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)⁠. 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, demonstrate 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 and functions, if any, for the majority of lincRNAs remain unknown (Rinn & Chang, 2012)⁠.

Most active enhancers are transcribed, often generating bidirectional, short and unstable noncoding RNAs (ref). Recently enhancer activity has also been shown to be the origin of a large fraction of lincRNAs (ref) (Guil & Esteller, 2012). Recently, lincRNAs associated with human trait variants have been shown to have enhancer-associated cis-regulatory roles on local gene expression. Importantly, these loci are correlated with higher frequencies of chromosomal interactions relative to other lincRNAs in a human lymphoblastoid cell line (LCL), suggesting that enhancer-associated lincRNAs (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, 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 are associated with high levels of active transcription, referred to as euchromatin. 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 where frequent DNA:DNA interactions occur as a result of their close spatial proximity, called topologically associating domains (TADs), which are key in modulating gene transcription programs. Recently that regions at the borders of TADs, called TAD boundaries, were shown to be essential for modulating and preventing gene regulatory interactions within and across TADs, respectively (REF). Deletion of TAD boundaries often disrupts those interactions, resulting in gene misexpression and disease phenotypes (Lupiáñez et al., 2015)⁠.

Some lincRNAs, in particular those arising at enhancers (elincRNAs) have been found to participate in the spatial organization of the genome. For example, (Yin et al., 2015)⁠⁠, can regulate expression of the HOXA gene cluster by modulating intrachromosomal interactions, specifically through 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, supporting 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 identified based on overlap at their putative promoter regions, defined as 1kb upstream from their transcriptional start site, with LCL enhancers predicted based on histone marks (GM12878, (ENCODE Project et al., 2012)⁠). LincRNAs whose promoter region also overlapped other predicted regulatory regions, specifically promoter elements (ENCODE Project et al., 2012)⁠, in LCLs were excluded from the analysis (elincRNAs=236, other LCL-expressed lincRNAs=1756).

*elincRNAs show similar expression levels as other lincRNAs*

Unlike most enhancer-associated noncoding RNAs (eRNAs) which are often transcribed bidirectionally and then rapidly degraded by the nuclear exosome (Darrow & Chadwick, 2013)⁠ (Lam, Li, Rosenfeld, & Glass, 2014)⁠, elincRNAs are more stably and preferentially transcribed in one direction, and are often spliced and polyadenylated (Marques et al., 2013)⁠.

First, to investigate if LCL-expressed elincRNAs share similar expression profiles as the lowly expressed eRNAs relative to other expressed genes, I compared their expression levels to that of other lincRNAs and protein-coding genes (Figure 1A). I found that elincRNAs are not more lowly expressed that are other lincRNAs in human LCLs (GM12878, two-tailed Mann-Whitney U test, p=0.258), suggesting elincRNAs may have distinct properties compared to other enhancer-associated transcripts.

*elincRNA transcripts are less conserved than other lincRNAs*

To gain insights into elincRNAs evolution, 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.

*elincRNAs promoter regions are enriched at loop anchors and cohesin binding*

Next, to examine whether elincRNAs are associated with the regulation of chromosomal architecture, I investigated the co-localization of elincRNAs with loop anchors where enhancer-promoter gene regulatory interactions occur (Ji et al., 2016). These regulatory regions are essential to establishing chromosomal interactions within topologically associating domains (TADs).I compared the overlap of elincRNAs with these regions relative to what would be expected if tese loci were randomly distributed across the intergenic regions of the genome (Methods). I found that elincRNAs are significantly enriched at LCL loop anchors than expected by chance (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., 2014a)⁠. Specifically, the cohesin protein-complex has been previously shown to be 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 or CTCF and cohesin collectively are important for the structural maintenance and have insulating properties, preventing interactions between TADs . In contrast, loops containing only cohesin binding sites are crucial in mediating regulatory chromosomal interactions. Cohesin depletion is associated with disrupted promoter-enhancer interactions within TADs (Seitan et al., 2013)⁠.

I found that elincRNA promoters are significantly enriched in LCL CTCF and cohesin binding sites (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 enrichement is greater than the one found for other lincRNAs (fold enrichment = 1.3 and 1.3, respectively, q<0.05, Figure 2C).

As most binding sites for CTCF and cohesin overlap in the human genome (Figure 2E) and loops mediated by CTCF or cohesin independently, or both CTCF and cohesin collectively are thought to have different roles in chromosomal organization (Ji et al., 2016)⁠, I further determined the independent enrichment of elincRNAs in CTCF and cohesin binding by obtaining mutually exclusive binding sites for these proteins. This revealed a greater enrichment of cohesin binding sites (Figure 2D) at elincRNA loci (fold enrichment = 13.1 fold, q=1e-04) compared to CTCF binding sites (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.

*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 in LCLs (Figure 3) I determined TAD boundary regions through a sliding diamond approach on contact matrices (Figure 4A, methods). Boundaries were extended from TAD borders until the sum of interactions in the diamond reached an arbitrary threshold (Figure 5, methods). 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 2B). Despite the absence of significant elincRNA enrichment at TAD boundaries, dividing TADs into 10 equally sized bins revealed that elincRNAs tend to be more frequently found near the end of the TADs and are depleted at the center of the TADs (bin 5, fold enrichment = 0.37 fold, q=0.06) relative to other LCL-expressed lincRNAs (Figure 2F). 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 observed 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 extending inwards from borders of TADs (see methods for details), therefore all genes located outside of defined TADs but close to a TAD border are excluded from 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 higher density DNA:DNA interactions within elincRNA containing TADs. To this end, I measured the average amount of contact in elincRNA TADs (Figure4B, methods). I found that these TADs display a higher mean density of contacts relative to TAds containing other lincRNAs in GM12878 (Median fold difference=1.24, two-tailed Mann-Whitney U test, p<1e-04, Figure 6). In addition, the fold difference in the amount of DNA contacts within TADs harbouring these elincRNA loci 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 mechanisms through which elincRNAs may regulate chromosomal architecture, their associated high DNA:A contacts, together with their enrichment of cohesin binding suggest a role for elincRNAs in the modulation of promoter-enhancer looping within topologically associating domains.

**3. Figures and tables**

**Figure 1:** ElincRNAs properties. A) Distribution of elincRNAs (red), other lincRNaA (green) and protein coding gene expression in LCL (GM12878).. B) Distribution of tissue specificity index (Tau). Tau ranges between 0 (low specificity) and 1 (high specificity). C) Average exonic sequence conservation through mammalian and primate evolution (phastCons score). The green horizontal line represents the median conservation of ancestral repeats, which are used here as a neutral proxy. Differences between distributions were tested using a two-tailed Mann-Whitney test, \*\*\*P<0.001; – not significant

**Figure 2: Association between elincRNAs and features relevant to chromosomal archictecture.** Fold-enrichment relative to random expectation of elincRNA and other lincRNAs promoter A) at loop anchors B) TAD boundaries, C) cohesin and CTCF and D) CTCF and cohesin exclusive binding sites. E) Proportions of overlap between RAD21, SMC3 and CTCF peaks in the human genome. F) Enrichment of elincRNAs and other lincRNAs 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. q represents the p-value adjusted for multiple testing using FDR (methods)

**Figure 3:** Schematic representation of TADs and loops and typical patterns observed in Hi-C matrices corresponding to these structures. Center: Example of a Hi-C matrix visualized in Juicebox (Durand et al., 2016)⁠. These matrices are symmetric and only upper/lower triangles are therefore used to simplify the visualization. The darker pixels on the matrix contain more interactions. Left: Two separate TADs are observed as high interactions triangle on the matrix. These are examplified by regions of compacted DNA where frequent interactions occur, while interactions across TADs (i.e. between the blue and green triangles) are less frequent. A and B are the borders of the first TAD while C and D are the borders of the second TAD. Boundaries are the rectangles expanding inwards from the borders. Right: Representation of a loop with A and B being the anchors of the loop where strong contact is observed. Unlike TADs, the contact is not particularly high in the region between the two anchors, therefore loops are seen as a sharp increase in contacts deviating from the matrix diagonal.

**Figure 4:** 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 across 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). B: Method used to compute the mean of all interactions in TADs. 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 P through a TAD~~. The value Vd is the sum of all interactions within a diamond placed at position d. All Vd values are stored in vector V. d represents the position on the diagonal of the matrix. d slides along the diagonal with a step of 5kb. Boundaries were extended until they reach a threshold, defined in the methods. 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 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 6:** Density of DNA:DNA contact within TADs. The distribution of DNA :DNA contacts for TADs containing elincRNAs and other lincRNAs across different cell lines. Set of genes as defined in GM12878 are used for all comparisons. Two tailed Mann-Whitney test, \*\*\*P<0.001; \*P<0.05; – non-significant

**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 conformation of chromatin is key to the regulation of gene transcription programs, features that underlie these chromatin organization dependent transcriptional regulation of remain relatively unclear (Bonev & Cavalli, 2016)⁠. Following the development of chromosome conformation capture techniques over a decade ago (J Dekker, 2002)⁠, have allowed the quantification of chromosomal interaction frequencies between genomic loci in close 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 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 genomic data to investigate the prevalence of elincRNAs regulation in modulating chromosomal architecture. Together with strong enrichment of cohesin binding at elincRNAs promoter regions and their frequent localization at loop anchors, the 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 transcript 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 these enhancer-associated regulation of chromosomal conformation will require genetical manipulation of candidate elincRNA transcript abundance, for example by using RNAi and inhibiting their transcription with CRISPR-Cas9 (Li et al., 2013)⁠ .

Furthermore, extending the analysis from intra-topological domain interactions to between-TAD contacts (Fraser et al., 2015)⁠, inter-chromosomal contacts, and eventually association with nuclear lamina would provide a more complete overview of elincRNA-dependent chromosomal interactions that shape the nuclear architecture. These analyses could also be performed on different 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 impossible to look at contacts between elincRNAs that are smaller than the highest resolution (5 Kb) (Rao et al., 2014b)⁠. Instead, only an estimation of the region surrounding the gene can be measured. In addition,, even at high contact resolutions, bulk Hi-C data can only provide an average estimate of all chromosomal contacts happening in a cell population, thus masking the dynamics and variability between cell-to-cell contacts. Recently, a single-cell Hi-C protocol has been developed (Nagano et al., 2013)⁠, allowing one to examine DNA contact profiles at a single cell resolution. Such single cell techniques, combined with a higher contact resolution, would provide greater power to shed light on how 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)⁠.

*elincRNA definition*

LincRNAs and protein-coding genes sets were provided by Jennifer Tan.. Generally those sets were obtained by combining ENSEMBL annotations for the Human genome (version XXXX) with a set of lincRNAs annotated de novo using LCL transcriptome data obtained from the ENCODE consortium. The list of genes used in all analyses corresponds to genes expressed in the GM12878 human lymphoblastoid cell line defined as XXXXX. LincRNAs were classified 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 not promoters marks in their promoter region, and other lincRNAs defined as overlapping neither enhancer nor promoters marks at their promoter region (Figure 7).

*TAD definition*

The list of TADs used in the computations was obtained from Rao et al (2014). These TADs were called 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 $M^\*\_{i,j}$ the corresponding entry in the normalized matrix.

*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 a sliding 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).

*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 Jennifer Tan. using RNA sequencing data generated by the ENCODE consortium (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.

*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