Do enhancer-associated long intergenic noncoding RNAs contribute to nuclear architecture ?

1. 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)⁠. Amongst all lncRNAs, those that do not overlap protein-coding genes are the most abundant, referred to as long intergenic noncoding RNAs (lincRNAs). 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, Guenzl, Barlow, & Pauler, 2013)⁠. However, the mechanisms that underlie functional roles, if any, of the majority of lincRNAs remain unknown (Rinn & Chang, 2012)⁠.

A large proportion of lincRNAs originate from active enhancers, which are widely transcribed and generating noncoding products, broadly classified as enhancer-association noncoding RNAs (eRNAs). These noncoding transcripts include lincRNAs and are referred to as enhancer-associated lincRNAs (elincRNAs) (Guil & Esteller, 2012). Recently, a set of such lincRNAs, identified in human lymphoblastoid cell lines (LCLs), are 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).

Indeed, spatial organization of the genome impacts gene regulation (Engreitz, Ollikainen, & Guttman, 2016)⁠⁠. Specifically, the folding of genomic DNA into variably compact chromosomal structures can strongly influence expression of the embedded genes (Gorkin, Leung, & Ren, 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 and 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 often disrupts 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.

2. Results

Enhancer-associated lincRNAs (elincRNAs) in human lymphoblastoid cell lines (LCLs) were determined from a set of LCL-expressed lincRNAs (Tan et al, 2016, under revision). They were defined based on overlap between their putative promoter region (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 often transcribed bidirectionally and then rapidly degraded by the nuclear exosome, thereby are often difficult to detect their low expression levels (Darrow & Chadwick, 2013)⁠ (Lam, Li, Rosenfeld, & Glass, 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 the expression levels of elincRNAs 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 (REF), elincRNAs are no differently tissue-specific from other lincRNAs either, although their median tissue specificity index is slightly higher, signifying marginally higher tissue specificity (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)⁠, as 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, recent studies on 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.

elincRNAs 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 regulatory regions that are essential in establishing chromosomal interactions within topologically associating domains (TADs). Specifically, significant enrichment of elincRNAs, compared to what would be expected if they were randomly distributed across the intergenic regions of the genome (Methods), was found at loop anchors (fold enrichment = 2.79, q=1e-04, Figure 2A), where enhancer-promoter gene regulatory interactions are high (Ji et al., 2016).

Previous studies have demonstrated that loop anchors within TADs are enriched for TAD-regulatory proteins, including the cohesion multi-protein complex and the CTCF transcription factor (Rao et al., 2014a)⁠. Specifically, cohesin is important for intra-TAD gene regulation in a cell type-specific manner (Hadjur et al., 2009)⁠, whereas CTCF bindings tend to be largely conserved across cell types and species (REF). According to a recent model (Ji et al., 2016)⁠, loops mediated by CTCF only or those by CTCF together with cohesin are important for the structural maintenance of TADs. These loops act as insulators and prevent interactions across TADs. In contrast, loops that contain cohesion only 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)⁠.

Significant enrichments were found for CTCF and cohesin binding peaks (subunits of cohesion, specifically SMC3 and RAD21, were used to measure cohesin binding, Methods) within elincRNAs promoter regions (fold enrichment=5.2 and 8.1, respectively, q=1e-04) relative to other lincRNAs (fold enrichment = 1.3 and 1.3, respectively, q<0.05, Figure 2C).

As a large proportion of binding sites for CTCF and cohesin overlap in the human genome (Figure 2D) and since loops mediated by CTCF and cohesin collectively are thought to have different roles in chromosomal organization from those mediated by only cohesin (Ji et al., 2016)⁠, I further determined the independent enrichment of elincRNAs in CTCF and cohesin binding using mutually exclusive binding sites of the proteins. This revealed a much stronger enrichment of cohesin binding sites (Figure 2E) 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 may be much more frequently involved in the formation of cohesin-only loops, supporting their roles in modulating promoter-enhancer looping.

ElincRNAs 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 obtained the sum of interactions within each TAD (REF, Rao et al). Next, I estimated TAD boundaries, defined as regions extending from each side of the previously predicted TAD borders (REF Rao et al) that contain no more than 10 percent of the maximum interactions within that TAD (Methods). Boundaries were extended from TAD borders until the sum of interactions in the diamond reached an arbitrary threshold (Methods). Although elincRNAs promoter regions are enriched at loop anchors relative to other LCL-expressed lincRNAs (Figure 2A), 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, they tend to be more frequently found near the peripherals and are significantly depleted in the center of the TADs (bin 5, fold enrichment = 0.37 fold, 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, supplementary files).

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., 2014a)⁠, 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 are close to a TAD border would be unaccounted for in the analysis.

elincRNAs 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 higher 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, 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) compared to GM12878. This provides evidence 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 at enhancer-promoter loop anchors suggest a role for elincRNAs in the modulation of chromosomal looping within topologically associating domains.

3. Figures and tables

Figure 1: Characteristics of elincRNAs. Median values are displayed in the boxes. All tests are performed using the two-tailed Mann-Whitney *U* test, \*\*\*P<0.001; – not significant. (A) Median expression levels, (B) Tissue specificity index (measured by the Tau index) 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) [REF]. Averaged phastCons score is used as a measure for nucleotide conservation (REF). Ancestral repeats are sequences expected to be evolving neutrally (green line).

Figure 2: elincRNAs are enriched at loop anchors and cohesin binding sites. Enrichment of promoter regions of elincRNA (orange) and other LCL-expressed lincRNAs (blue) at (A) loop anchors, (B) all cohesin and CTCF binding sites in human LCLs (GM12878). (C) Overlap between cohesin (RAD21 and SMC3) and CTCF binding sites in LCLs. (D) Enrichment of promoter regions of elincRNA (orange) and other LCL-expressed lincRNAs (blue) at cohesin- and CTCF-only binding sites. Values displayed on the bars represent gene set fold enrichment compared to random expectation. q represents the p-value adjusted for multiple testing using FDR.

Figure 3: Schematic representation of topologically associating domains and loops using Hi-C matrices. (A) TADs and (B) chromatin loops represented using Hi-C matrices visualized as symmetrical square matrices using Juicebox (Durand et al., 2016)⁠ (center panel) and as triangles, which is a simplified representation as the square matrices by removing duplicated contact information (side panels). Colour intensity of matrix pixels on the matrix signify density of chromosomal interactions. Left panel: Two TADs are illustrated as two triangular sub-matrices (triangles with blue and green borders) where frequent interactions within these regions occur and where interactions across the 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 often observed only at the contacting point between loop anchors, where less frequent contact occurs in the region between the two anchor points.

Figure 4. A. Enrichment of promoter regions of elincRNA (orange) and other LCL-expressed lincRNAs (blue) at (A) boundaries and (B) segments across TADs. Each bar represents 10% of TAD length. Colour of enrichment bars represent log10 of q-values.

5elincRNA-containing TADs are associated with significantly higher density of. (A) Average of interactions within a TAD is computed by summing all DNA contacts within a TAD (blue triangle) and normalized by TAD length. (B) Average contact frequency within TADs that encompass elincRNA and other LCL-expressed lincRNAs four, GM12878 (orange), HUVEC (green), K562 (blue), and NHEK (purple)Difference between groups are performed using the t*U*

Figure 6: Visual representations of the algorithms used to compute contacts in Hi-C matrices. As they are symmetric, matrices are represented as upper triangles for simplicity reasons. A: Schematic representation of the algorithm used to measure insulation. A diamond (blue) of width w set to 100kb is slid on all position along the diagonal. For each position, the sum in the diamond is computed and later used to define boundaries. The sum in the diamond at position d (dotted line) represents a measure of all interactions across position d (i.e. between elements before and after position d). Each TAD is taken as a submatrix (upper triangles of the submatrices are depicted in blue) and the mean value in the submatrix is computed.

Figure 5: Example of the calculated sums of interactions in vector V through a TAD. V value represents the value of interactions across each position in the TAD, while V index represent the position on the matrix diagonal, corresponding to a segment of 5kb in the genome. Boundaries were extended until they reach a threshold. The corresponding TAD in the Hi-C matrix is depicted above, visualized in Juicebox. The solid vertical lines represent the TAD borders, the horizontal dashed lines represent the threshold required to stop extending boundaries (methods) 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.

Figure 7: Genome browser view showing an example of an elincRNA (putative promoter region in pink, right) and two other lincRNAs (putative promoter region in blue, left). The promoter regions of other lincRNAs have low levels of promoter (H3K4Me3) and enhancer (H3K4Me1) histone marks, while the promoter region of the elincRNA has high H3K4Me1, but low H3K4Me3. Gene bodies are depicted in green, with solid boxes as exons and thin lines as introns.

Figure S1: Expression of elincRNAs compared with other lincRNAs and PCG in different cell lines. Sets of genes as defined in GM12878 are used in all comparisons. Two-tailed Mann-Whitney test, \*\*\*P<0.001; – not significant

4.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 & Cavalli, 2016)⁠. Following the development of chromosome conformation capture techniques over a decade ago (J 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) (Job Dekker, Marti-Renom, & Mirny, 2013)⁠.

LincRNAs originated 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 anchor and high amount of intra-TAD contacts within elincRNA-containing topological domains support their putative role in regulating enhancer-promoter loops.

There is ongoing debate on the biological relevance of elincRNA transcription as to whether their associated functional roles are transcript- or transcription-dependent. Notably, although I observed an association between elincRNA transcription and chromosomal organization, evidence for transcript-mediated functions remains to be investigated. This can be tested by comparing active enhancers producing stably and uni-directionally transcribed elincRNAs with others that produce unstable and bidirectionally-transcribed eRNAs. Ultimately, genetic manipulations of elincRNA transcript abundance and their transcription, for example by using RNAi and CRISPR-Cas9, can be applied to dissect the molecular mechanisms that underly these enhancer-associated regulation of chromosomal conformation (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 chromosomal interactions that shape the nuclear architecture. These analyses could also be extended to additional cell-lines to investigate whether elincRNAs’ roles 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 elincRNA loci that are smaller than the highest resolution (i.e. 5 Kb) (Rao et al., 2014b)⁠. 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 averaging out the dynamics and variabilities between cells. Recently, single-cell Hi-C protocol was developed (Nagano et al., 2013)⁠, allowing one to examine DNA contact profiles at a single cell resolution. Indeed, such single cell techniques, combined with higher contact resolution, would provide much more power to examine the impact of elincRNAs on gene expression regulation through modulating chromosome architecture.

5. 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 & Hall, 2010)⁠or the “intervals” package (Bourgon, 2015)⁠ in R. Manipulations on Hi-C contact matrices were performed using the “Matrix” package (Bates & Maechler, 2016)⁠. All position of genetic elements used are based on the hg19 human assembly.

elincRNA definition

LincRNAs and protein-coding genes sets were provided by Tan et al (2016, under revision). Those sets were obtained using data from the ENCODE website. The list of genes used in all analyses corresponds to genes expressed in the GM12878 human lymphoblastoid cell line. Subcategories of genes were defined based on overlap between their putative promoter region, defined as the 1kb region upstream of 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 promoters and all enhancers were considered. The 2 categories of lincRNAs that are used throughout this report are elincRNAs, defined as overlapping enhancers marks but no promoters marks in their promoter region, and other lincRNAs defined as overlapping neither enhancer nor promoters marks in their promoter region (Figure 7).

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 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. 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 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.

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 boundaries are insulated regions. In other words, there are few interactions between elements located before and those after the boundaries. The insulation is measured by sliding a diamond (Figure 4A) 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:

$\left\{\begin{matrix}1\leq w\leq \frac{n}{2}+1 \\ \forall d\in\left \{ w , ... , n-\left ( w-1 \right ) \right \}\end{matrix}\right. V\_{d}=\sum\_{i=d-(w-1)}^{d}\sum\_{j=d}^{d+(w-1)}M\_{i,j}$

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 an arbitrary threshold defined as the starting value (at the border) plus 10% of the maximum value in the TAD (Figure 5). The boundaries were filtered afterwards to remove all those extending beyond their TAD. This happened in cases where the TAD was embedded in a larger one and the border was already among the highest values in the vector V. The algorithm would need to account for background noise in order to properly process the boundaries in these special cases.

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 (Tau) was computed following the described procedure in Kryuchkova & 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 4B).

Chip-seq data

Chip-seq peaks for CTCF, RAD21 and SMC3 in GM12878 were retrieved from the ENCODE website (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 cohesin and CTCF.

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.

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 analysis were performed at the Vital-IT Center for High Performance Computing of the Swiss Institute of Bioinformatics (www.vital-it.ch).

6. References:

Bates, D., & Maechler, M. (2016). Matrix: Sparse and Dense Matrix Classes and Methods. manual.

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics, 17(11), 661–678. http://doi.org/10.1038/nrg.2016.112

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

Darrow, E. M., & Chadwick, B. P. (2013). Boosting transcription by transcription: enhancer-associated transcripts. Chromosome Research, 21(6), 713–724. article. http://doi.org/10.1007/s10577-013-9384-6

Dekker, J. (2002). Capturing Chromosome Conformation. Science, 295(5558), 1306–1311. http://doi.org/10.1126/science.1067799

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews. Genetics, 14(6), 390–403. http://doi.org/10.1038/nrg3454

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Systems, 3(1), 99–101. http://doi.org/10.1016/j.cels.2015.07.012

ENCODE Project, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74. http://doi.org/nature11247 [pii]\n10.1038/nature11247

Engreitz, J. M., Ollikainen, N., & 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. http://doi.org/10.1038/nrm.2016.126

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., … Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol, 11, 1–14. article. http://doi.org/10.15252/msb

Gorkin, D. U., Leung, D., & Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell, 14(6), 771–775. http://doi.org/10.1016/j.stem.2014.05.017

Guil, S., & Esteller, M. (2012). Cis-acting noncoding RNAs: friends and foes. Nature Structural & Molecular Biology, 19(11), 1068–1075. http://doi.org/10.1038/nsmb.2428

Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., … Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature, 460(7253), 410–413. http://doi.org/10.1038/nature08079

Heger, A., Webber, C., Goodson, M., Ponting, C. P., & Lunter, G. (2013). GAT: A simulation framework for testing the association of genomic intervals. Bioinformatics, 29(16), 2046–2048. http://doi.org/10.1093/bioinformatics/btt343

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., … Chinnnaiyan, A. M. (2015). The Landscape of Long Noncoding RNAs in the Human Transcriptome. Nat Genet., 47(3), 199–208. http://doi.org/10.1038/ng.3192.The

Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., … Young, R. A. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell, 18(2), 262–275. http://doi.org/10.1016/j.stem.2015.11.007

Kornienko, A. E., Guenzl, P. M., Barlow, D. P., & Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC Biology, 11(1), 59. http://doi.org/10.1186/1741-7007-11-59

Kryuchkova, N., & Robinson-Rechavi, M. (2015). A benchmark of gene expression tissue-specificity metrics. bioRxiv, (January), 27755. http://doi.org/10.1093/bib/bbw008

Lam, M. T. Y., Li, W., Rosenfeld, M. G., & Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. Trends in Biochemical Sciences, 39(4), 170–182. http://doi.org/10.1016/j.tibs.2014.02.007

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., … Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature, 498(7455), 516–20. http://doi.org/10.1038/nature12210

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., … Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 161(5), 1012–1025. article. http://doi.org/10.1016/j.cell.2015.04.004

Marques, A. C., Hughes, J., Graham, B., Kowalczyk, M. S., Higgs, D. R., & 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. article. http://doi.org/10.1186/gb-2013-14-11-r131

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, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. http://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2016). R: A Language and Environment for Statistical Computing. manual, Vienna, Austria.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014a). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680. article. http://doi.org/10.1016/j.cell.2014.11.021

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014b). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680. http://doi.org/10.1016/j.cell.2014.11.021

Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. Annual Review of Biochemistry, 81, 145–166. http://doi.org/10.1146/annurev-biochem-051410-092902

Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., … Merkenschlager, M. (2013). Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Research, 23(12), 2066–2077. http://doi.org/10.1101/gr.161620.113

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 1034–1050. http://doi.org/10.1101/gr.3715005

Tamaru, H. (2010). Confining euchromatin/heterochromatin territory: Jumonji crosses the line. Genes and Development, 24(14), 1465–1478. article. http://doi.org/10.1101/gad.1941010

Yin, Y., Yan, P., Lu, J., Song, G., Zhu, Y., Li, Z., … 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. http://doi.org/10.1016/j.stem.2015.03.007