1. Introduction :

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

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

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

Unlike most enhancer-associated RNAs, which are often transcribed bidirectionally and then rapidly degraded (Darrow & Chadwick, 2013)⁠, elincRNAs are transcribed preferentially in one direction (Marques et al., 2013)⁠ and are likely good candidates for studying the involvement of lincRNAs in the regulation of DNA-DNA contacts. There are already some characterized elincRNAs, such as HOTTIP that have been shown to act as a link between chromosomal interactions and transcription. My analysis shows elincRNAs are associated with high density of chromosomal contacts within TADs and are significantly enriched in cohesin and CTCF binding sites, supporting the idea that they may contribute to gene regulation by modulating chromosomal organization.

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

2. Results

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

Expression levels

Enhancer RNAs in general are known to have relatively low transcript abundances as they tend to be rapidly degraded by the exosome (Lam, Li, Rosenfeld, & Glass, 2014)⁠. I first tested if that is the case for my set of elincRNA compared to other lincRNAs and protein-coding genes (Figure 1). There was no significant difference between the expression levels of elincRNAs and other lincRNAs (two-tailed Mann-Whitney test, p=0.258) in GM12878. However, when comparing the expression of the same sets of genes, elincRNAs had significantly lower expression levels than other lincRNAs in two other cell lines (Figure 2). These results suggest that the expression levels of elincRNA are generally lower, but very dependent on the cell line.

Conservation

ElincRNAs are less conserved than both other lincRNAs and protein coding genes (Figure 3). These differences are consistent both when looking at conservation in mammals, or in primates. Again, these observations concur with previous findings in mice (Marques et al., 2013) showing that elincRNAs are almost under neutral evolution. This lower selective constraint also suggests that similar sets of genes will tend to be enhancer associated in different cell lines.

Enrichment at TAD boundaries

There was no direct signal for enrichment of elincRNAs promoter regions at TAD boundaries(Figure 4), however the promoter regions of elincRNAs were found to be enriched at anchors. Those anchors are themselves enriched at TAD boundaries (1.74 fold, q<0.001). The lack of direct signal for enrichment at TAD boundaries may be a consequence of the method used to defined them. Notably, I restricted boundaries to the inside of TADs, therefore genes that are close to a TAD, but outside the border are not detected.

Testing for enrichment of elincRNAs and other lincRNAs promoter regions through TADs divided into 10 bins does not reveal any enrichment near boundaries either (Figure 5). Surprisingly, we observe a sudden depletion (0.37 fold, q=0.06) of elincRNAs in the middle of TADs (bin number 5). However, none of these results are significants and therefore, one should not try to draw conclusions from these.

Architectural proteins binding sites

CTCF and cohesin are often called architectural or insulator proteins, as they are thought to prevent TADs from interacting with each other while increasing interactions within TADs. Enrichment tests for the binding sites of these proteins in elincRNAs revealed that CTCF, SMC3 and RAD21 binding sites 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) (Figure 6).

Most binding sites for CTCF and cohesin overlap in the genome (Figure 7) and to determine if the enrichment of CTCF binding sites is a consequence of the overlap with cohesin binding sites, I performed enrichment tests for CTCF and cohesin exclusive binding sites (Figure 8). The difference in fold enrichment for cohesin binding sites in elincRNAs compared to other lincRNAs much stronger when looking at exclusive binding sites, while it decreased for CTCF.

DNA-DNA contacts

Since enhancer associated RNAs are thought to be involved in chromosome looping, To investigate wether elincRNAs are associated with a higher amount of DNA-DNA contact, I computed the average amount contact per 5kb in the corresponding TAD for each gene (see material and methods for details). I find that elincRNAs are associated with TADs presenting higher amounts of contacts across 3 different cell lines, but not in the 4th. The strength of this association seems to be very cell line-dependent, but always in the same direction. Interestingly, the amount of contact vary dramatically across cell lines.

3. Figures and tables

Name

Overlap status

N. lincRNA

elinc

Overlaps enhancer only.

236

plinc

Overlaps promoter only

480

other

Overlaps neither promoters, nor enhancers

1756

Figure 1: Expression of elincRNAs compared with other lincRNA and protein-coding genes (PCG). Numbers on the boxes are median values. other lincRNAs. Two-tailed Mann-Whitney test, \*\*\*P<0.001

Figure 2: Expression of elincRNAs compared with other lincRNAs and PCG in different cell lines. Sets of genes are the same in all comparisons. Two-tailed Mann-Whitney test, \*\*\*P<0.001

Figure 3: Sequence conservation of elincRNAs is significantly lower than other lincRNAs and protein coding genes. The green horizontal line represents the median conservation of ancestral repeats, which are assuming to be evolving neutrally. Two tailed Mann-Whitney test, \*\*\*P<0.001.

Figure 4: Enrichment of elincRNA promoter regions at TAD boundaries and loop anchors compared to other lincRNAs. Fold enrichments and associated q-values are written on the bars.

Figure 5: Enrichment of elincRNAs across TADs. Each bar represent a bin of 10% TAD length. The log10 of q-values are put in color codes to give an estimation of the confidence in each value.

Figure 6: Enrichment in architectural proteins in elincRNA promoter regions, compared to other lincRNA. Fold enrichment and associated q-values are written on the bars. SMC3 and RAD21 are subunits of cohesin.

Figure 7: Proportions of overlap between RAD21, SMC3 and CTCF binding sites in the human genome.

Figure 8: Enrichment of CTCF and cohesin exclusive binding sites in promoter regions of elincRNAs compared to other lincRNAs. Fold enrichment and associated q-values are written on the bars.

Figure 9: Amount of within-TAD contact for elincRNAs compared to other lincRNAs across different cell lines. Two tailed Mann-Whitney test, \*P<0.05.,\*\*\*P<0.001.

Figure 10: Tissue specificity of elincRNA compared to other lincRNA and protein-coding genes. Median values are written on the boxes. Two tailed Mann-Whitney test, \*\*\*P<0.001.

Figure 11: Schematic representation of the algorithm used to estimate TAD boundaries based on intrachromosomal contact matrices. The algorithm takes the sum for each square submatrix of width w along the diagonal. The sums are then stored and used to compute TAD boundaries.

4.Discussion:

Found/didn’t found association between elincRNAs and….

 Although these results suggest a role for elincRNAs in TAD organization, they provide no information on their exact function of mechanistic role.

Most active enhancers are transcribed (Arner et al., 2015)⁠ and the resulting transcripts are hypothesized to recruit transcription factors to increase transcriptional activity.

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)⁠.

Genes

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

TAD definition

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

TAD boundaries definition

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

Conservation and tissue specificity

The sequence conservation was previously calculated (Tan Yihong, J., unpublished) using phastCons scores (Siepel et al., 2005)⁠. Tissue specificity index (Tau) was computed following the described in Kryuchkova & Robinson-Rechavi (2015⁠), considering only genes with expression above a 0.1 [RPKM?] cutoff.

Expression levels

Median expression levels were computed from [the GEUVADIS 1000 Genomes project expression matrix. The matrix contains expression data for 373 individuals.] → encode, source ?

DNA-DNA contacts

Contacts were calculated using Hi-C contact matrices from Rao et al (2014). All computations are performed on 5kb resolution matrices with a MAPQ score of at least 30. The matrices were normalized using the KR normalization vector provided by the authors, when analyzing only GM12878. When comparing between different cell lines, SQRTVC (square root vanilla coverage) was used for chromosome 9 of all cell lines, because the KR algorithm did not converge for chromosome 9 of K562.

For each gene overlapping a TAD, the mean contact inside the respective TAD was used as a measure. For single genes that overlap several TADs, the contacts are computed for each TAD independently. The mean contact in a TAD is computed by taking the arithmetic mean in a square submatrix spanning from the beginning to the end of the TAD in the intrachromosomal matrix.

Chip-seq

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

Enrichment of genetic elements

All enrichment tests were performed using the genome association tester (GAT) (Heger et al., 2013)⁠ version 1.2. All tests for enrichment of lincRNAs were performed using the intergenic space of the genome as a workspace. When testing for enrichment of protein-coding genes, the protein coding space of the genome was used as the workspace. Similarly, when looking for enrichmentFor all tests, the number of samples was set to 10,000 and the number of buckets was consequently adjusted to 270,000.

6. References:

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., … Hayashizaki, Y. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science (New York, N.Y.), 347(6225), 1010–4. http://doi.org/10.1126/science.1259418

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

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

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

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

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

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

Lupiáñez, D. G., Spielmann, M., & Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends in Genetics, 32(4), 225–237. http://doi.org/10.1016/j.tig.2016.01.003

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. http://doi.org/10.1186/gb-2013-14-11-r131

Ong, C., & Corces, V. G. (2014). CTCF : an architectural protein bridging genome topology and function. Nature Publishing Group, 15(4), 234–246. http://doi.org/10.1038/nrg3663

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, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., … Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature, 515(7527), 402–405. http://doi.org/10.1038/nature13986

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. (2014). 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

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. http://doi.org/10.1101/gad.1941010