

Characterization of distinct epigenomic features associated with Enhancer-like promoters in human cell-lines

A dissertation submitted for partial fulfillment of BS-MS dual degree in Science

Shagun Puri

MS16074



Indian Institute of Science Education and Research Mohali
Knowledge City, Sector 81, SAS Nagar, Manauli, PO 140306

April 2021

Certificate of Examination

This is to certify that the dissertation titled “Characterization of distinct epigenomic features associated with Enhancer-like promoters in human cell-lines” submitted by Shagun Puri (Reg. No. MS16074) for the partial fulfillment of BS-MS dual degree program of IISER Mohali has been examined by the thesis committee duly appointed by the institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.



Dr Shashi Bhushan Pandit



Dr Rajesh Ramachandran



Dr Kuljeet Singh Sandhu
(Supervisor)

Dated: April 30th, 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Singh Sandhu at the Indian Institute of Science Education and Research, Mohali.

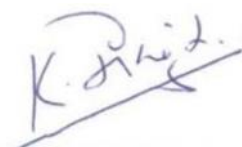
This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.



Shagun Puri
(Candidate)

Dated: April 30th, 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.



Dr Kuljeet Singh Sandhu
(Supervisor)

Dated: April 30th, 2021

Acknowledgement

I would like to express sincere gratitude to my supervisor Dr Kuljeet Singh Sandhu for all the useful comments and remarks throughout the project.

I would also like to extend thanks to my thesis committee members – Dr Shashi B. Pandit and Dr Rajesh Ramachandran. Big thanks to IISER Mohali for providing academic facilities throughout the curriculum and to KVPY for providing me with the fellowship, encouraging students like myself to pursue Science.

I would like to thank my family for being my support system throughout this rocky year, and providing me with a comfortable life at home while I did my MS project remotely. Special thanks to my sister – Anmol, for keeping things light whenever I was feeling down.

I am also thankful to the members of Genome Biology Lab – Yachna, Meenakshi, Mohan, Jui and ex-lab members – Sunandini, Lipika and Arashdeep for always being ready to extend a helping hand, and to Arpit – my fellow MS thesis student in the lab for healthy discussions about our projects.

Shagun

List of Figures

Figure 3.1: Infection point of fold change values for Enhancer activity of promoters for HeLa-S3 and K562 (CapStarr-seq) in (A) HeLa S3 , (B) K562

Figure 3.2: Correlogram for histone modifications and enhancer potential of promoters in (A) HeLa S3, (B) K562

Figure 3.3: Partial least squares regression biplots for PLSR performed on Histone modification matrix and Enhancer of ELPs in (A) HeLa S3, (B) K562

Figure 3.4: Barplots of X-loadings along component 1 for histone modifications in (A) HeLa S3, (B) K562

Figure 3.5: Correlogram for Transcription factors and enhancer potential of promoters in (A) HeLa S3, (B) K562

Figure 3.6: Partial least squares regression biplots for PLSR performed on Transcription factor matrix and Enhancer potential of ELPs in (A) HeLa S3, (B) K562

Figure 3.7: Barplots of X-loadings along component 1 for Transcription factors in (A) HeLa S3, (B) K562

Figure 3.8: Line plot for Time course gene expression data of ELPs and PLPs in (A) K562 cells stimulated by Hemin, (B) HeLa S3 cells stimulated by EGF

Figure 3.9: Stacked bar plots showing the proportion of ELPs and PLPs associated with Transcription factor encoding genes in (A) K562, (B) HeLa S3

Figure 3.10: Plot for enrichment of ELPs and PLPs within and outside CTCF loops in (A) K562, (B) HeLa S3

Contents

| | |
|--|------|
| 1. Certificate of Examination..... | iii |
| 2. Declaration..... | v |
| 3. Acknowledgement..... | vii |
| 4. List of figures..... | viii |
| 5. List of databases and tools..... | x |
| 6. Abstract..... | xi |
| 7. Chapter 1: Introduction..... | 1 |
| 8. Chapter 2: Materials and methods..... | 3 |
| 2.1: Dataset for Enhancer potential of promoters | |
| 2.2: Histone and Transcription factor CHIP-SEQ datasets | |
| 2.3: Quantile normalization of the datasets | |
| 2.4: Partial Least Squares Regression analysis | |
| 2.5: Time course gene expression | |
| 2.6: Calculating enrichment in CTCF loops using CTCF CHIA-PET data | |
| 9. Chapter 3: Results..... | 8 |
| 3.1: Inflection point of CapStarr-seq datasets | |
| 3.2: Histone modifications and TFs associated with ELPs | |
| 3.3: ELPs are inducible promoters and are associated with early response genes | |
| 3.4: ELPs are enriched inside CTCF loops | |
| 10. Chapter 4: Discussion..... | 17 |
| 11. Appendix..... | 19 |
| 12. Bibliography..... | 20 |

List of databases, tools and softwares

1. ENCODE project : <https://www.encodeproject.org/>
2. GEO (NCBI) : <https://www.ncbi.nlm.nih.gov/geo/>
3. ENSEMBL Biomart : <https://www.ensembl.org/biomart/martview/>
4. FANTOM : <https://fantom.gsc.riken.jp/>
5. BEDTOOLS
6. BEDOps

Abstract

The human genome vastly consists of non-coding DNA, and this portion contains variants associated with diseases, as shown by GWAS studies. Many studies have shown that the SNPs for diseases often lie in the regulatory elements of other genes. This creates a gaping hole in our understanding of how Trans regulation among genes happens inside our genome and how these phenotypes manifest as a result of this regulatory propagation in the genome. Recent studies have shown that certain promoters interact with each other, much like enhancers and promoters interact spatio-temporally. Several other studies have shown that a small proportion of promoters display enhancer activity. This evidence along with studies that blur the classical architectural demarcations between enhancers and promoters indicate that gene regulation is much more complex than earlier thought of, and how these promoter-promoter interactions could amplify in the genome due to network effect.

In this thesis, we study the epigenomic markers associated with Enhancer-like promoters, namely Histone modification and Transcription factor marks in two ENCODE human cell lines – K562 and HeLa-S3. We also analyse how Enhancer-like promoters are temporally expressed with respect to regular promoters in the presence of environmental stimuli, and their position in the 3-D genome with respect to CTCF loops. We hypothesize that these promoters with enhancer activity are associated with inducible genes and kick-start the developmental program in the cell. They are enriched within CTCF loops; which constrain the transcriptional induction of ELPs and stop ripple effect.

Chapter 1

Introduction

Gene expression in eukaryotes is regulated spatio-temporally through gene-regulatory elements – Enhancers and promoters. Classically, promoters have been defined as TSS-proximal gene regulatory elements, which consist of TATA box, initiator and downstream core promoter element and interact with RNA polymerase II. These elements have the capacity to induce gene expression locally. Whereas, enhancers have been defined as TSS-distal gene regulatory elements which induce gene expression at a distance, independent of the distance and the orientation with respect to the promoter (Atchison, 1988). This classification has also been supported by studies that showed that several epigenomic marks are associated with these regulatory elements. A 2007 study showed that H3K4me3 (H3 lysine 4 trimethylation) is enriched in regions proximal to the 5' end of genes ,i.e., around active gene promoters, whereas H3K4me1 was found to be enriched in the distal gene regulatory elements (Heintzman et al., 2007). This implies that the signal ratio (H3K4me1:H3K4me3) could be used to distinguish between enhancers and promoters and H3K27ac has been shown to be associated with active promoters and enhancers (Creyghton et al., 2010). However, many studies challenge these conventional ideas and have shown that broad similarities exist between these regulatory elements. Although association with RNAPII has been defined to be the key property of promoters, it has been shown that enhancers bind with RNAPII as well, leading to the formation of non-coding enhancer RNAs (eRNAs). These eRNAs are often transcribed bi-directionally, which is a feature of many mammalian promoters resulting in the formation of short anti-sense non-coding RNAs (Kim et al., 2010), (Seila et al., 2008). Besides these structural similarities between enhancers and promoters, it has also been shown that some promoters exhibit enhancer activity as well. In a 2012 study, the authors used genome-wide CHIA-PET methodology to map long range chromatin interactions associated

with RNAPII, and found that there exists a widespread phenomenon of promoter-centred intragenic, extragenic and intergenic interactions in the human genome (Li et al., 2012). These interactions led to the formation of clusters, which led to the interaction of proximal and distal genes through promoter-promoter interactions. More recently, many novel techniques have been developed to identify enhancers and quantify their activity in the genome. One such technique is CapStarr-seq (Vanhille et al., 2015), which captures genomic sequences of interest for high-throughput assessment of enhancer activity in mammals. Using this technique, a research group carried out a genome-wide assessment of promoters in mammalian cell lines to quantify the enhancer activity of mammalian promoters, discovering a set of promoters displaying enhancer potential in the studies cell lines (Dao et al., 2017).

A major implication of the presence of these Enhancer-like promoters is that it adds one more layer of complexity to the already complex process of transcriptional regulation of genes and to our understanding of genotype to phenotype associations in complex disorders. Genome Wide Association Studies (GWAS) studies have been used in the recent past as an effective method to identify genomic loci associated with common diseases, which have identified a lot of noncoding variants associated with common diseases/ phenotypes (Maurano et al., 2012). This also alludes to the fact that if these variants are found inside enhancer like promoters (ELPs) they could lead to disease related phenotypes associated with distal genes, via promoter-promoter interactions.

A variant associated with Type 2 Diabetes – rs11603334 lies within the ARAP1 promoter and affects the PAX6/PAX4 binding in pancreatic islets. Interestingly enough, the ARAP1 promoter also displayed Enhancer activity in STARR-seq assays (Kulzer et al., 2014), (Dao et al., 2017).

Since Enhancer like promoters can have such a strong network effect in the genome, in this study, we analyse in two mammalian cell-lines, HeLa-S3 and K562, the features of these promoters displaying enhancer activity and compare them with regular promoters. We also study their temporal expression patterns when induced by external factors, in order to gain insights about their role in the development.

Chapter 2

Materials and Methods

2.1 Dataset for Enhancer potential of promoters

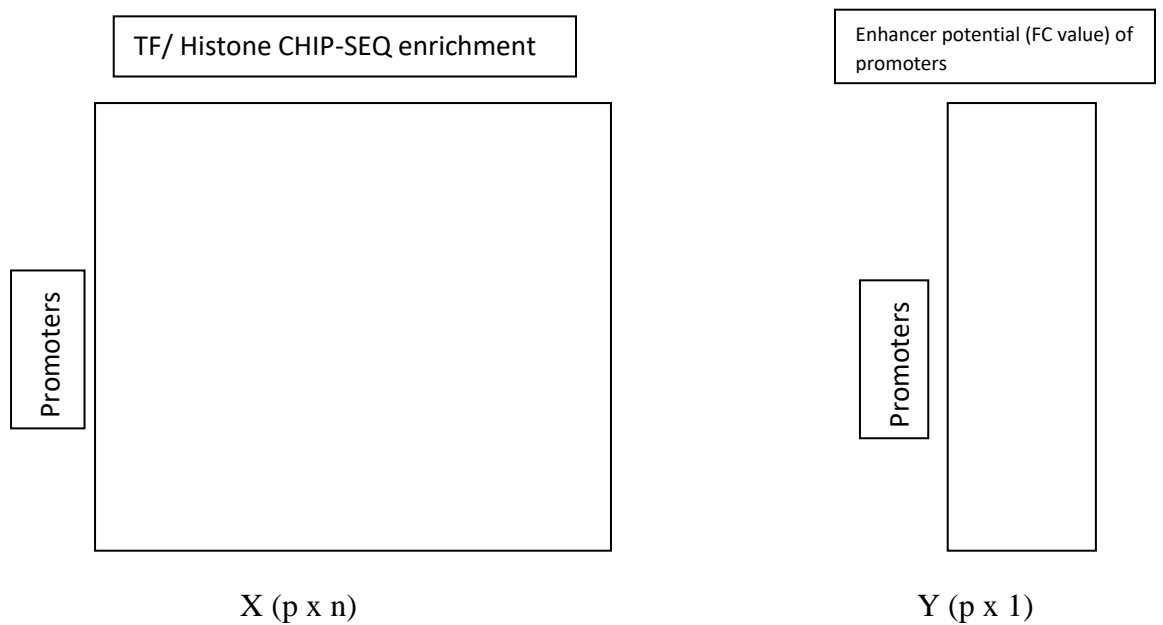
The dataset for enhancer activity of promoters was taken from a 2017 paper (Dao et al., 2017), wherein the authors had performed CapStarr-seq assay for all promoters of RefSeq-defined human coding genes to characterize ELPs (Enhancer-Like Promoters) in an unbiased manner for two ENCODE cell lines – HeLa-S3 (Human Cervix carcinoma cell line) and K562 (Human myelogenous leukemia cell line). The fold change of the CapStarr-seq signal over input for two replicates was considered to be the quantification of enhancer potential of these promoters. We used this data such that we took the average of this fold change value for each gene (for both replicates) and defined the promoter region to be $TSS \pm 500$ bp.

For classifying the 20720 promoters as ELPs or PLPs, we found the inflection point of the average FC dataset for both the cell lines. This was done by plotting the FC values versus gene rank, and determining the diagonal line of the curve from the end points. This line was then slid on the curve to see where it is tangential and that point was considered to be the inflection point (Vanhille et al., 2015).

2.2 Histone and Transcription factor CHIP-SEQ datasets

In order to find out which histone modifications and transcription factors are associated with ELPs in both the cell lines, we downloaded the CHIP-SEQ datasets from the ENCODE consortium (Davis et al., 2018) for both the cell lines in the BigWig format (Accession IDs : Ref. Table A1, A2), which were converted into to continuous BedGraph files using BEDTools (RRID:SCR_006646). Due to

the large size of the files, each file was fragmented chromosome wise using BEDOps (Neph et al., 2012) and mapping was done onto the promoter regions as described in section 2.1 using Bioconductor (Gentleman et al., 2004) in R. We converted the BedGraph data frame into Genomic ranges (GRanges) and took the mean of enrichment values across the length of the promoter region. These chromosomally segregated files for each histone modification and Transcription factor binding dataset were then combined using BEDTools. Finally, X and Y matrices were obtained for each cell line such that:



2.3 Quantile normalization of the datasets

Normalization is essential before any analysis which involves comparison of multiple datasets. Here, the final matrices (obtained after mapping the histone and transcription factor CHIP-SEQ datasets to the refseq promoter coordinates) were quantile normalized in R, using the preprocesscore library (Bolstad B 2021).

This method allows us to optimally compare and analyse datasets generated from different platforms. Following is the algorithm (Bolstad, Irizarry, Åstrand, &

Speed, 2003) for normalizing a set of data vectors by giving them the same distribution:

- i) Given n arrays of length p , form X of dimension $p \times n$ where each array is a column
- ii) Sort each column of X to give X_s
- iii) Take means across rows of X_s and assign the mean to each element in the row to get X'_s .
- iv) Get X_{norm} by rearranging each column of X'_s to have the same ordering as original X matrix.

2.4 Performing Partial Least Squares regression analysis

PLS is a great analysis tool for high-dimensional datasets and has been widely used for chemometric data analysis. It is a technique which reduces predictors into a smaller set of uncorrelated components, especially for such datasets where the predictors are often highly correlated with each other. We used the pls package (CRAN) in R to carry out PLS regression on our datasets. The algorithm is as follows (Mutalik & Venkatesh, 2005):

- i) First, the design matrix (X_{norm}) and the response matrix (Y) are centred to column mean zero, resulting in matrices X' and Y' .
- ii) Then, using linear dimension reduction $T = XR$, and the n predictors of X are mapped onto $c \leq \text{rank}(X') \leq \min(n, p)$ latent components in T ($p \times c$). SIMPLS algorithm was used to achieve this.
- iii) Assuming the model $Y' = TQ' + E$, Y' is regressed by Ordinary least squares regression against T (the X -scores) to get the loadings Q ($m \times c$), where $Q = Y'^T(T'T)^{-1}$.
- iv) Then, the estimate of coefficients B in $Y' = X'B + E$ is computed from estimates of the weight matrix R and Y -loadings Q via $B = RQ'$.

- v) Lastly, the coefficients for the original equation are computed by rescaling B.

2.5 Time course gene expression

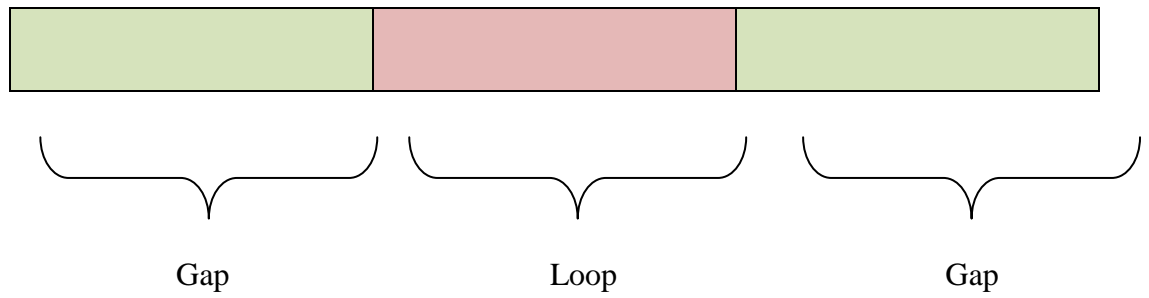
The time series RLE normalized CAGE dataset for K562 (subject to hemin stimulation) was obtained from FANTOM's TET (Table Extraction Tool) for the hg19 assembly (Lizio et al., 2019). The expression values were obtained for all the genes across regular time intervals.

Time course expression data for HeLa-S3 (subject to EGF stimulation) was obtained from NCBI's GEO database (Amit et al., 2007). This dataset consisted of expression values across time intervals using microarray expression profiling. Since the experiment was conducted on the 'GPL96 Affymetrix Human Genome U133A Array' platform, we used the Affy package in R Bioconductor to extract the time course gene expression matrix from the .CEL format. After extracting the matrix, the affymetrix ID for each gene was substituted with gene names using the GPL96 platform's annotation file. These time course matrices for both the cell lines were mapped onto the RefSeq promoters (ref section 2.1) in order to obtain segregated datasets for time course gene expression for genes associated with ELPs and PLPs.

2.6 Calculating enrichment in CTCF loops using CTCF CHIA-PET data

We used CTCF CHIA-PET data for K562 cell line from the ENCODE portal (Dunham et al., 2012), File accession : ENCFF001THV and obtained the CHIA-PET data for HeLa-S3 from NCBI's GEO database (Tang et al., 2015). From the CHIA-PET datasets for both the cell lines, we didn't consider Trans loops (inter chromosomal CHIA-PET interactions) and Cis (intra chromosomal CHIA-PET interactions) beyond 1 Mb, hence only Cis interactions with length < 1 Mb were taken into account as the number of observed interactions at or below this length significantly exceeds the number of interactions expected from stochastic interaction events or random ligation (DeMare et al., 2013).

In order to investigate the CTCF loops as well as their loop flanking regions, each loop was extended on either side by the size of the loop itself.



These extended regions were then divided into bins equal to 10% of the loop length and average enrichment of ELPs and PLPs was calculated in these regions.

Statistical analysis was done in R and Python was used for handling files.

Chapter 3

Results

3.1 Inflection point of the CapStarr-seq dataset

In order to classify the promoters of all the RefSeq defined human coding genes as Enhancer-Like Promoters (ELPs) or Promoter Like Promoters (PLPs), we plotted the average fold change values on the y-axis and rank-wise promoters on the x-axis to find the inflection point of the dataset. The promoters having a fold change value greater than or equal to this cut-off were defined as ELPs, whereas the ones having a fold change value less than the cut-off were defined as PLPs.

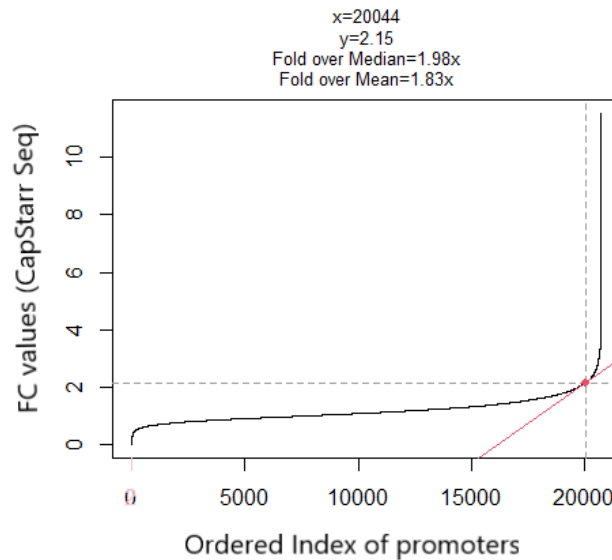


Fig. 3.1 (A) Inflection point of fold change values for Enhancer activity of promoters for the cell line HeLa-S3.

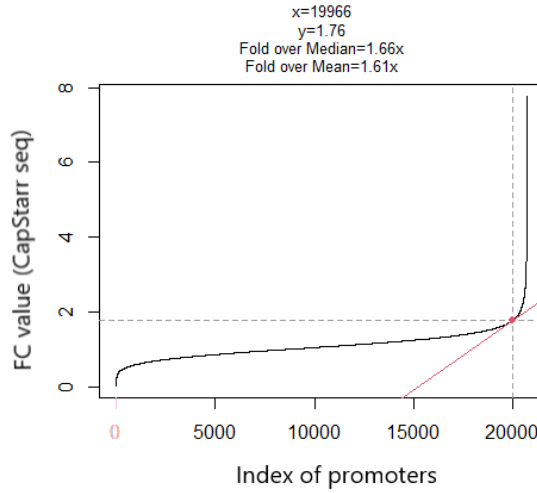


Fig.3.1 (B) Inflection point of fold change values for Enhancer activity of promoters for the cell line HeLa-S3.

We observed that the inflection point for HeLa-S3 turned out to be 2.15, with 676 ELPs and 20044 PLPs (~3.26 % ELPs). For the cell line K562, the inflection point was 1.76 with 754 ELPs and 19966 PLPs (~ 3.64 % ELPs).

3.2 Histone modifications and Transcription factors associated with ELPs

To look at the trends and correlations in the dataset graphically, we made correlograms (Hierarchical clustering method) using the X_{norm} and Y matrices (ref. section 2.3), which consist of the Histone modifications / Transcription factor enrichment, and FC values respectively, reflecting enhancer activity of the promoters respectively.

We observed that for HeLa-S3, enhancer activity showed high correlation with H3K27ac, H3K9ac, H3K4me3, H2AF.Z and H3K79me2 (Fig. 3.2(A)), whereas in the case of K562 we only observed very weak correlations among enhancer activity and histone modifications.

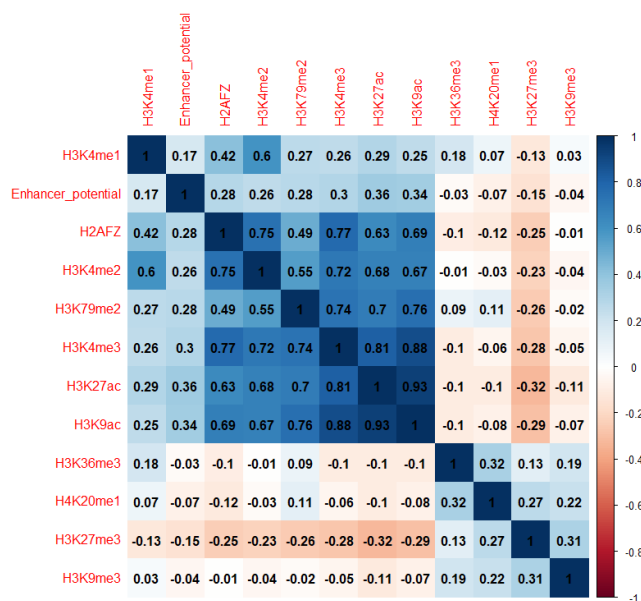


Fig. 3.2 (A)

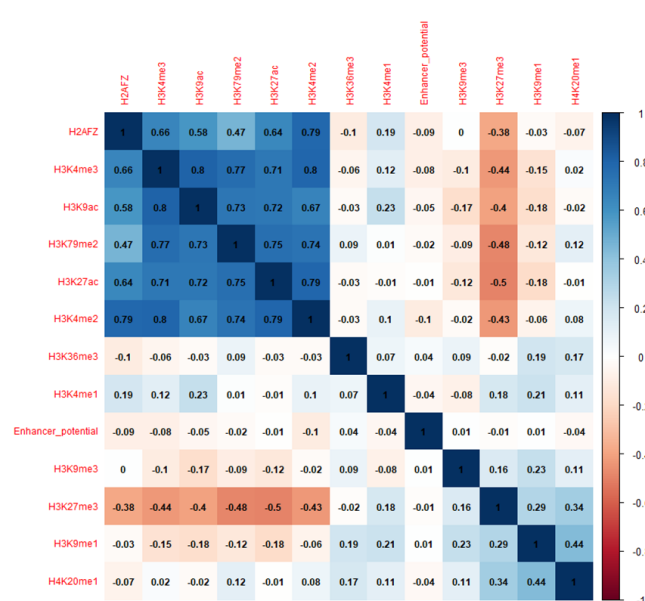


Fig. 3.2 (B)

Fig. 3.2 : Correlogram for various histone modifications and enhancer potential of promoters for (A) HeLa-S3 (B) K562

In figures 3.2 and 3.5, we see that some of the predictors (Histone modifications/ Transcription factors) are highly correlated with each other as well. We therefore used PLS regression, which reduces the predictors into a set of smaller set of uncorrelated components and then performs regression on these components. We made biplots with component 1 and 2, plotting the X loadings (Predictor variables) and Y loadings (response variables) so obtained with respect to Component 1 and 2 (Fig. 3.3). Since component 1 explained maximum variance of the response variable, we plotted the X-loadings along the first component for all the predictors to see which ones are best associated with enhancer activity of promoters (Fig 3.4). For HeLa-S3, H3K27ac, H3K9ac, H3K4me3, H2Af.Z and H3K4me2 turn out to be the best predictors with highest values for X-loadings. However, in case of K562, the trend is completely opposite as compared to HeLa-S3, wherein H3K27ac, H3K9ac, H3K4me3, H2Af.Z and H3K4me2, have magnitudes of X-loading similar to HeLa-S3 but they're negative. This is an interesting result; however we don't know how to explain it yet.

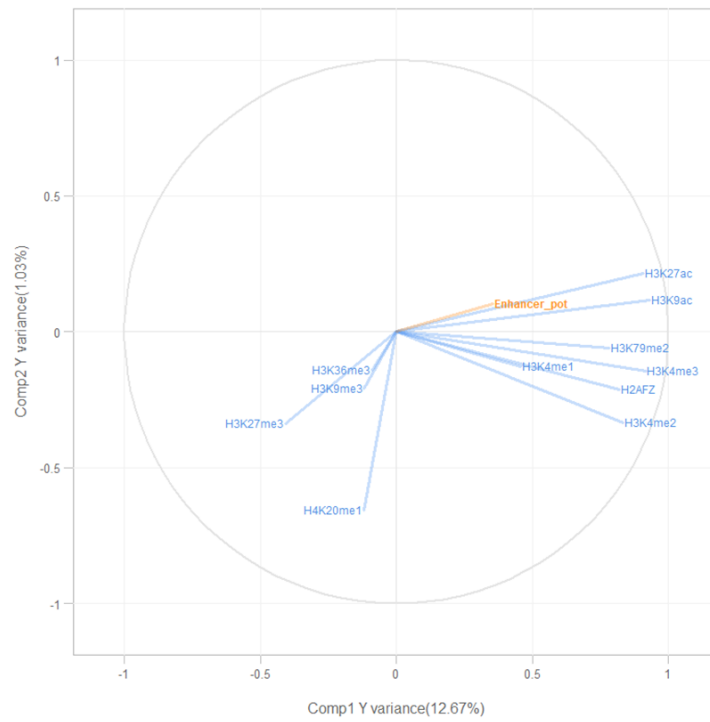


Fig 3.3 (A)

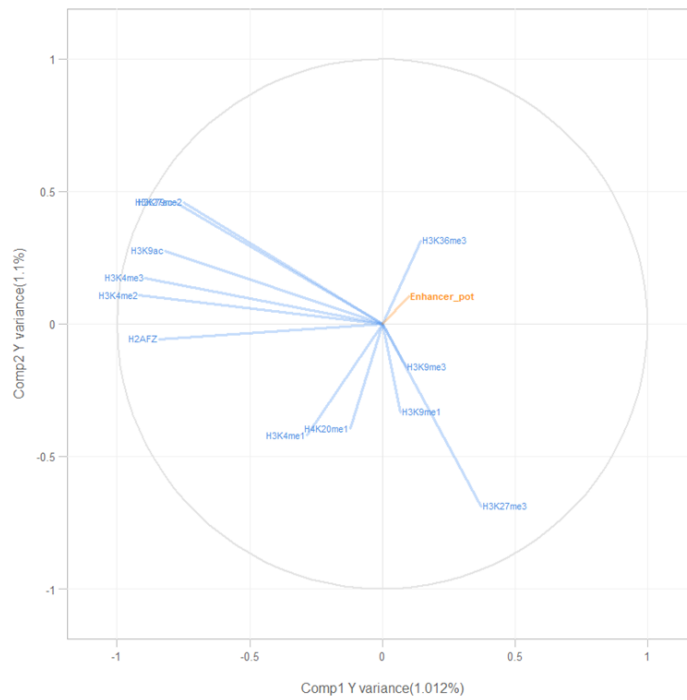


Fig. 3.3 (B)

Fig. 3.3 : Partial least squares regression biplots for PLSR performed on Histone modification matrix and Enhancer of ELPs in (A) HeLa-S3 (B) K562

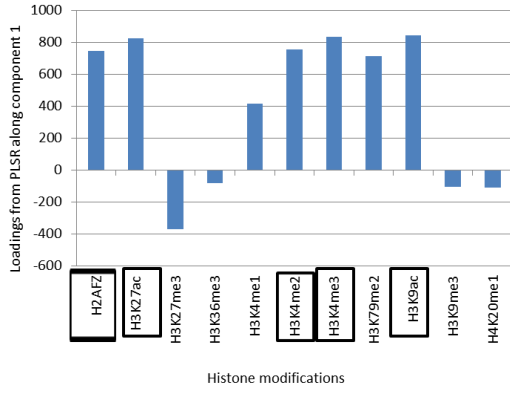


Fig. 3.4 (A)

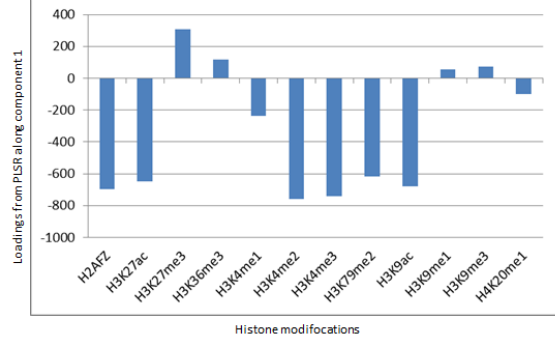


Fig. 3.4 (B)

Fig. 3.4 : Barplots of X loadings along component 1 for histone modifications in (A) HeLa-S3 and (B) K562

The same procedure as above was followed for Transcription factor datasets and it was found that in the case of HeLa-S3, transcription factors like ZHX1, EP300, NFE2L2, MAFF, TAF1, GABPA and POLR2A were correlated with the enhancer potential, and in K562, EP300 and STAT5A showed a relatively higher correlation than others (Fig 3.5).

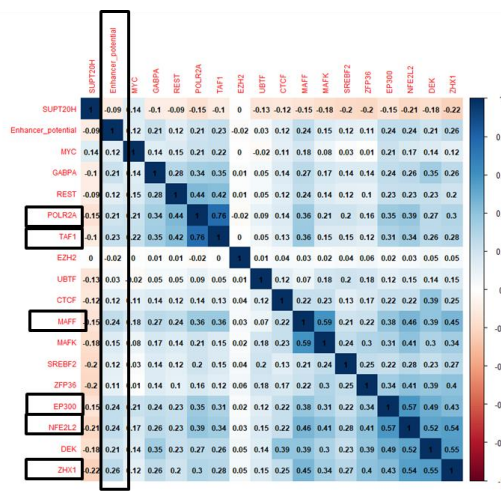


Fig. 3.5 (A)

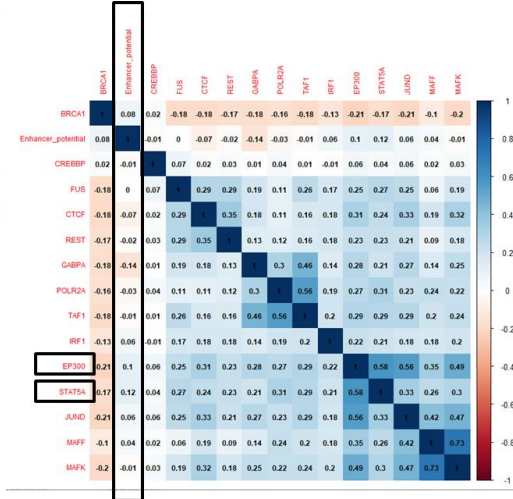


Fig. 3.5 (B)

Fig. 3.5: Correlation matrix for Transcription factors and enhancer potential of promoters for (A) HeLa-S3 (B) K562

From PLSR results, we obtained that NFE2L2, DEK, EP300, MAFF, ZHX1 are the key Transcription factors associated with ELPs in HeLa-S3, and EP300 and STAT5A were found to be important in the case of K562 (Fig. 3.6, 3.7).

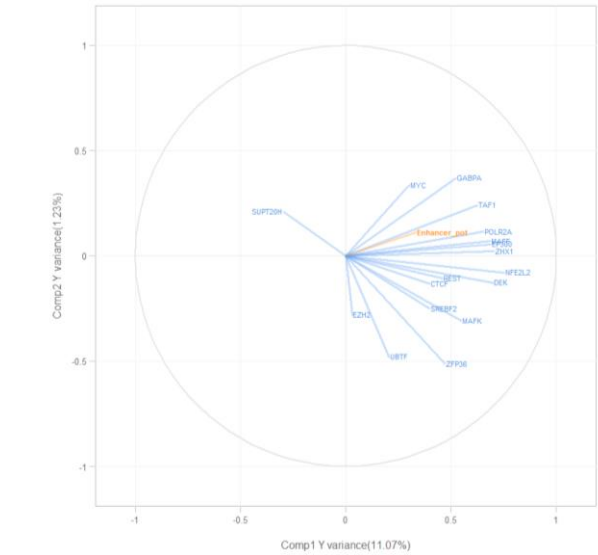


Fig. 3.6(A)

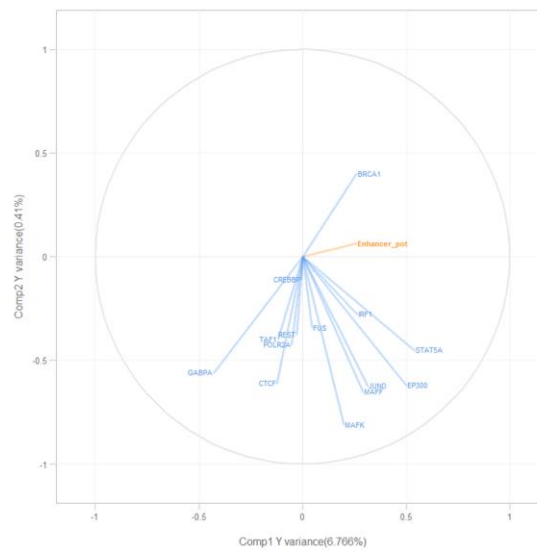


Fig. 3.6 (B)

Fig. 3.6 : Partial least squares regression biplots for PLSR performed on Transcription factor matrix and Enhancer potential of ELPs in (A) HeLa-S3 (B) K562.

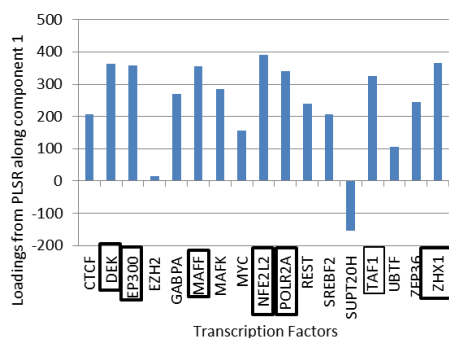


Fig. 3.7 (A)

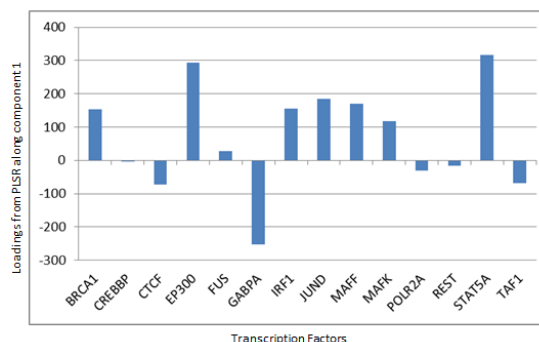


Fig. 3.7 (B)

Fig. 3.7: Barplots of X loadings along component 1 for Transcription factors in (A) HeLa-S3 and (B) K562

3.3 ELPs are inducible promoters and are associated with early response genes

Having obtained time course gene expression data for both the cell lines, we divided the dataset into two – for ELPs and PLPs. For both ELP and PLP datasets, we calculated Average fold change expression for all time points across promoters. It was observed that ELPs are more inducible than PLPs in both cell lines and are associated with early expressing genes (Fig. 3.8).

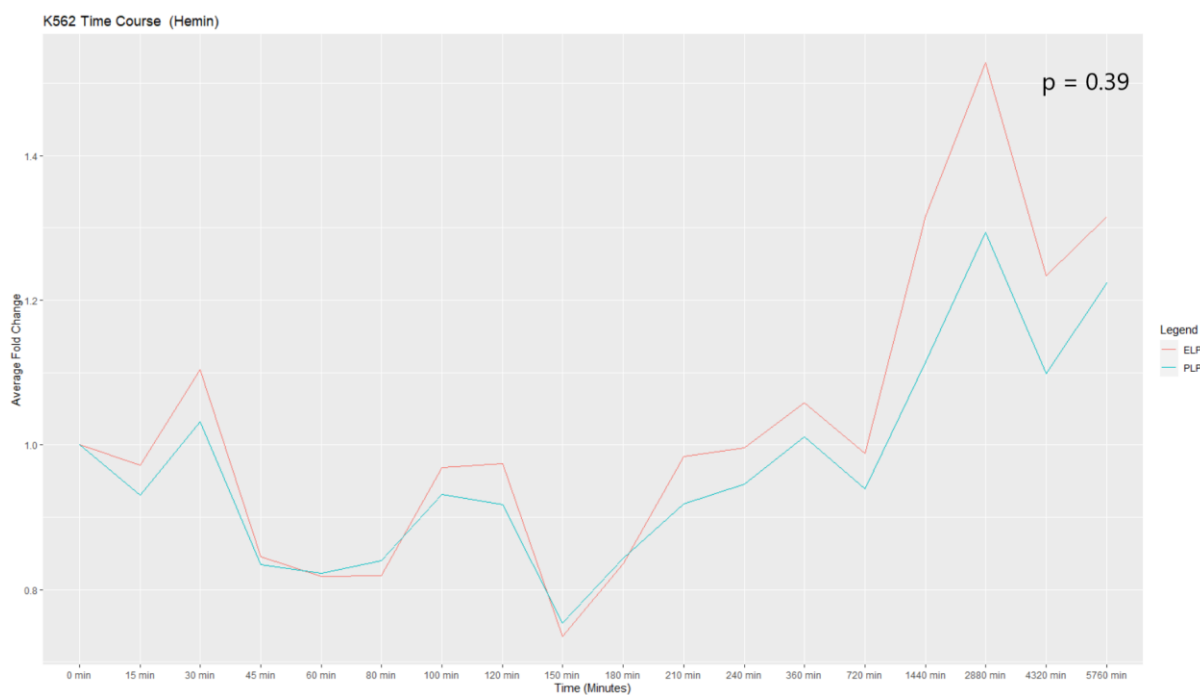


Fig. 3.8 (A) : Time course gene expression plots of ELPs and PLPs for K562 cell line stimulated by Hemin. Wilcoxin signed rank test was used to test the difference between them.

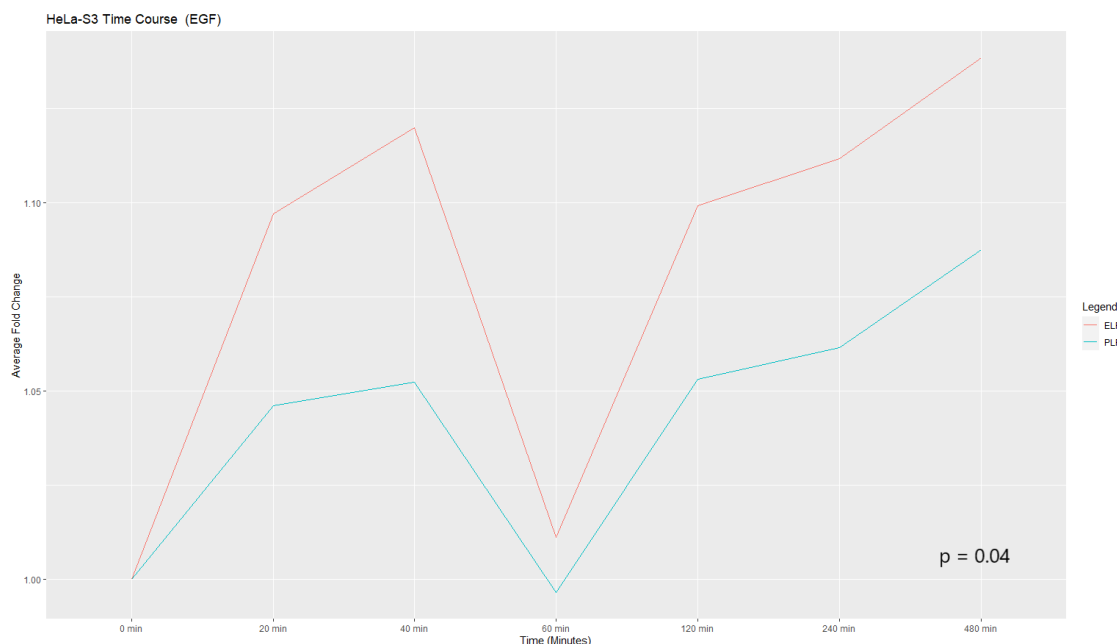


Fig. 3.8 (B): Time course gene expression data of ELPs and PLPs for HeLa-S3 cell line stimulated by EGF. Wilcoxin signed rank test was used to test the difference between them.

We further wanted to understand what could be the possible action mechanism of these ELPs. Since they're associated with early expressing genes, we wanted to check if they're involved with genes coding for transcription factors. We pulled down data for genes coding for transcription factors from FANTOM to find out what proportion of ELPs and PLPs are associated with TF coding genes. We found that there is no significant difference in either cell lines (Fig. 3.9). Hence, we conclude that ELPs act through chromatin interactions instead of gene products.

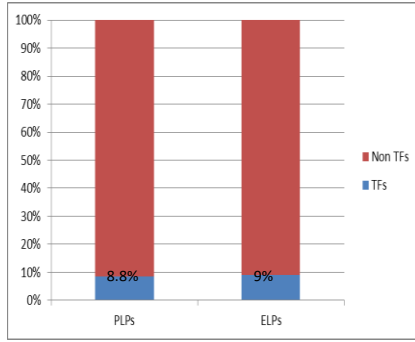


Fig. 3.9 (A)

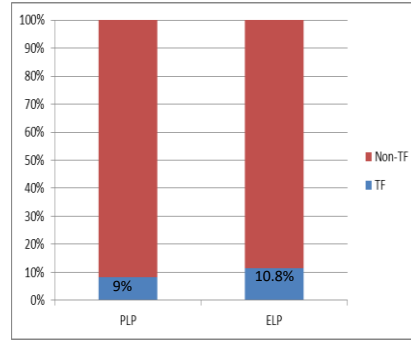


Fig. 3.9 (B)

Fig. 3.9: Stacked bar plots showing the proportion of ELPs and PLPs associated with Transcription factor encoding genes for (A) K562 and (B) HeLa-S3

3.4 ELPs are enriched inside CTCF loops

From our previous results, we found out that ELPs are inducible promoters, have enhancer potential, and act through chromatin-chromatin interactions; we wanted to check if they are restricted by boundary elements. Enrichment of ELPs and PLPs was calculated using the method as described in section 2.6, and the results indicate that ELPs are enriched in CTCF loops in both the cell lines.

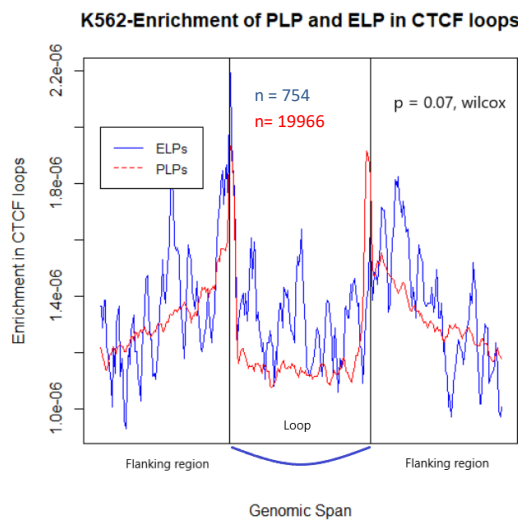


Fig. 3.10 (A)

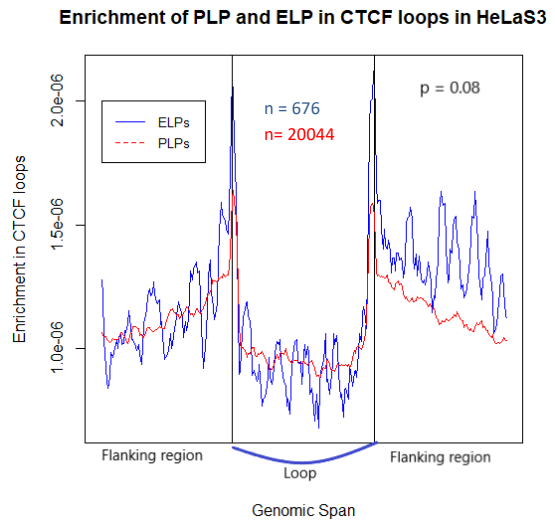


Fig. 3.10 (B)

Fig. 3.10: Plot for enrichment of ELPs and PLPs within and outside CTCF loops for (A) K562 and (B) HeLa-S3. Plots were made using sliding averages (window = 5) across 300 bins. p-values were calculated by taking 20% of the loop length on either boundaries and applying the wilcoxin signed rank test in both the cases.

Chapter 3

Discussion

From these results, it is evident that in both the cell lines – HeLa-S3 and K562, Enhancer like promoters are more inducible as compared to regular promoters when stimulated by environmental factors. These are also associated with early response genes which are among the first ones to be transcribed during the development of an organism and in a way kick-start the developmental program. They initiate a downstream wave of transcriptional response and pave way for other genes to be expressed sequentially.

Enhancer-like promoters seem to affect the transcription of other downstream genes through promoter-promoter interactions, as the case for their action mechanism through protein products (i.e., via transcription factors) was ruled out. There seemed to be virtually no difference between the proportion of ELPs or PLPs which were associated with genes encoding for transcription factors, hence reaffirming that these ELPs indeed act as true enhancers.

ELPs are also observed to be highly enriched within the boundaries of CTCF loops throughout the genome. CTCF protein is involved in organisation of the genome and also mediates three dimensional chromatin interactions. Enrichment of ELPs inside CTCF loops could be facilitating their action mechanism in two ways. One - that the target promoters with which ELPs interact don't necessarily reside in their vicinity and their early transcription could lead to a ripple effect propagating in the genome which would mean untimely transcription of other genes. The fact that ELPs are bound inside CTCF loops prevents this ripple effect as CTCF act as insulator proteins. Second – CTCF loops also facilitate the interaction of distal genomic elements in space hence bringing ELPs and their actual targets closer.

Histone modifications like H3K27ac, H3K9ac, H3K4me3, H2Af.Z and H3K4me2 were found to be associated with enhancer potential of ELPs in HeLa-S3. H3K4me1, which is considered to be the hallmark of enhancers, did not turn out to be strongly associated with ELPs however H3k4me2 which is again a mark for enhancers (Lokody, 2013) was one of the strong predictors according to PLSR. H3K27ac is known to be associated with active regulatory elements and H3K4me3 is a mark associated with promoters. The case for K562 was a bit curious, because the trends of histone modifications were completely opposite as compared to HeLa-S3. Transcription factors like NFE2L2, DEK, EP300, MAFF, ZHX1 were found to be best associated with ELPs in HeLa-S3, and EP300 and STAT5A in the case of K562. These Transcription factors have been known to be associated with enhancers in literature. It is difficult to extrapolate these associations based on these two datasets because of lack of coherence in the results. More data from various other cell-lines need to be analysed to see whether trends exist or if these associations are dependent on the nature of biological samples used.

Appendix

| | |
|-------------|----------|
| ENCFF527JUP | H3K9ac |
| ENCFF439ZCI | H3K4me1 |
| ENCFF526UWC | H3K9me1 |
| ENCFF834YLI | H3K9me3 |
| ENCFF689TMV | H3K4me3 |
| ENCFF486DJY | H3K27me3 |
| ENCFF745HXR | H3K36me3 |
| ENCFF143CUR | H4K20me1 |
| ENCFF003CLZ | H3K79me2 |
| ENCFF010PHG | H3K27ac |
| ENCFF191EXE | H2AFZ |
| ENCFF118MMT | H3K4me2 |
| ENCFF279WBA | BRCA1 |
| ENCFF813TKO | CREBBP |
| ENCFF933ZLL | CTCF |
| ENCFF357GGL | EP300 |
| ENCFF380ABG | FUS |
| ENCFF869WSZ | GABPA |
| ENCFF992YWK | IRF1 |
| ENCFF223ARA | JUND |
| ENCFF931BLV | MAFF |
| ENCFF964KPV | MAFK |
| ENCFF348UKA | POLR2A |
| ENCFF571ECM | REST |
| ENCFF832FFP | STAT5A |
| ENCFF488POX | TAF1 |

Table A1: Encode Accession IDs
Histone and TF CHIP-SEQ
experiments.

| | |
|-------------|----------|
| ENCFF442LQQ | H2AFZ |
| ENCFF131OIJ | H3K27ac |
| ENCFF837PIH | H3K27me3 |
| ENCFF236IFZ | H3K36me3 |
| ENCFF065HMK | H3K4me1 |
| ENCFF790RUR | H3K4me2 |
| ENCFF045NNJ | H3K4me3 |
| ENCFF216CWP | H3K79me2 |
| ENCFF019ETS | H3K9ac |
| ENCFF603RCV | H3K9me3 |
| ENCFF094AQR | H4K20me1 |
| ENCFF980IVM | ZFP36 |
| ENCFF078EDJ | EZH2 |
| ENCFF308NHO | CTCF |
| ENCFF428DLR | MAFF |
| ENCFF630LNK | SREBF2 |
| ENCFF739HNF | MAFK |
| ENCFF710PXV | REST |
| ENCFF893YQO | UBTF |
| ENCFF721OKD | ZHX1 |
| ENCFF043NAS | GABPA |
| ENCFF776MXK | DEK |
| ENCFF701YXF | TAF1 |
| ENCFF796RLG | EP300 |
| ENCFF522IZP | MYC |
| ENCFF360MQA | SUPT20H |

Table A2: Encode Accession IDs for K562 for
HeLa S3 Histone and TF CHIP-SEQ
experiments.

BIBLIOGRAPHY

- Amit, I., Citri, A., Shay, T., Lu, Y., Katz, M., Zhang, F., ... Yarden, Y. (2007). A module of negative feedback regulators defines growth factor signaling. *Nature Genetics*, 39(4), 503–512. <https://doi.org/10.1038/ng1987>
- Atchison, M. L. (1988). Enhancers: Mechanisms of action and cell specificity. *Annual Review of Cell Biology*, 4, 127–153. <https://doi.org/10.1146/annurev.cb.04.110188.001015>
- Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), 185–193. <https://doi.org/10.1093/bioinformatics/19.2.185>
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., ... Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), 21931–21936. <https://doi.org/10.1073/pnas.1016071107>
- Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., ... Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, 49(7), 1073–1081. <https://doi.org/10.1038/ng.3884>
- Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., ... Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1), D794–D801. <https://doi.org/10.1093/nar/gkx1081>
- DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R., & Noonan, J. P. (2013). The genomic landscape of cohesin-Associated chromatin interactions. *Genome Research*, 23(8), 1224–1234. <https://doi.org/10.1101/gr.156570.113>
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. <https://doi.org/10.1038/nature11247>
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... Zhang, J. (2004). Bioconductor: open software development for computational biology and

- bioinformatics. *Genome Biology*, 5(10). <https://doi.org/10.1186/gb-2004-5-10-r80>
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., ... Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3), 311–318. <https://doi.org/10.1038/ng1966>
- Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., ... Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295), 182–187. <https://doi.org/10.1038/nature09033>
- Kulzer, J. R., Stitzel, M. L., Morken, M. A., Huyghe, J. R., Fuchsberger, C., Kuusisto, J., ... Mohlke, K. L. (2014). A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *American Journal of Human Genetics*, 94(2), 186–197. <https://doi.org/10.1016/j.ajhg.2013.12.011>
- Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., ... Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1–2), 84–98. <https://doi.org/10.1016/j.cell.2011.12.014>
- Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C. C., ... Kawaji, H. (2019). Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Research*, 47(D1), D752–D758. <https://doi.org/10.1093/nar/gky1099>
- Lokody, I. (2013). Gene regulation: Chromatin editing reveals enhancer targets. *Nature Reviews Genetics*, 14(11), 749. <https://doi.org/10.1038/nrg3601>
- Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195. <https://doi.org/10.1126/science.1222794>
- Mutalik, V. K., & Venkatesh, K. V. (2005). Quantification of the glycogen cascade system: The ultrasensitive responses of liver glycogen synthase and muscle phosphorylase are due to distinctive regulatory designs. *Theoretical Biology and Medical Modelling*, 2, 1–12. <https://doi.org/10.1186/1742-4682-2-19>
- Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., ... Stamatoyannopoulos, J. A. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics*, 28(14), 1919–1920. <https://doi.org/10.1093/bioinformatics/bts277>

- Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., ... Sharp, P. A. (2008). Divergent Transcription from Active Promoters. *Science*, 322(5909), 1849 LP – 1851. <https://doi.org/10.1126/science.1162253>
- Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., ... Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7), 1611–1627. <https://doi.org/10.1016/j.cell.2015.11.024>
- Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., ... Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, 6, 1–29. <https://doi.org/10.1038/ncomms7905>