to analyze publicly available multi-omics data from human lymphoblastoid cell lines (LCLs). I investigated the molecular properties of elincRNAs. Specifically, I examined their enrichment in regions that are key in TAD regulation and their association with the amount of chromosomal interactions to gain initial insight into their roles in gene regulation within topological domains. My analyses show that elincRNAs are associated with high density of DNA:DNA contacts within TADs and are significantly enriched in protein binding sites important for TAD regulation. Importantly, elincRNAs are strongly enriched at chromosomal loop anchors where promoter-enhancer interactions occur. Together, my findings support the hypothesis that they may contribute to gene regulation by establishing contacts between gene regulatory elements and modulating chromosomal organization.

Results

Enhancer-associated lincRNAs (elincRNAs) in human lymphoblastoid cell lines (LCLs) were identified from a set of LCL-expressed lincRNAs (Tan et al, 2016, under revision). They were defined based on overlap between their putative promoter regions (estimated as 1kb upstream from their transcriptional start site) and enhancers predicted from histone marks in LCL (GM12878, [ENCODE Project et al., 2012]). LincRNAs whose promoter region also overlapped other predicted regulatory elements, specifically active promoters [ENCODE Project et al., 2012], in LCLs were excluded from the analysis (elincRNAs=236, other LCL-expressed lincRNAs=1756).

elincRNAs show similar expression profiles as other lincRNAs

Most enhancer-associated noncoding RNAs (eRNAs) are transcribed bidirectionally and then rapidly degraded by the nuclear exosome, which makes them hard to detect due to their low expression levels [Darrow and Chadwick, 2013] [Lam et al., 2014]. In contrast, elincRNAs are more stably and preferentially transcribed in a single direction, and are often spliced and polyadenylated [Marques et al., 2013]. First, to investigate if LCL-expressed elincRNAs share similar expression profiles relative to other expressed genes as the lowly expressed eRNAs, I compared their expression levels to that of other lincRNAs and protein-coding genes (Figure 1A). I found that elincRNAs are expressed at no more lower levels compared to other lincRNAs in human LCLs (GM12878, two-tailed

Mann-Whitney U test, $p = 0.258$). In addition, contrary to the highly tissue specific eRNAs [Wu et al., 2014], elincRNAs are no differently tissue-specific from other lincRNAs either, although their median tissue specificity index (τ) is slightly higher (two-tailed Mann-Whitney U test, $p = 0.38$, Figure 1B). These results suggest that elincRNAs may have distinct properties relative to other enhancer-associated transcripts.

elincRNA transcripts are less conserved than other lincRNAs

To gain insights into the molecular evolution of elincRNAs, I investigated the nucleotide conservation of their exons in primates and placental mammals using phastCons scores [Siepel et al., 2005], a measure of nucleotide conservation (Methods). I found that exons of elincRNAs are less conserved than other LCL-expressed lincRNAs (two-tailed Mann-Whitney U test, mammals: $p < 1e-07$, primates: $p < 1e-05$) as well as protein coding genes (two-tailed Mann-Whitney U test, mammals: $p < 1e-97$, primates: $p < 1e-86$) (Figure 1C).

Interestingly, a recent study of a set of trait-relevant and enhancer-associated lincRNAs showed that although exons of these lincRNAs did not seem to have evolved under purifying selection relative to other LCL-expressed lincRNAs across mammalian and primate evolution, their sequences are constraint specifically during recent human evolution (Tan et al, 2016, under revision). Therefore, although my result may suggest that elincRNA transcripts were not evolving under constraint across broad mammalian evolution, their conservation across modern human evolution remains to be investigated.

elincRNA promoter regions co-localize with loop anchors and cohesin binding sites

Next, to examine whether elincRNAs are associated with the regulation of chromosomal architecture, I investigated their co-localization with loop anchors, where enhancer-promoter gene regulatory interactions are frequent [Ji et al., 2016]. These regulatory regions are essential to establish chromosomal interactions within topologically associating domains (TADs). I compared the overlap of elincRNAs with these regions relative to what would be expected if these loci were randomly distributed across the intergenic regions of the genome (Methods). I found that elincRNAs are significantly enriched at LCL loop anchors (fold enrichment = 2.79, $q = 1e-04$, Figure 2A).

Architectural proteins, such as cohesin and CTCF, are also enriched at the loop anchors [Rao et al., 2014]. Specifically, the cohesin protein-complex is important for cell type-specific intra-TAD gene regulation [Hadjur et al., 2009]. The transcription factor CTCF is another central player in the regulation of chromatin architecture and gene expression. According to a recent model [Ji et al., 2016], loops mediated by CTCF only or CTCF and cohesin collectively are important for the structural maintenance of TADs. These loops act as insulators, preventing interactions between TADs. In contrast, loops containing only cohesin binding sites are crucial in mediating regulatory intra-chromosomal interactions, supported by the evidence that cohesin depletion is associated with disrupted promoter-enhancer interactions within TADs [Seitan et al., 2013].

I found that elincRNA promoter regions are significantly enriched in CTCF and cohesin binding sites in LCL (binding of SMC3 and RAD21, 2 cohesin subunits, were used to identify cohesin binding, Methods) (fold enrichment = 5.2 and 8.1, respectively, $q = 1e-04$) This enrichment is greater than the one found for other lincRNAs (fold enrichment = 1.3 and 1.3, respectively, $q < 0.05$, Figure 2B).

As a large proportion of binding sites for CTCF and cohesin overlap in the human genome (Figure 2C) and since loops mediated by CTCF and cohesin collectively are thought to have different roles in chromosomal organization from those mediated only by cohesin [Ji et al., 2016], I further determined the independent enrichment of elincRNAs in CTCF and cohesin binding, using mutually exclusive binding sites for these proteins. This revealed a greater enrichment of cohesin binding sites (Figure 2D) in elincRNA loci (fold enrichment = 13.1, $q = 1e-04$) compared to that of CTCF (fold enrichment = 3.92, $q = 1e-04$). This suggests elincRNAs are more frequently involved in the formation of cohesin-only loops, supporting their roles in modulating promoter-enhancer looping.

elincRNA promoter regions are not enriched at TAD boundaries

As loop anchors are frequently found at TAD boundaries, I next investigated whether similar enrichment would be observed for elincRNAs at TAD boundaries. Using Hi-C data (Figure 3) I defined TAD boundaries as regions extending inside TADs from the borders until reaching local intra-TAD contacts that exceeds a cut-off threshold (see Methods for details). Although elincRNAs promoter regions are enriched at loop anchors relative to other LCL-expressed lincRNAs, no significant enrichment was found for these loci at TAD boundaries (fold enrichment = 1.2, $q = 0.08$, Figure 4A). Despite the absence of significant elincRNA enrichment at TAD boundaries, dividing TADs into 10 equally sized bins reveals that elincRNAs tend to be more frequently found near the end of the TADs and are depleted in the center of the TADs (bin 5, fold enrichment = 0.37, $q = 0.06$) relative to other LCL-expressed lincRNAs (Figure 4B). The trend is consistent with their enrichment at loop anchors, which are enriched at TAD boundaries (fold enrichment =1.74 , $q < 1e-3$, not shown).

This lack of significant enrichment of elincRNAs at TAD boundaries may be a consequence of poor resolution of the current Hi-C technology, which is restrained to a maximum of 5kb [Rao et al., 2014], as well as limitations in the method used to define boundary regions. Particularly, I defined TAD boundaries by only extending inwards from TAD borders (see methods for details), therefore all genes located outside of defined TADs but close to a TAD border would be unaccounted for in the analysis.

elincRNA are associated with high DNA:DNA contacts within TADs

To further support their role in regulating promoter-enhancer contacts, I investigated whether elincRNAs are associated with regions with higher DNA:DNA interactions. To this end, I measured the average amount of contact within their respective TADs (Figure 5A, methods). I found that elincRNAs are frequently embedded within TADs with a higher mean density of contacts compared to other lincRNAs in GM12878 (Median fold difference=1.24, two-tailed Mann-Whitney U test, $p < 1e-04$, Figure 5B). In addition, the fold difference in the amount of DNA contacts within TADs that harbour these elincRNA loci relative to other LCL-expressed lincRNAs were less pronounced in 3 other cell lines (Median fold difference=1.05, 1.07, and 1.04, p=$0.02, 0.04, 0.472$ in HUVEC, K562 and NHEK, respectively). This suggests that the association between elincRNA expression and chromosomal contacts is likely cell line-dependent. Although my findings do not provide insights into the molecular mechanisms through which elincRNAs may regulate chromosomal architecture, their associated high DNA:DNA contacts, together with their enrichment for cohesin binding suggest a role for elincRNAs in the modulation of enhancer-promoter looping within topologically associating domains.

Figures, tables and legends

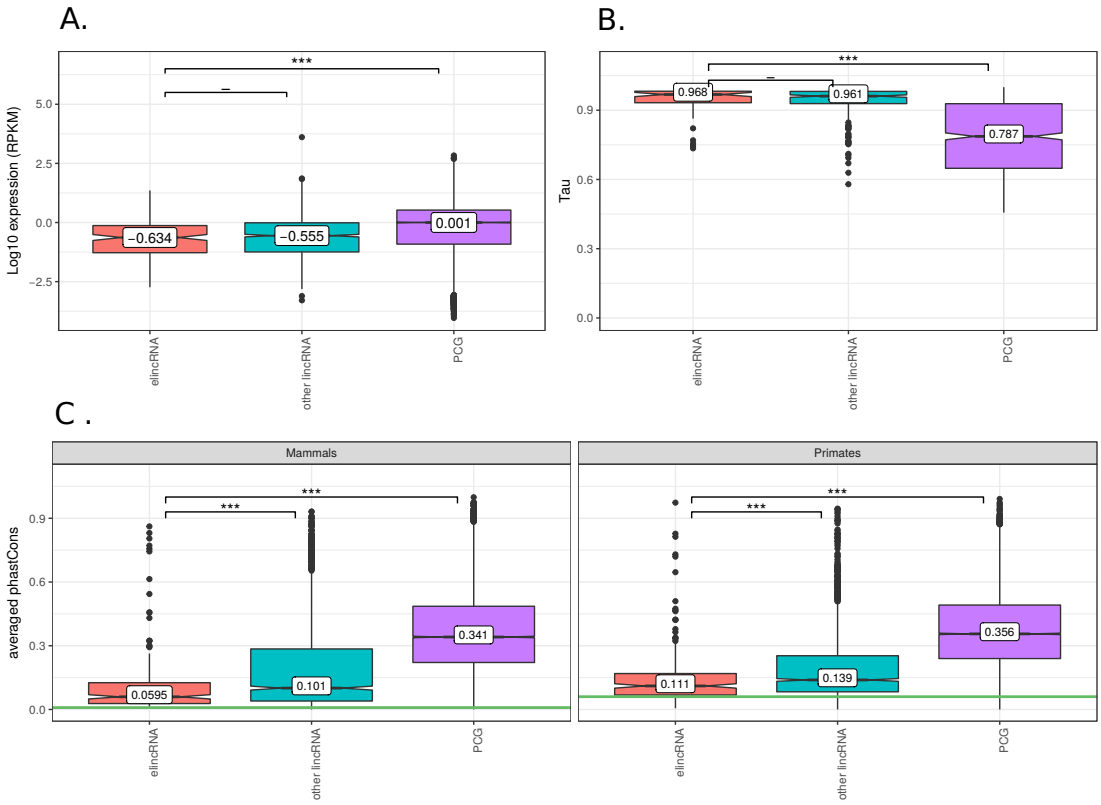

Figure 1: Properties of elincRNAs. Median values are displayed in the boxes. **A)** Distribution of median expression levels in LCL (GM12878), **B)** tissue specificity index (Tau) and **C)** Average exonic sequence conservation across mammalian and primate evolution of elincRNAs (orange), other LCL-expressed lincRNAs (blue) and protein-coding genes (purple). The tau index, a measure of gene expression tissue specificity, ranges from 0 (low specificity) to 1 (high specificity) [Kryuchkova and Robinson-Rechavi, 2015]. Averaged phastCons score is used as a measure for nucleotide conservation [Siepel et al., 2005]. The green horizontal line represents the median conservation of ancestral repeats, which are used as a proxy for neutral evolution. Differences between distributions were tested using a two tailed Mann-Whitney U test, $***P < 0.001$; - not significant

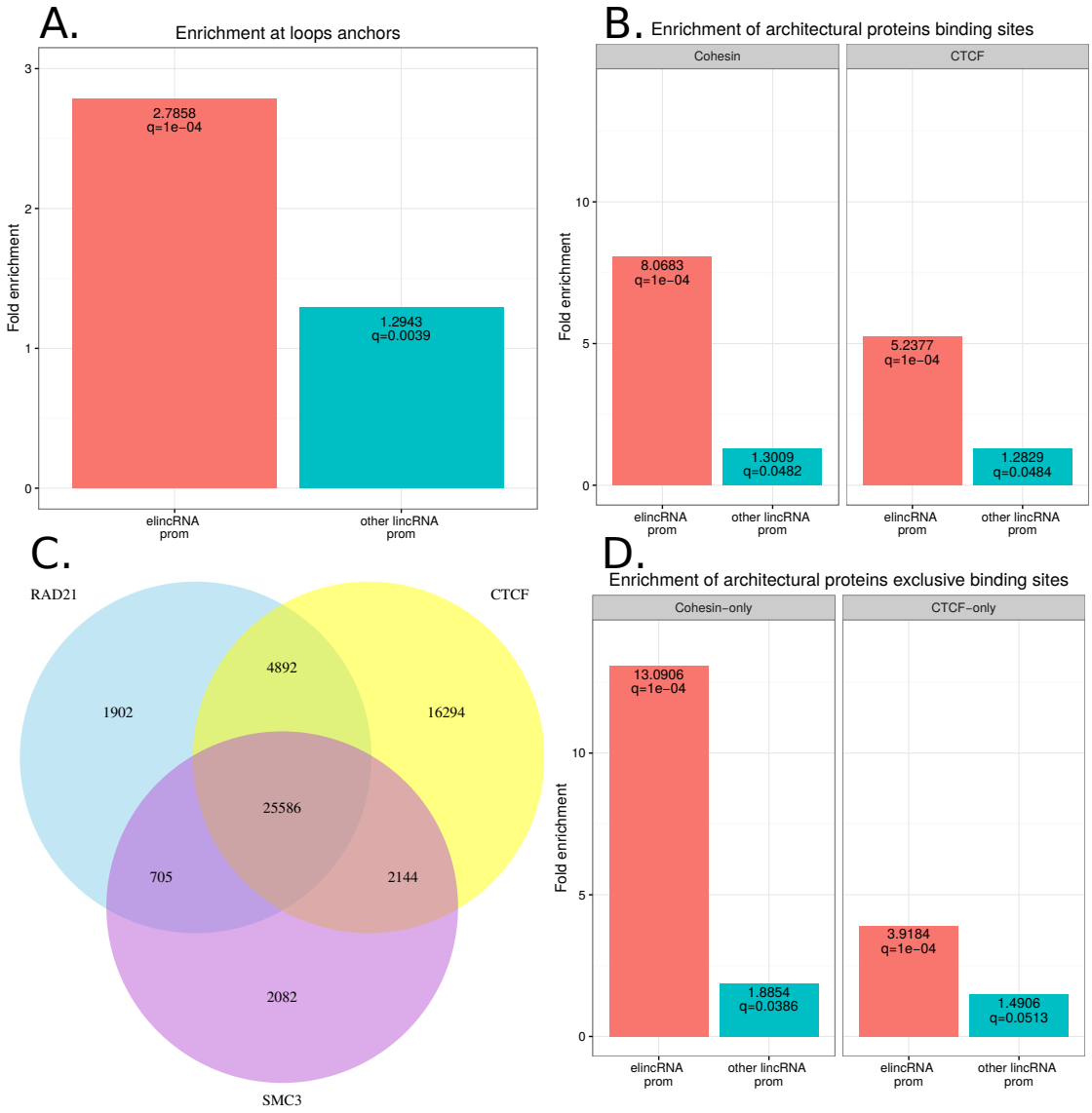

Figure 2: elincRNAs are enriched at loop anchors and cohesin binding sites. Fold enrichment relative to random expectation: **A.** elincRNA (orange) and other lincRNA (blue) promoter regions at loop anchors. **B.** all cohesin and CTCF peaks in human LCLs at elincRNA and other lincRNA promoter regions. **C.** Proportions of overlap between cohesin (RAD21 and SMC3) and CTCF peaks in human LCLs. **D.** cohesin- and CTCF-only peaks at elincRNA and other lincRNAs promoter regions.

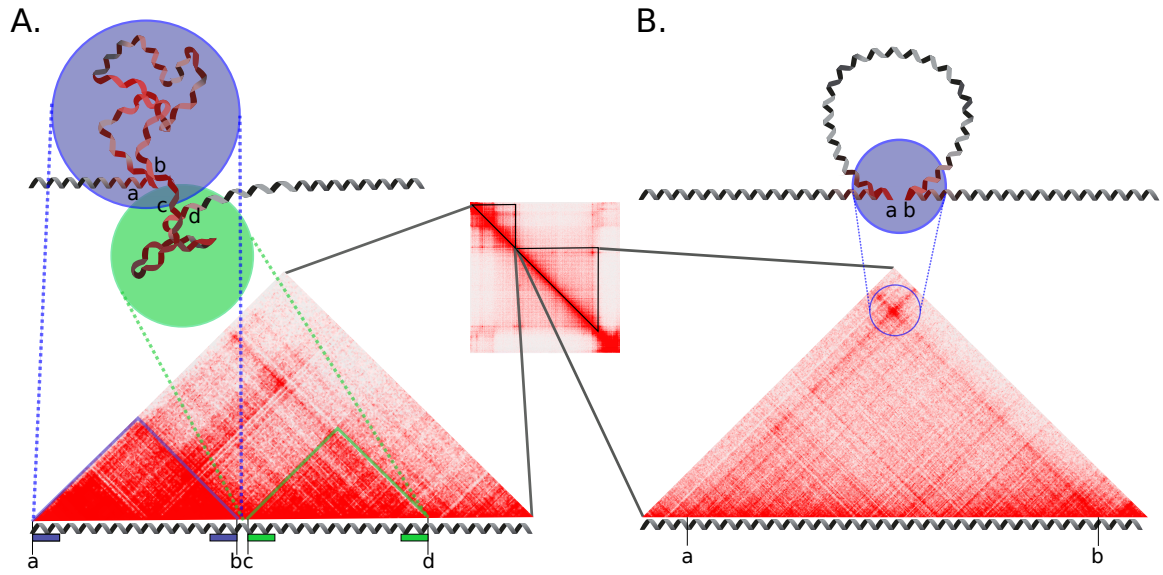

Figure 3: Schematic representation of topologically associating domains (TADs) and loops using Hi-C matrices. **A)** TADs and **B)** chromatin loop represented using a Hi-C intra-chromosomal contact matrix visualized as a symmetrical square matrix (center panel) using Juicebox (Durand et al., 2016) and as upper triangles, for simplified representation (side panels). Colour intensity of matrix pixels reflect the density of interactions. Left panel: Two TADs are illustrated as two triangular sub-matrices (triangles with blue and green outlines). Frequent interactions occur within these regions occur while interactions across TADs are less frequent. These two TADs correspond to the two globular chromatin structures within the blue and green circles. a and b denote the borders of the first TAD while c and d represent the borders of the second TAD. Boundaries are the rectangles expanding inwards from the borders. Right: Chromosomal loop with a and b representing the loop anchors. Unlike TADs, strong DNA:DNA contact is observed only at the contacting point between loop anchors, while less frequent contact occurs in the region between the two anchor points.

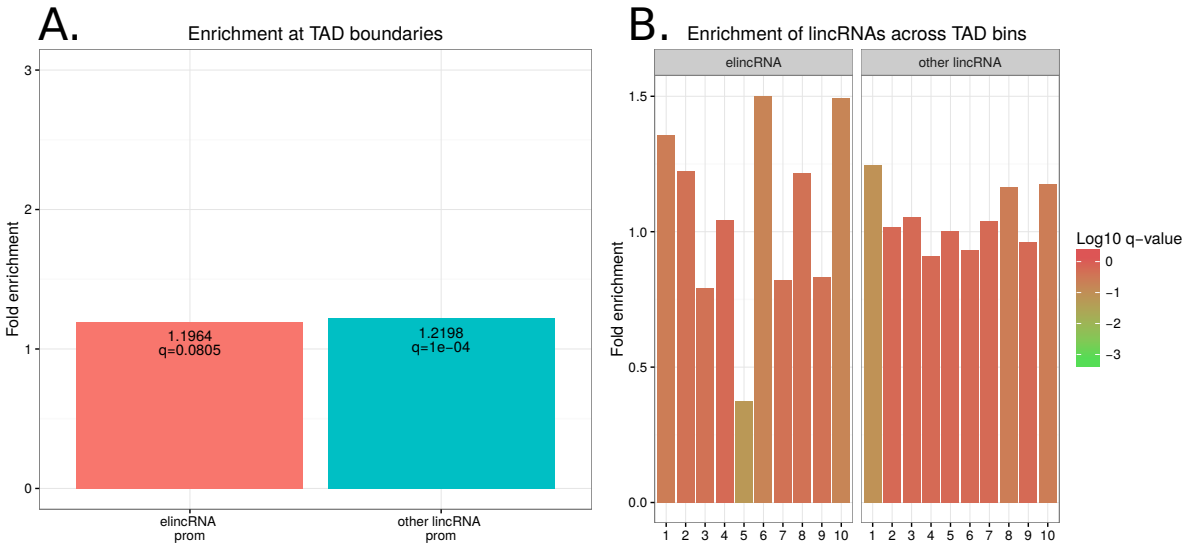

Figure 4: elincRNAs are not enriched at TAD boundaries. **A.** Enrichment of elincRNA (orange) and other LCL-expressed lincRNAs (blue) promoter regions at TAD boundaries. **B.** Enrichment of elincRNA and other LCL-expressed lincRNA promoter regions across TADs. Each bar represent a bin of 10% of TAD length. Colour of enrichment bars represent log10 of q-values.

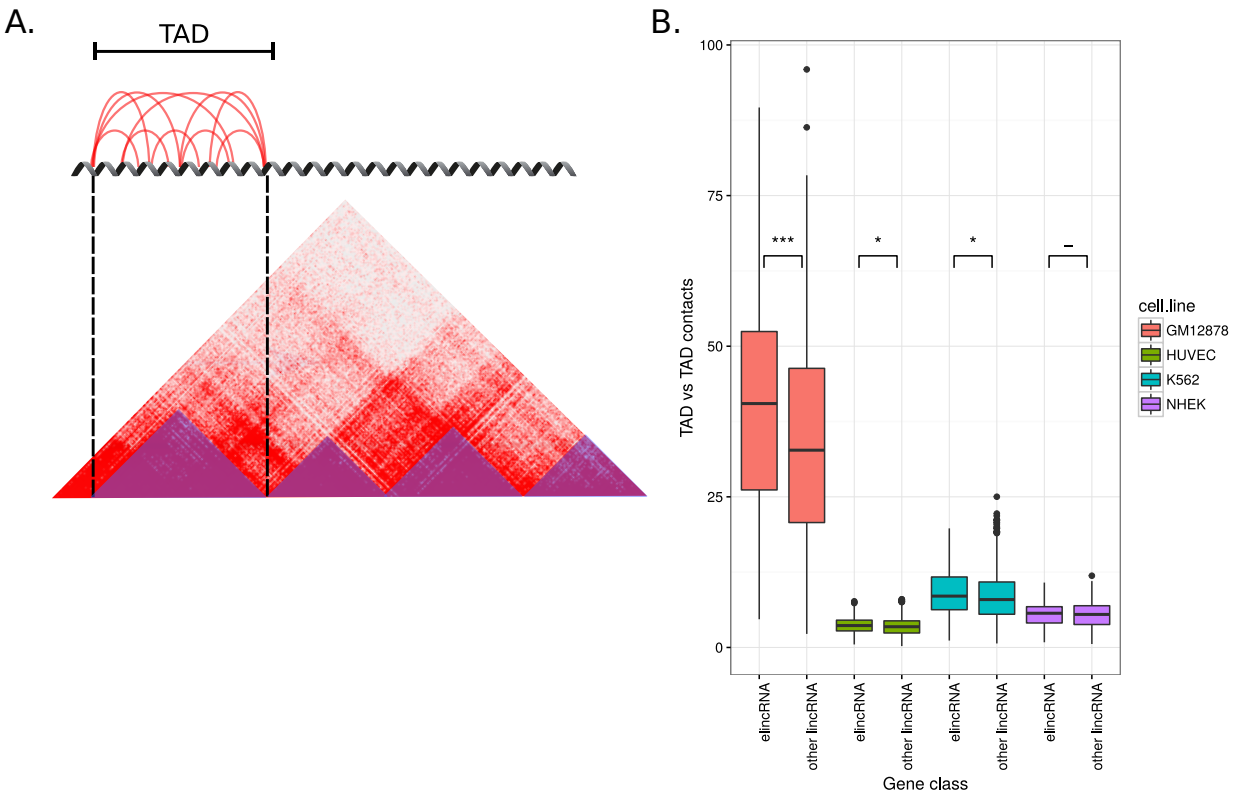

Figure 5: elincRNA-containing TADs are associated with significantly higher density of DNA:DNA contacts. **A.** Average of interactions within a TAD is computed by summing all DNA contacts within the TAD (blue triangles) and dividing by TAD length. **B.** Average contact frequency within TADs that encompass elincRNA and other LCL-expressed lincRNAs across four cell lines, GM12878 (orange), HUVEC(green), K562 (blue) and NHEK (purple). Set of genes as defined in GM12878 are used for all comparisons. Difference between groups are tested using the two tailed Mann-Whitney test, $***P < 0.001; *P < 0.05$; $-$ non-significant

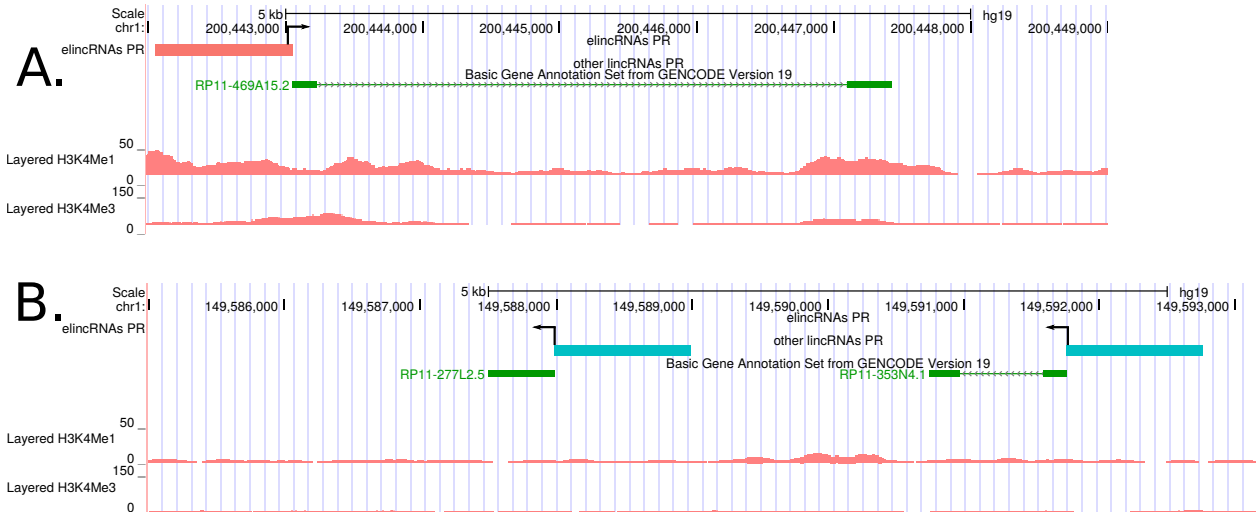

Figure 6: Genome browser view of elincRNA and other lincRNA genes. **A.** elincRNA with putative promoter region in pink **B.** two other lincRNAs with putative promoter region in blue. The promoter regions of the elincRNA has low levels of promoter (H3K4Me3) and high level of enhancer (H3K4Me1) histone marks, while the promoter region of the other lincRNAs have low levels of both. Gene bodies are depicted in green, with solid boxes as exons and thin lines as introns. black arrows indicate direction of transcription

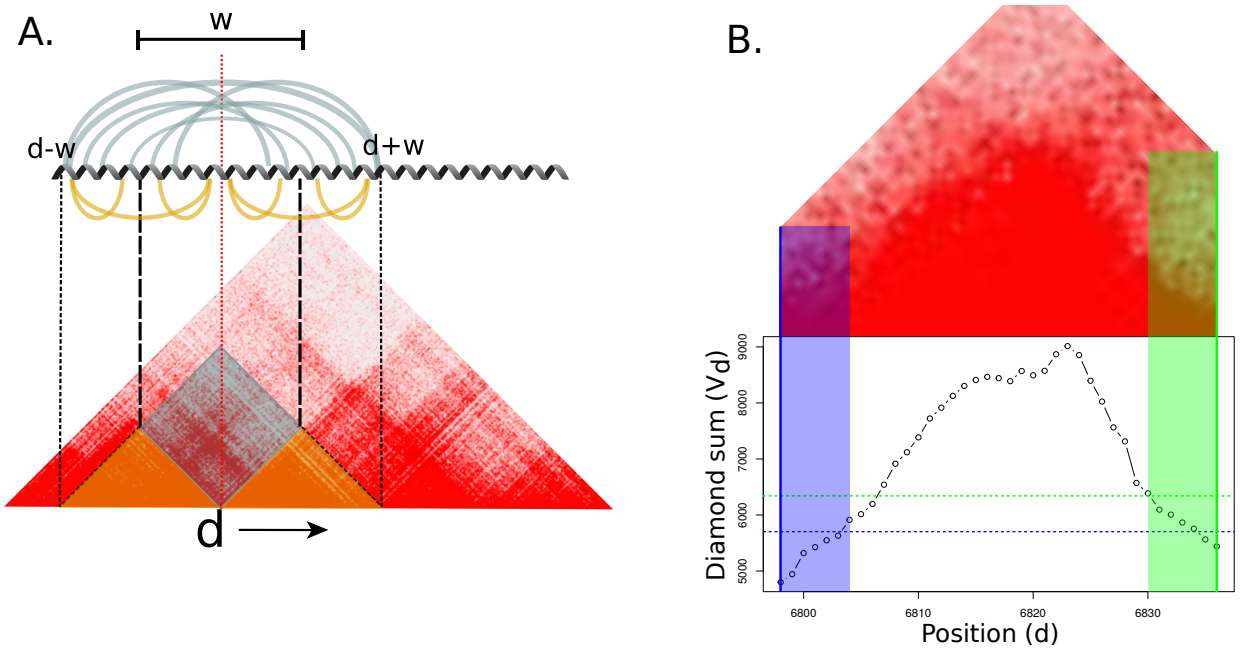

Figure 7: Visual representations of the algorithm used to compute TAD boundaries in Hi-C matrices. **A.** A diamond (blue) of width w set to 100kb is slid on all position (d) along the diagonal, representing a step-size of 5kb in the genome, given by the matrix resolution. For each position d, the sum of all pixels in the blue diamond (V_d) is computed and later used to define boundaries. The sum in the diamond at position d (red vertical dotted line) represents a measure of all interactions across position d (i.e. elements upstream of d contacting those downstream of d) at a maximum range of w (blue curves at the top). In other words, the diamond measures contacts between regions [d-w;d] (upstream) and [d;d+w] (downstream) (blue), excluding interactions happening inside those regions (orange). **B.** Example of the calculated sums of interactions in vector V through a TAD. V_d represents the sum of interactions across position d (i.e. the sum in the diamond at position d, stored as the d^{th} value of vector V), while d are the positions along the matrix diagonal (corresponding to 5kb segments of the genome). Boundaries were extended inwards from TAD borders until V_d reached an arbitrary threshold defined as the value of V at the border, plus 10% of the maximum value of V inside the TAD. The solid vertical lines represent the TAD borders, the horizontal dashed lines represent the thresholds required to stop extending boundaries and the transparent areas represent the final boundaries. All blue elements relate to the left side, while all green elements relate to the right side.

Discussion

Although it has been long recognized that chromosomal conformation has strong impact on the regulation of gene transcription programs, features that underlie the regulation of chromatin organization remain relatively unclear [Bonev and Cavalli, 2016]. Following the development of chromosome conformation capture techniques over a decade ago [Dekker, 2002], advances in other 3C-based technologies have allowed the quantification of chromosomal interaction frequencies between genomic loci in close spatial proximity, both locally (i.e. 3C and 4C) and genome-wide (i.e. Hi-C and ChIA-PET) [Dekker et al., 2013].

LincRNAs originating from enhancers (elincRNAs) have been shown to modulate chromosomal architecture [Yin et al., 2015] and are frequently located within topologically associated domains (TADs) associated with high levels of genomic interactions (Tan et al, 2016, under revision). Here, I used publicly available genomics data to investigate the prevalence of elincRNA regulation in modulating chromosomal architecture. Together with strong enrichment in cohesin binding at elincRNAs promoter regions, their frequent localization at loop anchors and high amount of intra-TAD contacts associated with elincRNAs support their putative role in regulating enhancer-promoter loops.

There is ongoing debate on the biological relevance of elincRNA transcription in particular with regards to whether their associated regulatory roles are transcript- or transcription-dependent. Notably, the association between elincRNAs and chromosomal organization does not constitute a proof for their role in promoting contact, as it might be a consequence of the presence of active enhancers at their promoter region, rather than the elincRNA itself. With this in mind and to address this key question, it would be important to perform an analysis similar to that presented here but where enhancers producing elincRNAs are compared to other active enhancers, that generally produce unstable bidirectionally-transcribed eRNAs. Ultimately, the dissection of the molecular mechanisms underlying enhancer-associated regulation of chromosomal conformation will require genetic manipulation of elincRNA candidate transcript abundance, for example by using RNAi and inhibiting their transcription with CRISPR-Cas9 [Li et al., 2013].

Furthermore, extending the analysis from within-topological domain interactions to between-TAD contacts [Fraser et al., 2015], to inter-chromosomal contacts, and eventually to DNA associations with the nuclear lamina would provide a more complete overview of elincRNA-dependent chromosomal interactions that shape the nuclear architecture. These analyses could also be extended to additional cell-lines to reveal if the effect of elincRNAs in nuclear architecture is cell-line specific, or more generalized.

The current resolution of global chromosomal conformation capture techniques (including Hi-C) remains a limiting factor in measuring accurate interaction frequencies as it is presently not possible to look at contacts at elincRNAs loci that are smaller than the highest resolution (i.e. 5 Kb) [Rao et al., 2014]. Instead, only an estimation of the region surrounding the gene can be measured. In addition, even using data with high contact resolutions, bulk Hi-C technique can only provide an average estimate of all chromosomal contacts happening in a cell population, thus masking the dynamics and variabilities between cells. Recently, a single-cell Hi-C protocol has been developed [Nagano et al., 2013], allowing one to examine DNA contacts profiles at a single cell resolution. Such single cell techniques, combined with a higher contact resolution, would provide greater power to shed light on the impact of elincRNAs on gene expression regulation through modulating chromosome architecture.

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]. All positions of genetic elements used are based on the hg19 human assembly.

elincRNA definition

LincRNAs and protein-coding genes sets were provided by Jennifer Tan (Tan et al, 2016, under revision). Generally, those sets were obtained by combining ENSEMBL annotations for the human genome (build 70) with a set of lincRNAs annotated *de novo* using LCL transcriptome data obtained from the ENCODE consortium [ENCODE Project et al., 2012]. The list of genes used in all analyses corresponds to genes expressed in the GM12878 human lymphoblastoid cell line. LincRNAs were classified based on overlap between their putative promoter region, defined as the 1kb region upstream from the transcription start site and predicted regulatory elements available on ENCODE [ENCODE Project et al., 2012]. These regulatory elements are predicted computationally from histone marks by a hidden Markov-model. Predicted active promoters and all enhancers were considered. The 2 categories of lincRNAs that are used throughout this report are elincRNAs, defined as overlapping predicted enhancers but no predicted active promoters in their putative promoter region (i.e. high H3K4Me1/H3K4Me3 ratio, Figure 6A), and other lincRNAs defined as overlapping neither enhancers nor active promoters in their promoter region (i.e. low H3K4Me1 and H3K4Me3 marks, Figure 6B).

Expression levels

Processed median expression data for elincRNAs and protein-coding genes in 4 different cell lines were calculated by Jennifer Tan (Tan et al, 2016, under revision) using RNA sequencing data generated by the ENCODE consortium [ENCODE Project et al., 2012]. Specifically, they used the number of reads overlapping exons of lincRNAs and protein-coding genes (GENCODE version 19) to estimate the expression level (RPKM) in each sample. Genes with 0 RPKM in at least half of the samples were not considered.

Conservation and tissue specificity

The sequence conservation was previously calculated in exons (Tan et al, 2016, 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 (τ) was computed following the procedure described in [Kryuchkova and Robinson-Rechavi, 2015], considering only genes with expression above a cutoff of 0.1 RPKM.

Chip-seq data

Chip-seq peaks for CTCF, RAD21 and SMC3 in GM12878 were obtained from the ENCODE consortium [ENCODE Project et al., 2012]. Cohesin peaks were defined as the union between the

peaks of RAD21 and SMC3 subunits. The CTCF and cohesin exclusive peaks were obtained by using the intersect and subtract tools from the bedtools suite, removing all peaks that are common to CTCF and cohesin.

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 regions of the human 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 to average expected number of segments overlapping 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 controlled with the Benjamin-Hochberg procedure.

TAD definition

The list of TADs used in the computations was obtained 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 algorithm. 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., 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. To correct for bias induced by chromatin accessibility and restriction sites density [Rao et al., 2014], 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, as described by the authors, consists in dividing each entry in the raw contact matrix M by corresponding values in the normalization vector N:

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

Where $M_{i,j}$ is an entry from the raw matrix and $M^*_{i,j}$ corresponding normalized entry and res is the resolution of the contact matrix (i.e. 5000, in this study).

TAD boundaries definition

Boundaries are extended from TAD borders towards the interior of TADs using a custom algorithm. The method used to define boundaries relies on the assumption that there are few interactions between elements located before and those after the boundaries. These interactions are measured by sliding a diamond (Figure 7A) on every position along the matrix diagonal and computing the sum in the diamond at each position. Boundaries should present lower values. 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. 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 $1 + w$ and $n - w$. 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:

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

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 an arbitrary threshold defined as the starting value (at the border) plus 10% of the maximum value in the TAD (Figure 7B).The boundaries were filtered afterwards to remove all those extending beyond their TAD. This happened in cases where the border was already among the highest values in the vector V. This shows the algorithm is not optimal and would need be improved to properly process the boundaries in these special cases. Lowering the threshold would reduce the number of boundaries that fail to stop inside their TAD, but it would also reduce the length of all boundaries. To accurately define boundaries, one would need a more complex algorithm that takes additional factors into account.

intra-TAD 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 contact matrix (Figure 7B).

Acknowledgements

I wish to thank Jennifer Yihong Tan for her precious advices and help throughout the project and writing of the report. I also wish to thank Ana Claudia Marques for her suggestions, general guidance and corrections in 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 and critical reading of the report. Computationally demanding analyses 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.

[Bonev and Cavalli, 2016] Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678.

[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.

[Dekker, 2002] Dekker, J. (2002). Capturing Chromosome Conformation. *Science*, 295(5558):1306–1311.

[Dekker et al., 2013] Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403.

[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.

[Fraser et al., 2015] Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., Dostie, J., Pombo, A., and Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, 11:1–14.

[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.

[Heger et al., 2013] Heger, A., Webber, C., Goodson, M., Ponting, C. P., and Lunter, G. (2013). GAT: A simulation framework for testing the association of genomic intervals. *Bioinformatics*, 29(16):2046–2048.

[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.

[Li et al., 2013] Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H.-S., Glass, C. K., and Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–20.

[Lupiáñez et al., 2015] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

[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.

[Nagano et al., 2013] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

[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.

[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.

[Seitan et al., 2013] Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A. G., Flicek, P., Dekker, J., and Merkenschlager, M. (2013). Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*, 23(12):2066–2077.

[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.

[Wu et al., 2014] Wu, H., Nord, A. S., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., Pennacchio, L. A., and Visel, A. (2014). Tissue-specific rna expression marks distant-acting developmental enhancers. *PLOS Genetics*, 10(9):1–12.

[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.