



Faculty of Resource Science and Technology

**Identification and Characterization of Liver-Specific Enhancer Elements
in Chromosome 18 of the *Nasalis larvatus* Genome**

Lim Whye Kit Leonard

**Master of Science
2019**

Identification and Characterization of Liver-Specific Enhancer Elements in
Chromosome 18 of *Nasalis larvatus* Genome

Lim Whye Kit Leonard

A thesis submitted

In fulfilment of the requirements for the degree of Master of Science

(Molecular Biology)

Faculty of Resource Science and Technology
UNIVERSITI MALAYSIA SARAWAK
2019

DECLARATION

I hereby declare that conclusions made based on materials collected by me and this dissertation are my original work. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Lim Whye Kit Leonard

16020176

Faculty of Resource Science and Technology

UNIVERSITI MALAYSIA SARAWAK

March 2019

ACKNOWLEDGEMENT

This thesis has finally come to fruition with the help and support I have received throughout the execution of this research project. There are a few individuals that I wanted to thank personally below.

First of all, I thank God for all the wisdom, strength and good health that have been granted to me in the time of needs. My deepest gratitude goes to my supervisor, Dr Chung Hung Hui and co-supervisor, Dr Chong Yee Ling for their continuous guidance and insights, my intellectual debts are to them for their relentless sharing of knowledge and advices. Special thanks to UNIMAS for supporting my studentship for the past two years under the ZAMALAH scholarship. My sincere appreciation goes to Dr Lee Nung Kion for his countless support in terms of knowledge and materials.

Next, I would like to extend my heartfelt gratitude to my Animal Biotechnology Lab mates: Aimi, Jill, Cynthia, Lee Ying, Vivien and Yuki for their aid, companionship and knowledge sharing. They are the wonderful joy and pillars of strength for my delightful lab experiences throughout these two years. I am grateful to have received various forms of motivations from my peers: Delvino, John, Carlson, Anarrin, Aina, Mugu, Ivy and Zy Ying.

Last but not least, I am highly indebted to my family: my dad Lim Loong Kien, my mum Liew Hong Yen and my sister Lim Zhi Vern Joanne for their selfless prayers and financial support. This work was fully supported by Fundamental Research Grant Scheme (FRGS) from the Ministry of Education Malaysia (Grant number: FRGS/1/2014/SG03/UNIMAS/02/2). The collection of samples was conducted using permit number NCCD.907.4.4(JLD.11) and park permit number 537/2014 and 538/2014.

ABSTRACT

Enhancers play major roles in the regulation of gene expression at various developmental stages and biological processes temporally and spatially. Albeit the first enhancer was discovered for more than four decades ago, the identification of enhancers is still at its infancy as there is no single universal enhancer mark revealed to date. With the booming emergence of genome sequences of various non-human primates, the study of enhancer landscape across these primates had been made possible recently. Therefore, this study aimed to identify, isolate and characterize enhancers from the genome of one of the most ancient colobine primate, the proboscis monkey (*Nasalis larvatus*). A total of 13 liver-specific enhancers were identified using three different *in silico* approaches and they were inserted into expression plasmids before tested further for their activities in HepG2 and A549 cell lines. Two of the strong enhancers and one weak enhancer were subjected to deletion analysis and activity determination subsequently in both cell lines. This study has discovered several clusters of adjacently located transcription factor binding sites that influence enhancer activities and tissue specificities greatly. The results from this research will drive future investigations on the enhancer landscape of the proboscis monkey and comparison of the landscape across different primates.

Keywords: Enhancer, identification, proboscis monkey, *in silico*, expression

Identifikasi dan klasifikasi elemen-elemen penambah yang berkhhusus dalam hati dari kromosom 18 genom *Nasalis larvatus*

ABSTRAK

*Penambah memainkan peranan utama dalam pengawalan ekspresi gen dipelbagai peringkat pertumbuhan dan proses biologi secara temporal dan spatial. Walaupun penambah pertama dijumpai sejak empat dekad dahulu, identifikasi terhadap penambah masih berada di fasa muda kerana masih tiada tanda penambah yang tunggal dan universal didedahkan sampai kini. Dengan kemunculan urutan genom daripada pelbagai primata bukan manusia secara mendadak, penyelidikan landskap penambah merentasi primata tersebut dimungkinkan akhir-akhir ini. Justeru itu, penyelidikan ini bertujuan mengenalpasti, mengasingkan dan mencirikan penambah-penambah daripada genom primata colobine yang paling kuno, iaitu kera bekantan (*Nasalis larvatus*). Sebanyak 13 penambah yang berkhhusus dalam hati telah dikenalpasti melalui tiga kaedah in silico yang berbeza dan mereka dimasukkan ke dalam plasmid ekspresi sebelum aktiviti mereka diuji di dalam sel HepG2 dan A549. Dua penambah kuat dan satu penambah lemah telah dijalani analisis penghapusan dan aktiviti mereka turut diuji dalam kedua-dua sel tersebut. Kajian ini telah menemui beberapa kumpulan tapak pengikat factor transkripsi bersebelahan yang mempunyai pengaruh hebat terhadap aktiviti penambah dan kekhususan tisu. Keputusan daripada kajian ini akan memacu penyiasatan terhadap landskap penambah kera bekantan dan perbandingan landskap tersebut antara primata yang berbeza.*

Kata kunci: *Penambah, identifikasi, kera bekantan, in silico, ekspresi*

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
<i>ABSTRAK</i>	iv
TABLE OF CONTENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: LITERATURE REVIEW	4
2.1 Regulation of gene expression	4
2.1.1 Gene regulation in eukaryotes	5
2.2 <i>Cis</i> -regulatory elements	7
2.2.1 Enhancer	9
2.2.2 Enhancer mechanism	10
2.3 Identification of enhancer	12
2.3.1 Reporter assay approach	13
2.3.2 High-throughput assay approach	14
2.3.3 <i>In silico</i> approach	15
2.3.3.a Chromatin signature identification by artificial neural network (CSI-ANN)	18

2.3.3.b Gapped k-mer support vector machine (gkm-SVM)	19
2.3.3.c DeepBind	19
2.3.3.d iEnhancer-2L	20
2.3.3.e Gapped motif feature representation-convolutional neural networks (GMFR-CNN)	21
2.4 Liver and its main transcription factors	21
2.4.1 AP-1	23
2.4.2 C/EBP- β	23
2.4.3 GATA-3	24
2.4.4 HNF-1	25
2.4.5 HNF-3 Beta	26
2.4.6 NF-1	26
2.5 Significance of non-human primates in biomedical researches	27
2.6 Enhancers in primates	28
2.7 Proboscis monkey and its significance	29
CHAPTER 3: RESEARCH METHODOLOGY	32
3.1 Enhancer identification	32
3.1.1 Enhancer identification using iEnhancer-2L	32
3.1.2 Enhancer identification using conservation from other primates	32
3.1.3 Enhancer identification using 5 different enhancer predictor tools	33
3.1.4 Selection of enhancers	34
3.2 Primer design and synthesis	35
3.3 Stool DNA extraction using QIAamp® DNA Stool Mini Kit (Qiagen, Germany)	36

3.4 Gradient PCR temperature optimization	36
3.5 PCR product purification or gel extraction	38
3.6 Cloning of enhancers into pGEM®-T Easy Vector	38
3.6.1 Chemically competent cell preparation	38
3.6.2 Ligation of enhancer into pGEM®-T Easy Vector	39
3.6.3 Bacterial transformation into XLI-Blue <i>E. coli</i> competent cells	40
3.6.4 Blue/white primary colony screening and secondary screening	41
3.6.5 Colony PCR	41
3.6.6 Plasmid Miniprep	42
3.7 Subcloning of enhancer insert into pGL4.23 [luc2/minP] vector	43
3.7.1 Restriction digestion	43
3.7.2 Ligation of enhancer insert into pGL4.23 [luc2/minP] vector	44
3.7.3 Transformation, primary colony screening, secondary screening and colony PCR	44
3.7.4 Sequencing of plasmids	45
3.8 Cell culture maintenance	45
3.8.1 Culture medium preparation	45
3.8.2 Primary cell culture	46
3.8.3 Subculture of cell line	46
3.9 Transient transfection	48
3.9.1 Cell passaging	48
3.9.2 Cell counting	49
3.9.3 Transient transfection using Lipofectamine 2000	50
3.10 Dual-luciferase reporter assay	51

3.11 Motif distribution analysis	52
3.12 Transcription factor binding site analysis	53
3.13 Deletion analysis	53
CHAPTER 4: RESULTS	55
4.1 Enhancer identification	55
4.1.1 Enhancer identification using iEnhancer-2L	55
4.1.2 Enhancer identification using conservation from other primates	55
4.1.3 Enhancer identification using 5 different software packages	56
4.2 Selection of enhancers and primer design	61
4.3 Genomic DNA extraction	63
4.4 Amplification of enhancers	66
4.5 Cloning of enhancers	68
4.6 Motif distribution	68
4.7 Transcription factor binding site (TFBS) analysis	69
4.7.1 Overall TFBS abundance	70
4.7.2 TFBS composition	70
4.7.3 TFBS distribution pattern	73
4.8 Luciferase reporter assay	74
4.9 Deletion analysis	75
4.10 Comparison of enhancer activity across two cell lines	82
CHAPTER 5: DISCUSSION	84
5.1 <i>In silico</i> enhancer prediction and enhancer activity	84
5.2 Enhancer activity	86
5.3 Motif distribution and enhancer activity	87

5.4 Transcription factor binding site (TFBS) and enhancer activity	88
5.4.1 Overall TFBS abundance and enhancer activity	89
5.4.2 TFBS composition and enhancer activity	91
5.4.3 TFBS distribution pattern and enhancer activity	92
5.5 Deletion analysis	95
5.6 Comparison of enhancer activity across two cell lines	98
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS	100
6.1 Conclusions	100
6.2 Recommendations	101
REFERENCES	103
APPENDICES	116

LIST OF TABLES

		Page
Table 2.1	A summary of features considered in <i>in silico</i> enhancer predictions (Cao & Yip, 2016).	17
Table 3.1	PCR reaction components for enhancer isolation and amplification from the proboscis monkey genome.	37
Table 3.2	Thermal cycling parameters used for PCR reactions.	37
Table 3.3	PCR reaction components for <i>cytochrome b</i> gene amplification from the proboscis monkey genome.	38
Table 3.4	List of ligation components for enhancer PCR products.	40
Table 3.5	List of colony PCR reaction mixtures for enhancer insert	42
Table 3.6	List of restriction digestion reagents.	43
Table 3.7	List of ligation mixtures for enhancer insert.	44
Table 4.1	The outcome of iEnhancer-2L depicting the number of strong, weak and non-enhancers from chromosome 18 of <i>N. larvatus</i>	55
Table 4.2	Bedtools “merge” results.	57
Table 4.3	Bedtools “intersect” outcomes for 2 software.	58
Table 4.4	Bedtools “intersect” outcomes for 3 software.	59
Table 4.5	Bedtools “intersect” outcomes for 4 software.	60
Table 4.6	Bedtools “intersect” outcomes for 5 software.	61
Table 4.7	The TFBS analysis outcome of individual candidate enhancers.	62
Table 4.8	Gene-specific primers designed to isolate enhancers and their characteristics.	63
Table 4.9	The optimum annealing temperatures for each enhancer.	67
Table 4.10	The enumeration of all transcription factors predicted within candidate enhancer sequences.	72
Table 4.11	The list of internal primers used for enhancer fragments amplification via PCR.	77

LIST OF FIGURES

		Page
Figure 2.1	The overview of complex gene regulation with various regulatory stages.	5
Figure 2.2	The mechanistic actions of enhancer-promoter interactions in transcription initiation involving various general transcription factors, multi-subunit cofactors and chromatin remodelling complexes	7
Figure 2.3	Gene regulation at chromatin level (Adapted from Zaidi <i>et al.</i> , 2010).	7
Figure 2.4	An overview of <i>cis</i> -regulatory elements that regulates transcription activities (Adapted from Levine & Tjian, 2003).	8
Figure 2.5	The four enhancer mechanism models where the hexagon represents transcription factor, grey oval represents polymerase, and grey halo represents nuclear subcomponents (Adapted from Kolovos <i>et al.</i> , 2012).	11
Figure 3.1	Gridlines for haemocytometer.	49
Figure 4.1	The enhancer alignment between human, chimpanzee and rhesus macaque for enhancer pme001.	56
Figure 4.2	Genomic DNA extraction from stool of proboscis monkey. Lane L: TransGen 1 kb DNA ladder; Lane 1-4: Genomic DNA extracted from the same stool sample (four technical replicates).	64
Figure 4.3	Amplification of housekeeping gene, cytochrome b gene. Lane L: TransGen 1 kb DNA ladder; Lane 1: <i>cytochrome b</i> gene (~1500 bp) amplified at 50.5°C annealing temperature.	65
Figure 4.4	Part of the pairwise aligned sequences from BLASTn analysis of cytochrome b gene of proboscis monkey.	65
Figure 4.5	PCR amplification of enhancers. Lane L: TransGen 100 bp ladder; Lane 1-4: pme001-pme004; Lane 5-7: pme101-pme103.	66
Figure 4.6	PCR amplification of enhancers. Lane L: TransGen 100 bp ladder; Lane 1 and 2: pme104 and pme105; Lane 3-6: pme201-pme204.	67
Figure 4.7	The motif distribution of each enhancer.	69

Figure 4.8	The predicted TFBS frequency per nucleotide bar graph across 13 enhancers.	70
Figure 4.9	The breakdown of predicted TFBS amount across 13 enhancers.	71
Figure 4.10	The TFBS distribution per 100 bp interval of all 13 enhancers	74
Figure 4.11	The measured normalized relative enhancer luciferase activity (relative to NIC=1) across 13 enhancers	75
Figure 4.12	PCR amplification of enhancer fragments.	78
Figure 4.13	Overview of selected TFBS distribution and deletion analysis of enhancer pme001 in pGL4.23 [<i>luc</i> /minP], activities as measured in HepG2 cell line.	79
Figure 4.14	Overview of selected TFBS distribution and deletion analysis of enhancer pme101 in pGL4.23 [<i>luc</i> /minP], activities as measured in HepG2 cell line.	80
Figure 4.15	Overview of selected TFBS distribution and deletion analysis of enhancer pme103 in pGL4.23 [<i>luc</i> /minP], activities as measured in HepG2 cell line.	81
Figure 4.16	The normalized relative luciferase unit (relative to NIC=1) of enhancers in both HepG2 and A549 cell lines.	82
Figure 4.17	The normalized relative luciferase unit (relative to NIC=1) of enhancer fragments in both HepG2 and A549 cell lines.	83

LIST OF ABBREVIATIONS

bp	Base pair
CO ₂	Carbon dioxide
DMSO	Dimethylformamide
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide phosphate
kb	Kilo base pair
LB	Luria Bertani
MEM	Minimum Essential Medium
Mg ²⁺	Magnesium ion
NIC	Non-Insert Control
OD	Optical Density
OptiMEM	Reduced-Serum MEM
PCR	Polymerase Chain Reaction
PenStrep	Penicillin Streptomycin
pme	Proboscis monkey enhancer

RLU

Relative Luciferase Unit

CHAPTER 1

INTRODUCTION

The major discovery of the influence of gene regulatory sequences on divergent evolutionary relationship in the animal kingdom by King and Wilson (1975) for over 40 years ago had greatly change the way we look at non-coding regions, especially cis-regulatory elements. One of the most popular cis-acting elements that could possibly create the evolutionary phenotypic differences in divergence is the enhancer. Enhancers are cis-regulatory elements that can increase gene transcription of adjacent and distant genes in the genome. Enhancers modulate expressions of genes at various developmental stages and they are important indicators to several biological processes to date (Kleftogiannis *et al.*, 2015). They are very vital in terms of development processes and are deemed to be one of the vehicles of evolution and natural selection.

Genome-wide enhancer identification is one of the most effective and comprehensive way to fully discover the enhancer landscape of a species. However, this has been a challenging task if it is conducted experimentally due to its cost- and labour-intensive methodologies. Furthermore, there is no single universal enhancer “golden standard” mark identified to date to accurately identify them from the genome (Lim *et al.*, 2018). The *in silico* approaches have speed up the path in predicting enhancers using various enhancer marks albeit some minor limitations and they are usually conducted before downstream experimental approaches are applied. Enhancer identification has gained more importance in this new genomic era as multiple lines of evidences has pointed out the link between enhancer sequences and genetic diseases, especially in humans (Kim *et al.*, 2011; Corradin & Scacheri, 2014; Miguel-Escalada *et al.*, 2015; Zhang *et al.*, 2018).

The liver is an essential organ found only in vertebrates and it plays major roles in biological processes such as detoxification of various metabolites, biochemical factory for the digestive system, hormone and protein synthesis, as well as decomposition of red blood cells. The role of liver in mediating drug responses has been extensively studied as adverse drug events has been one of the major causes of death in this century (Kim *et al.*, 2011). Studies have found that not only the nucleotide variations in protein-coding genes can lead to the heterogeneity in drug responses but the nucleotide variations in non-coding elements (including SNPs) as well as possibly the enhancers also had made substantial contributions to variability in drug responses (Kim *et al.*, 2011; Corradin & Scacheri, 2014; Miguel-Escalada *et al.*, 2015; Zhang *et al.*, 2018).

A wide range of prescription drugs were designed to target membrane transporters because these transporters are the major mediators in absorption, dissemination, removal, as well as translocation of various xenobiotics (Kim *et al.*, 2011). Liver membrane transporters play roles in the uptake and elimination of most drugs and xenobiotics in the liver. Previous studies have shown that nucleotide variants of promoters can lead to alteration in the transporter transcription and thus the affected gene expression resulting in modifications on the pharmacokinetic and pharmacodynamic properties of a drug (Sun *et al.*, 2010; Kim *et al.*, 2011). Furthermore, Kim *et al.* (2011) had postulated that enhancer nucleotide variations can too lead to similar outcomes.

Similar to what other researchers had done in the past on coding regions of the genome of primates for genetic diseases detection and treatment, this study was conducted on liver-specific enhancers from one of the most ancient primates in the world, the proboscis monkey (*Nasalis larvatus*). This primate is endemic to Borneo Island and was proven to be one of its kind in terms of its primitive chromosome number ($2n=48$) especially (Chiarelli, 1966; Soma

et al., 1974; Stanyon *et al.*, 1992). To date, there is no genome-wide enhancer study and enhancer activity determination on proboscis monkey (*Nasalis larvatus*) reported since the completion of the proboscis monkey genome sequencing project in the year 2014 (BioSpectrum Bureau, 2010; Abdullah *et al.*, 2014). Moreover, with the commencement of genome sequencing of various primates to date, it is interesting to shed light on the landscape of enhancers in these primates via comparative studies, which will in turn contributes to conservation researches. Furthermore, some enhancers are known to have evolutionarily acquiring significant functions following the speciation events and some of them are highly related to variable drug responses and onset of genetic diseases (Kleftogiannis *et al.*, 2015). Therefore, it is interesting look onto the gene regulations and enhancers in one of the most primitive colobine, the proboscis monkey, particularly in the liver, to discover the uniqueness of the enhancer landscape in this species and further functionally characterize these enhancers based on liver-specific transcription factor binding sites (TFBSs). It is hypothesized that these liver-specific enhancers identified via *in silico* methods can functionally express via reporter assays approach.

Hence, this study was aimed:

- i. To identify liver-specific enhancers from the chromosome 18 of proboscis monkey using *in silico* approaches.
- ii. To validate the expression of liver-specific enhancers of proboscis monkey in eukaryotic cell lines HepG2 and A549.
- iii. To identify determinants of strong enhancer activity via deletion analysis.

The limitations of this study are the enhancer scarcity in terms of genome coverage and the lack of *in vivo* studies, in which these are listed in the recommendations section for further studies.

CHAPTER 2

LITERATURE REVIEW

2.1 Regulation of gene expression

Gene expression is regulated by a broad range of mechanisms that controls the amount of gene products produced spatially and temporally. The regulation of gene expression beyond the gene level (RNA or protein level) is essential for various biological pathways and developmental stages to ensure the survivability of the host organism. The process of morphogenesis and cellular differentiation are greatly dependent on the regulation of the gene expression at the correct time and location, especially in multicellular organisms (Carlberg & Molnár, 2014; Jakubowski, 2016).

The gene expression can be regulated in various ways and at various control stages, from chromatin remodelling, transcription initiation, transcription, transcript processing and modifications, mRNA degradation, translational initiation, translation, posttranslational modifications to the stage of protein transport and protein degradation (Figure 2.1). The transcription regulation particularly, is one of the most utilized stages in gene regulation. This regulation of transcription is responsible for the management of the amount of RNA copies in a particular cell or tissue to ensure that the requirements of the host body are met, and it also enables the host to adapt and respond to environmental stimuli (Carlberg & Molnár, 2014).

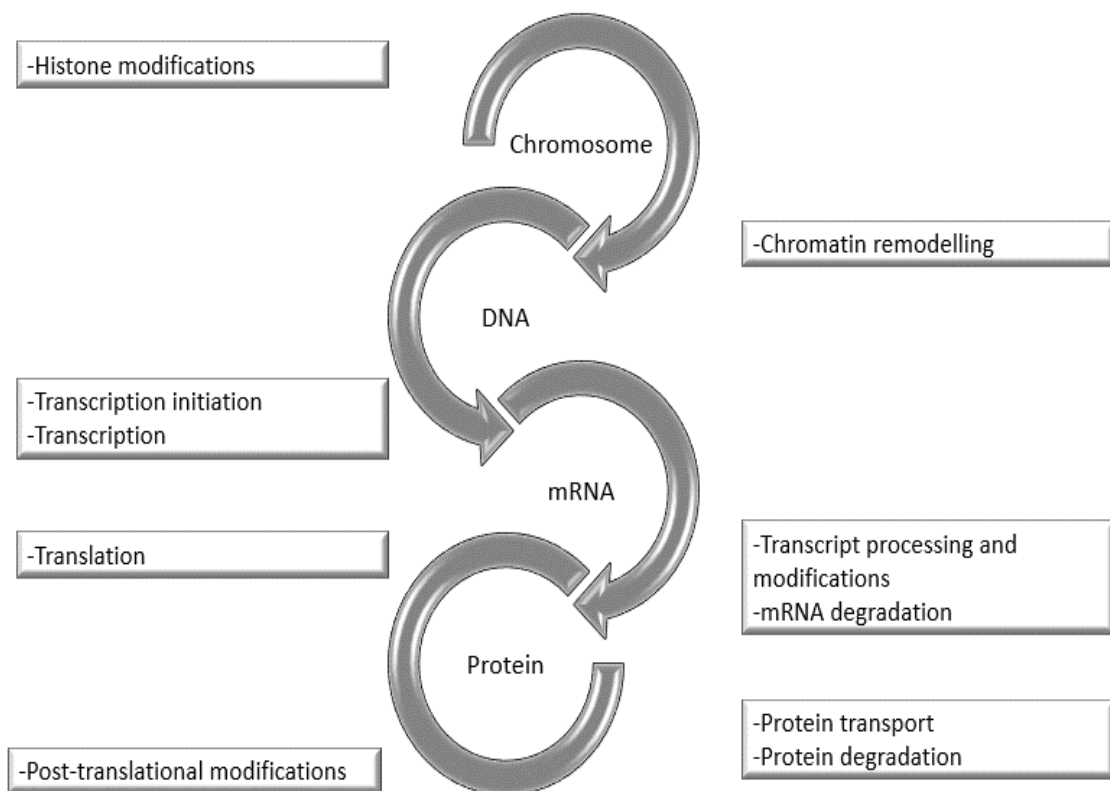


Figure 2.1: The overview of complex gene regulation with various regulatory stages (in boxes) at different cellular levels (in circle).

2.1.1 Gene regulation in eukaryotes

Several gene regulatory mechanisms have been identified previously and they contribute to the changes in gene expression patterns in eukaryotic organisms, namely regulation at chromatin state level, regulation via enhancers and transcription factors as well as regulation of pre-initiation complex and promoter escape.

Gene regulations, especially transcriptions, in eukaryotes are governed at the level of DNA and chromatin. At the level of DNA, *cis*-regulatory modules such as the promoters and enhancers orchestrate target genes in terms of their expression via transcription factors binding (Levine & Tjian, 2003; Wray *et al.*, 2003). The core promoter can be found in all eukaryotic genes and one commonly found core promoter is none other than the TATA box (TATAAAAAA). The core promoter is strongly conserved across all protein-coding genes

in terms of its binding factors and structure whereas other upstream promoters may have differing structures and binding factors across various types of genes (Wray *et al.*, 2003). On the other hand, enhancers are usually found in non-coding regions with strong sequence and functional conservation. The interactions between promoters and enhancers together with several general transcription factors (TFIIA, -B, -D, -E, -F and -H), multi-subunit cofactors (TRAP, mediator, ARC/DRIP, CRSP, etc.) and chromatin remodelling complexes (RSF, PBAF, SWI/SNF and ACF) are essential for the RNA polymerase II to initiate transcription of the target gene, including those tissue-specific genes (Figure 2.2). There are numerous studies on the effects of the dysregulation of promoters and enhancers on gene expression, evolution as well as genetic diseases, indicating their vital roles in various biological processes and development (Cao & Yip, 2016).

At the chromatin level, gene regulations in terms of transcriptions are mediated by core histones (H2A, H2B, H3 and H4) (Figure 2.3). Unlike naked DNA, nucleosome can impose positional restrictions to the access of transcription factors from binding to their recognition sequences, thus impeding transcription of target genes (Zaidi *et al.*, 2010; Lelli *et al.*, 2012). Besides, posttranslational modifications of transcription factors such as phosphorylation can occur in response to extracellular signals to alter its nuclear import via the nuclear pore complexes. Methylation of histones can also strongly affect the transcriptional state of a particular gene (Lelli *et al.*, 2012).

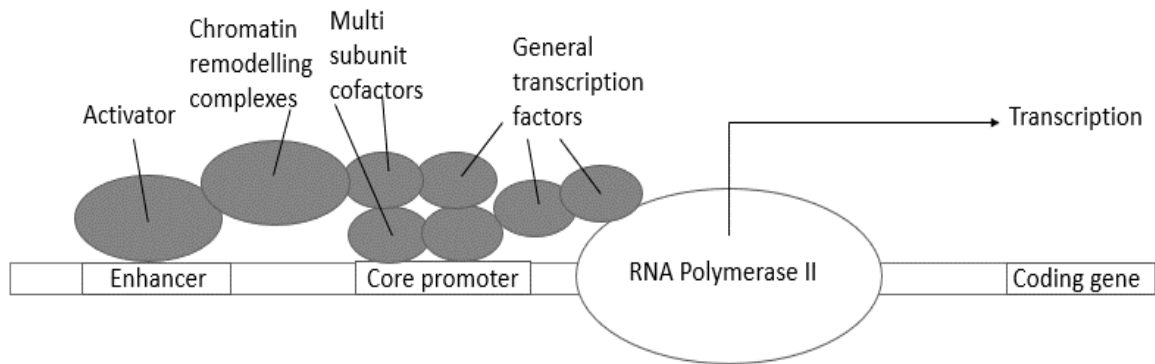


Figure 2.2: The mechanistic actions of enhancer-promoter interactions in transcription initiation involving various general transcription factors, multi-subunit cofactors and chromatin remodelling complexes.

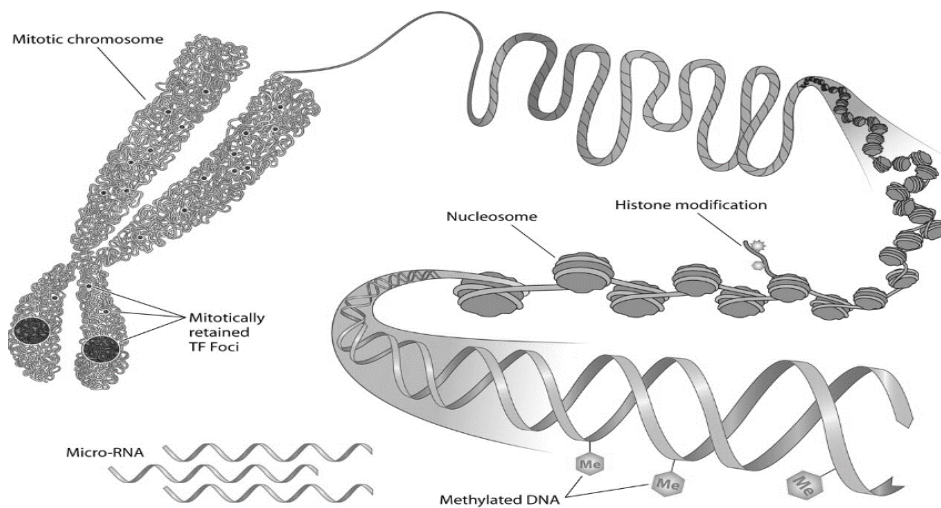


Figure 2.3: Gene regulation at chromatin level (Adapted from Zaidi *et al.*, 2010).

2.2 *Cis*-regulatory elements

Cis-regulatory elements are non-coding DNA sequences that act upon adjacent genes to bring about the regulation of gene transcription (Figure 2.4). These elements are usually located close to the genes they regulate, and they are highly responsible for essential

biological processes such as morphogenesis and anatomy development (Epstein, 2009). The examples of cis-acting elements are operators and some enhancers

Cis-regulatory elements are deemed to play major role in evolutionary processes, sometimes even more profound than that of coding regions. Coding regions including those protein-coding ones are influenced greatly by negative evolutionary process through selection where variations in functional sequences give lethal effects (Epstein, 2009). Often, the variations caused by mutations occurring within these noncoding cis-regulatory elements ultimately results in huge phenotypic consequences of the host such as chronic diseases such as atherosclerosis, β -thalassemia and haemophilia (de Vooght *et al.*, 2009).

The role of mutations of noncoding genetic regions (not limited to enhancers only) in the progression of genetic diseases are minimally explored but is picking up its pace. Previous studies had successfully proven the link between enhancer sequence variations and the onset of genetic diseases in human: *RET* proto-oncogene enhancer variant with susceptibility to Hirschsprung disease, *Sonic hedgehog (SHH)* limb enhancer point mutations with preaxial polydactyly as well as *beta-globin (HBB)* enhancers with thalassemia's (Visel *et al.*, 2009b).

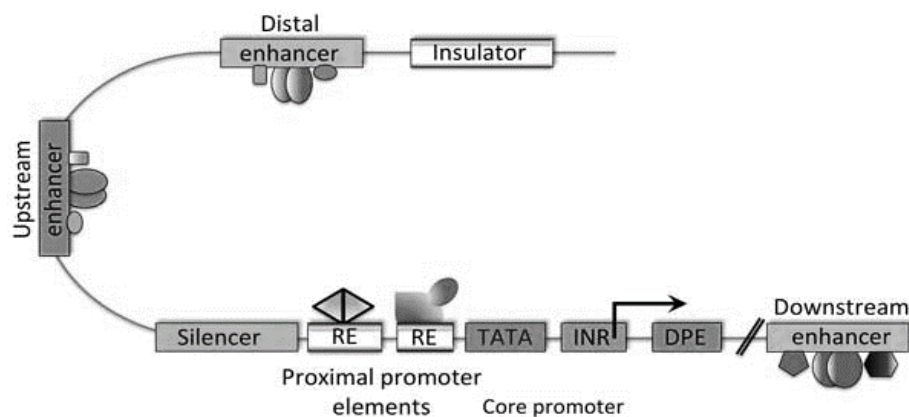


Figure 2.4: An overview of *cis*-regulatory elements that regulates transcription activities (Adapted from Levine & Tjian, 2003).

2.2.1 Enhancer

Back in the year 1981, the term ‘enhancer’ was first introduced by Walter Schaffner and his co-scientists to define DNA repeats of 72 bp in length that significantly activated β -globin gene in rabbit. Enhancer is primarily defined as DNA element that act across distances regardless of its orientation as well as its position from the gene it regulates (upstream or downstream). They also proposed that enhancer actions can cause change in superhelical density of DNA, facilitate RNA Pol II entry, as well as nuclear matrix binding (Schaffner *et al.*, 1981).

Initially, enhancer (particularly SV40 enhancer) is defined as element of tandem repeats (size covering 72 bp) residing around 100 bp away upstream from the cap site of early viral genes in which deletion of this element diminishes early gene expression of T antigen significantly. Not long after, it was found that SV40 enhancer can act on more than just natural adjacent genes but also genes that are heterologous. Enhancer was later found that it can act regardless of its locality (upstream or downstream of gene) and orientation, making this feature one of the major differences between enhancer and promoter (Banerji *et al.*, 1981). The gene enhancement activity of enhancer was also found in the year 1981 where presence of enhancer sequence elevates RNA and protein synthesis of gene products under the regulation of mouse mammary tumour virus promoter, which is difficult in the absence of enhancer (Banerji *et al.*, 1981).

Generally, enhancers are short portions, with lengths ranging from 50 to 1500 base pairs, of DNA that can act as binding platforms for transcriptional activators or suppressors to eventually increase or decrease the transcription rate of particular genes. Besides, enhancers are cis-acting regulatory elements that regulates gene transcription at various developmental stages spatially and temporally. In its active state, enhancers can recruit RNA

polymerase II as well as general transcription factors to synthesize enhancer-originating RNAs (eRNAs) which can in turn improve the efficiency of enhancers (Melo *et al.*, 2013; Natoli & Andrau, 2012).

The binding of proteins to the enhancer region is limited to only those having specific sequences, and thus this accounted for their specificity in their actions. However, not all enhancers are functionally active despite the fact that enhancer-related chromatin feature is abundant in the genome (Bulgar & Groudine, 2010).

It is interesting that we can now predict enhancer activity by the means of detecting the presence of enhancer markers such as H3K4me1, H3K4me3, H3K36me3 and H3K27ac (Creyghton *et al.*, 2010; Zhu *et al.*, 2013). This can be done via the coupling of chromatin immunoprecipitation as well as massively parallel sequencing (ChIP-seq) (Zhu *et al.*, 2013).

2.2.2 Enhancer mechanism

There are four models that describe how enhancers regulate genes as reviewed by Kolovos *et al.* (2012), namely the tracking model, linking model, relocation model as well as looping model (Figure 2.5).

The tracking model as proposed by Kolesky *et al.* (2002), postulates that enhancers tend to move along the track on chromatin fibre following its activation by a transcription factor. This process will then eventually bring upon the elevation of transcription level for the genes they regulate. It is also proven earlier by Herendeen *et al.* (1992) that this tracking model requires transcription-activating signalling between enhancers and promoter (in this case, the T4 late promoter).

The linking model is similar to that of the tracking model, only that the polymerization of protein takes place as the complex moves towards the promoter to drive transcription.

This model was proposed by Bulger and Groudine (1999) where it explained how stably integrated beta-globin transgenes are dependent on locus control regions in terms of site of integration (position effects) on gene transcription.

The relocation model involves the affinity of a gene of target to relocate towards the spot where enhancer-promoter interactions and subsequently favours the occurrence of transcription. In this model, repositioning of locus was found to facilitate the gene activation or silencing process by nuclear compartmentalization (Lancôt *et al.*, 2007).

The looping model explains how protein-protein interactions between enhancers and promoter exert effects on the DNA conformation by the formation of loop. The physical proximity of enhancer on promoter provide a strong way to activate genes at a distance. It is also found that looping can cause a great increment of enhancer activity of over 3 kb in yeast (Petrasccheck *et al.*, 2005).

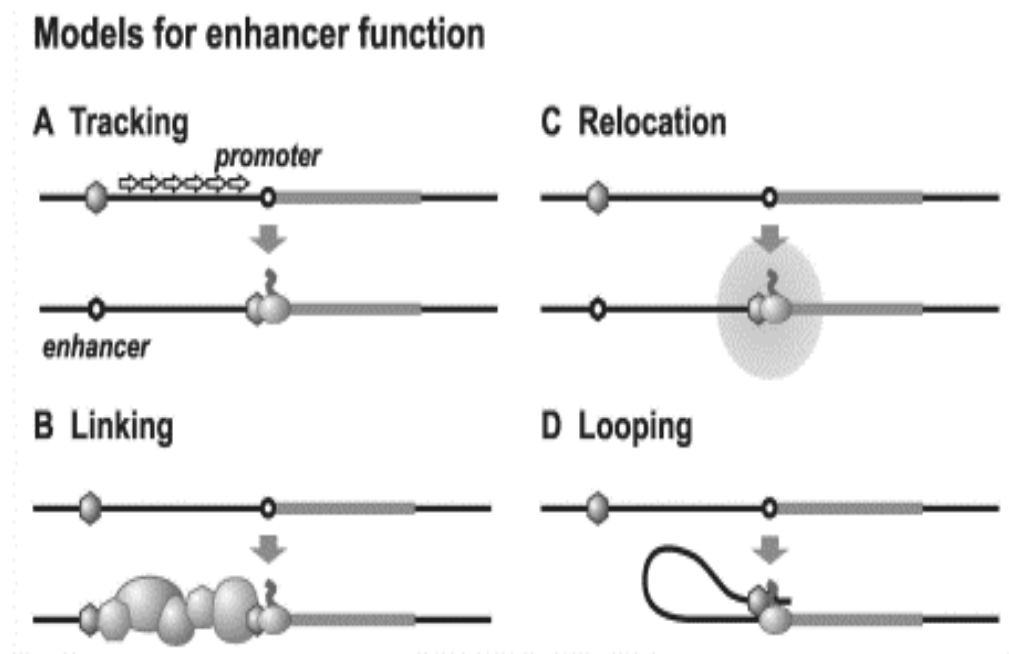


Figure 2.5: The four enhancer mechanism models where the hexagon represents transcription factor, grey oval represents polymerase, and grey halo represents nuclear subcomponents (Adapted from Kolovos *et al.*, 2012).

2.3 Identification of enhancer

The identification of enhancers can be conducted via two main approaches, namely the experimental (reporter assays and high-throughput assays) and *in silico* approach. There are a few reporter assays that have been proven successful in enhancer identification, namely enhancer trap, transient transgenesis assay and undirected integration of enhancer-reporter vector (Cao & Yip, 2016). The *in silico* approach has a much wider scope, ranging from training models to machine learning and it is free from any context restrictions (Cao & Yip, 2016).

One of the major drawbacks of reporter assays and high-throughput assay as compared to *in silico* approach is that the preservation of the whole native state of the predicted enhancers is impossible to achieve via these *modus operandi* (Cao & Yip, 2016). To put it in other words, the enhancer context tested varying from various disease state, developmental stage and cell or tissue types of which enhancers are tested, may differ from the others in terms of chromatin state, across different species or even differing genome structures, resulting in completely dissimilar enhancer-promoter looping (Cao & Yip, 2016). The enhancer activities data obtained may not accurately reflect its activities in nature context. Thus, this state of art grants *in silico* method a great advantage over the two abovementioned approaches.

Currently, there is an ongoing project by the National Human Genome Research Institute named the Genomics of Gene Regulation (GGR) which aims to bridge the gap between genomic data and gene regulatory elements where gene regulations can be predicted directly based on the DNA sequences alone. It also targets to decipher phenotypic consequences of non-coding genomic regions with varied genetic makeups as well as to comprehend the regulatory elements involved in a gene of a cell type at the particular

location and developmental time (National Institute of Health, 2017). This project is expected to be very powerful in aiding accurate predictions of gene regulatory networks and thus forging the advancement of genomic medicine.

2.3.1 Reporter assay approach

There are three ways enhancer reporter assays can be characterized via experimental approaches using a reporter, that is enhancer trap, undirected integration of enhancer-reporter vector as well as transient transgenesis (Kvon, 2015).

The first method is the enhancer trap method where vector designed to have a reporter gene and a minimal promoter is integrated randomly in the genome. The results generated from this approach is the resultant cumulative enhancer activity from the mediation of all regulatory elements of the target gene (Kvon, 2015). The activity of the enhancer trap transgene is very dependent on its site of integration, resulting in varying cumulative adjacent enhancer activities and these activities are not usually additive. The advantage of using this method is that it provides information about the regulatory architecture of a given locus in the genome (Schwarzer & Spitz, 2014). One of the major limitations of this method is that it does not reveal the exact location and activity of individual enhancers and therefore it requires validations from other direct methods of enhancer characterization.

The undirected integration of the enhancer-reporter vector involves inserting enhancer sequences upstream of reporter gene in a vector encompassing transposon recognition sequences. Embryo integration is the approach used to randomly incorporate the transgene in the genome (Visel *et al*, 2009a). This method allows for detection of individual enhancer activity as well as the concerted effects of multiple enhancers, thus provide information on the functional anatomy of enhancers and their modularities (Visel *et al*, 2009a). This method

has been proven to be a very useful tool in characterizing various human developmental enhancers in mouse and other model organisms (Pennacchio *et al.*, 2006).

Transient transgenesis is one of the most sensitive method in enhancer characterization. In this approach, a vector encompassing enhancer-reporter is injected into specific organs or cell lines for transient view of enhancer activities. The difference between this method and the previous is that this method does not involve the integration of enhancers into the genome, but it is allowed to act freely in the extrachromosomal space (Kvon, 2015). This approach had been proven feasible in *C. elegans* where injection of plasmid construct into the gonads provides overview of enhancer activities across several generations (Mello *et al.*, 1991).

2.3.2 High-throughput assay approach

High-throughput assay approach is an approach to scale up reporter assays that can be utilized to test multiple enhancers at the same time and context. An example of this approach is STARR-seq where it is used to test enhancer activities downstream of reporter gene in a manner that the enhancer sequence has been made part of the RNA transcript outcomes (Arnold *et al.*, 2013). The number of reads encompassing enhancer sequence can give clue on the quantification of enhancer activities via standard RNA-sequencing approach (Arnold *et al.*, 2013).

Another example of high-throughput assay for enhancer activity determination is the massively parallel reporter assays. This approach was first used on bacteriophage promoter assays by Patwardhan *et al.* (2009) where functional assays were conducted on all possible point mutations and small indels of these promoters (T3, T7 and SP6) synthesized in parallel

on a programmable microarray. The differential activity of distinct promoter variants was determined from pool of promoter barcode library via RNA-seq approach.

In another study by Kwasnieski *et al.* (2014) which focused on targeted motif disruptions of 2104 ENCODE-predicted enhancers and 3314 enhancer variants in HepG2 and K562 cells, the reduced enhancer activities were affected by the disruptions of certain transcriptional activators (HNF1, HNF4 and FOXA in HepG2; GATA and NFE212 in K562). HNF1 motif disruption resulted in decreased expression in HepG2 but null effect on K562 cells whereas GF11 motif disruption only caused reduced expression in K562 but not in HepG2. This study has proven that enhancer activity is strongly correlated by nucleosome exclusion, enhancer strength, evolutionary conservation as well as binding of other factors.

The major limitation of these approaches is the failure to preserve the native context of predicted enhancers wholly. As mentioned above, considering that the context of these experimental approaches deviates from the whole native context, active enhancers identified experimentally may have differing activity status (inactive or selectively active) in their native form.

2.3.3 *In silico* approach

The *in silico* method in enhancer identification is a great leap stone compared to other approaches in terms of its flexibility in utilization of various data obtained from different context and platforms to produce predictions comparable to their experimental counterparts in the most cost-minimal manner.

Despite its major benefits as mentioned in the previous section, it is obstructed by the ultimate predicaments that are unavoidable in enhancer identification, that is the enhancer locations, context, features as well as its mode of actions (Cao & Yip, 2016). First, the

relative location of an enhancer from the gene it regulates is not bounded by any rules yet to date. Enhancers can be found in anywhere within the genome, ranging from intron, intergenic region to even exons of genes located far apart. The flexibility of enhancer locations adds up to the difficulty in enhancer localization during enhancer predictions (Cao & Yip, 2016). Secondly, there is no single "enhancer mark" that can be used to accurately identify enhancers from non-enhancers in the genome (Shlyueva *et al.*, 2014). Thirdly, context specific can be seen in the enhancer activities as well as interactions between enhancer and promoter of a particular gene (Yip *et al.*, 2013). Fourthly, enhancer-promoter interactions can occur in a many-to-many association in which multiple enhancers can regulate one promoter or one enhancer can interact with various promoters (He *et al.*, 2014). Last but not least, the biggest challenge of all is that the list generated so far for experimentally verified enhancers is not comprehensive enough and the "gold standard" enhancers are heterogeneously defined across several publications resulting in differing training models (Cao & Yip, 2016).

There are several features that are used in *in silico* enhancer predictions using training models, namely transcription factor binding motifs, histone modifications, evolutionary conservation, chromatin accessibility, DNA methylation, eRNA and transcription factor binding sites based on ChIP-seq or ChIP-exo (Cao & Yip, 2016). The advantages and limitations of each features are summarized in Table 2.1. There are various *in silico* tools that had been developed recently to eventually predict enhancers accurately and at the same time searching for the universal enhancer mark that can be very useful in a genome-wide search (Cao & Yip, 2016; Lim *et al.*, 2018). In this study, five different enhancer prediction tools employing different enhancer features were selected to predict enhancers from the genome of the proboscis monkey.

Table 2.1: A summary of features considered in *in silico* enhancer predictions (Cao & Yip, 2016).

Features	Advantages	Limitations
TF binding motifs	Vastly available	Binding of TF in a given context is not guaranteed by presence of motif; Canonical sequence motifs are not necessary for TF binding
Histone modifications (HMs)	Can identify poised and active enhancers (positive or negative HMs); Each context only requires little amount of ChIP-seq experiments	No perfect HM correlations to enhancer activities can be traced based on single HMS or any combinations of HMs
Evolutionary conservation	Vastly available	Sequence conservations do not usually correlate highly with functional conservations in enhancers; Lack of context-specific data; Cannot differentiate enhancers from other conserved DNA elements
Chromatin accessibility	Only one feature is needed for each context	Regions of high chromatin accessibility do not usually correspond to regions of high enhancer activities
DNA methylation	Supplies complementary data to other features	Poor understanding of the quantitative relationship between target gene expression and methylation of enhancers; The regulation of enhancers differs accordingly with types of DNA methylation
eRNA	Vastly available transcriptome data; High accuracy in enhancer prediction	The mode of actions of eRNAs are not fully understood; Active enhancer may not synthesise eRNAs; Regions synthesising eRNA-like RNAs are not necessarily enhancers; eRNA signals may not be present in poly-A enriched RNA-seq data
TFBSs based on ChIP-seq or ChIP-exo	Direct measurements taken from the desired context	A huge number of ChIP-seq experiments are mandatory for a comprehensive list of binding sites for many TFs

2.3.3.a Chromatin signature identification by artificial neural network (CSI-ANN)

The CSI-ANN is a tool that can deploy chromatin signatures in the process of searching for functional DNA elements (Firpi *et al.*, 2010). It takes into account both data transformation and feature extraction prior to classification, thus adding significant level of accuracy in genome-wide scale prediction of regulatory DNA elements based on chromatin signatures (Firpi *et al.*, 2010). This tool considers more than just arithmetic mean but also skewness and kurtosis of the raw histone modification signal data. For instances, it could capture histone modification peaks of different shapes not being captured by previously designed approaches (Firpi *et al.*, 2010).

CSI-ANN uses HeLa cell ENCODE data (6 histone modification ChIP-Chip data namely: H3K4me2, H3K4me3, H3, H3K4me, H4Ac and H3) as well as human CD4⁺ cell data (T-cell histone modification ChIP-Seq data) as data source. This tool utilizes location data of 3 markers from ENCODE region namely: DnaseI Hypersensitivity Sites (DHS), TRAP220 and p300 which are known enhancer-binding co-activators (Firpi *et al.*, 2010).

About 75.2% of enhancers predicted by CSI-ANN were backed up by at least one of the 4 lines of evidence: DHS, sequence conservation, PreMod and p300. This tool has achieved 66.3% positive predictive value and 65.5% sensitivity with overall improvement of 5.9% and 11.6% over best approach previously (Hidden-Markov Model). One of the disadvantages of this tool is that it only considers chromatin histone modification data but not other data sources such as transcription factor binding sites, nucleosome positioning and gene expression profiles (Firpi *et al.*, 2010).

2.3.3.b Gapped k-mer support vector machine (gkm-SVM)

The gkm-SVM is a robust tool that are designed for estimation of k-mer (6-8 bp) frequencies for the activity prediction of 500 bp to 2 kb functional genomic sequence elements. This tool was proven to outperform the previously designed kmer-SVM where long TFBSs are taken into consideration instead of just short ones (Ghandi *et al.*, .2014). The gkm-SVM tool functions to predict enhancer activity based on transcription factor binding and also to determine how enhancers can differ in function due to specific common or rare genetic variation.

Ghandi *et al.* (2014) have proven that this tool outperformed its counterparts (the best single PWM) for its ability to predict all ENCODE ChIP-seq data from ChIP-seq regions. This tool has the advantage over position weight matrix (PWM) approach as scoring threshold and large data sets are required to be optimized using PWMs. This tool has shown increased accuracy significantly compared to their close counterparts in terms of prediction of tissue-specific enhancers as well as functional genomic regulatory elements (Ghandi *et al.*, .2014).

2.3.3.c DeepBind

DeepBind is a pattern discovery tool that can evaluate pools of data sets and metrics of wide variety, it can evaluate both ChIP-seq and HT-SELEX data sets even on a very large scale (Alipanahi *et al.*, 2015). This tool can be easily applied to both sequencing and microarray data and its graphics processing unit (GPU) can be used for parallel implementation purposes without inducing technology-specific biases. It is found that this tool is useful in the analysis of disease-associated genetic variants as well as their corresponding TFBS and gene expression.

DeepBind works by capturing binding specificities from raw sequence data and concurrently finding new sequence motifs and their binding specificities, subsequently it consolidates all these data into predictive binding score. Then, eventually DeepBind can be useful in predicting binding sites in test sequence and also in the discovery of the effects of novel mutations by scoring (Alipanahi *et al.*, 2015).

2.3.3.d iEnhancer-2L

The iEnhancer-2L is a tool programmed to formulate DNA elements with the "pseudo k-tuple nucleotide composition" (Liu *et al.*, 2015). It is deemed by its developer as one-of-a-kind tool that can determine the strength of enhancers instead of just enhancer identification. This tool functions on benchmark dataset which were constructed from chromatin state data (including various histone marks: H3K27ac, H3K4me3, H3K4me1, etc.) of nine cell lines namely HMEC, NHEK, HSMM, HpeG2, H1ES, GM12878, K562, HUVEC and NHLF (Liu *et al.*, 2015).

This tool operates via two predictors but both predictors work with SVM algorithm. The iEnhancer-2L tool processes the sample sequences into fragments of 200 bp (so that nucleosome and linker DNA are included), resulting in identification of 742 strong enhancers. CD-HIT software was used to reduce redundancy and biases from the result data sets (Liu *et al.*, 2015).

One of the major advantages of this tool is that it operates on two layers: the first layer involves identification of enhancers from the query DNA sequences while the second layer determines the strength of the enhancers identified (Liu *et al.*, 2015).

2.3.3.e Gapped motif feature representation-convolutional neural networks (GMFR-CNN)

The GMFR-CNN tool is a tool that relies solely on characteristics of sequence to differentiate enhancers from pools of non-enhancers and it was constructed using the extended position weight matrix. This tool was proposed based on the concept of k-mer sequence features (gapped motif feature) and it was designed to capture motifs undiscovered in the previously designed continuous sequence motifs algorithm (Wong *et al.*, 2016).

The GMFR can work location insensitively with various data despite their different locations and this possess advantage to identify enhancers from anywhere of the genome. The CNN was deployed to help in the detection of essential features and organise them into meaningful data. This tool was cross-validated with LS-GKM and DeepBind in terms of f-measure, accuracy, precision and recall (Wong *et al.*, 2016).

The combined approach of using GMFR coupling with CNN has achieved 98% accuracy for 10 ChIP-seq data sets describing transcription factors binding within enhancer regions from the literature of DeepBind (CEBPB, CEBPD, CTCF, ELK1, EP300, FOXA2, HNF4A, TBP, TFAP2A and YY1) (Wong *et al.*, 2016). This tool has been proven to outperform LS-GKM and DeepBind in the context of enhancer identification with its results producing near optimal f-measure and accuracy.

2.4 Liver and its main transcription factors

The liver is an essential organ found only in vertebrates and it plays major roles in biological processes such as detoxification of various metabolites, biochemical factory for the digestive system, hormone and protein synthesis as well as decomposition of red blood cells. The role of liver in mediating drug responses has been extensively studied as adverse

drug events has been one of the major causes of death in this century (Kim *et al.*, 2011). Studies have found that not only the nucleotide variations in protein-coding genes can lead to the heterogeneity in drug responses but the nucleotide variations in non-coding elements (including SNPs) as well as possibly the enhancers also had made substantial contributions to variability in drug responses (Kim *et al.*, 2011).

A wide range of prescription drugs were designed to target membrane transporters because these transporters are the major mediators in absorption, dissemination, removal as well as translocation of various xenobiotics (Kim *et al.*, 2011). Liver membrane transporters play roles in uptake and eliminating most drugs and xenobiotics in the liver. Previous studies have shown that nucleotide variants of promoters can lead to alteration in the transporter transcription and thus the affected gene expression resulting in modifications on the pharmacokinetic and pharmacodynamic properties of a particular drug (Sun *et al.*, 2010; Kim *et al.*, 2011). Besides, the enhancer nucleotide variations are also major contributors towards the differing drug responses across individuals (Kim *et al.*, 2011; Corradin & Scacheri, 2014; Miguel-Escalada *et al.*, 2015; Zhang *et al.*, 2018).

According to Kim *et al.* (2011), there are a total of nine liver membrane transporters that are known to have direct interactions with frequently utilized pharmaceuticals, namely *SLC22A1*, *SLCO2B1*, *SLCO1B1*, *SLC10A1*, *ABCC2*, *ABCB11* as well as *SLC22A7*. The transcription factors selected for transcription factor binding site analysis are based on those selected by Kim *et al.* (2011) which are major transcription factors that affects these liver membrane transporter gene transcriptions significantly in the liver. The six liver-specific transcription factors selected are HNF-1, AP-1, GATA-3, C/EBP- β , NF-1 as well as HNF-3 β .

2.4.1 AP-1

AP-1 was among the first transcription factors to be discovered in mammals, however its physiological functions are yet to be elucidated entirely. One of the challenges faced when assigning clearly-defined biological functions to this transcription factor is its diverse biological responses due to its heterogeneous functional nature.

The important functions of AP-1 can be seen through its diverse roles in neoplastic transformation, growth promotion, apoptosis, cell proliferation and death, bone biology, embryonic development as well as multistep tumorigenesis (Jochum *et al.*, 2001; Shaulian & Karin, 2001; Shaulian & Karin, 2002). AP-1 was recently found to function as transcriptional repressor of NOS-3 gene which is strongly related to cholestatic liver diseases (González-Rubio *et al.*, 2013).

Besides, it was also found that AP-1 can work with another growth-induced transcription factor (STAT3) as well as tissue specific transcription factor (HNF-1) in an interaction that leads to an adaptive stimulus to injury of liver organ. They are the essential transcription factors that play major role in the amplification of hepatic genes expressions in which these increased gene expressions contribute to liver organ regeneration via homeostatic stimuli (Leu *et al.*, 2001).

2.4.2 C/EBP- β

The CCAAT-enhancer binding proteins (C/EBPs) is a family of six transcription factor isoforms which are generally involved in energy metabolism, inflammation as well as cell differentiation (Strohmeyer *et al.*, 2014). The C/EBP- β plays vital roles in biological pathways leading to metabolic syndrome (van der Krieken *et al.*, 2015). As reviewed by van der Krieken *et al.* (2015), this transcription factor functions in adipose tissue differentiation,

insulin and glucose metabolism, inflammation, high density lipoprotein synthesis, triacylglycerol breakdown, hepatic steatosis and endoplasmic reticulum stress.

The expression of C/EBP- β transcription factor was found to be high in organs such as spleen, liver, kidney, pancreatic β -cells, adipose tissue and intestine. The C/EBP- β is one of the major regulators of cAMP levels in the liver as well as adipose tissue besides its role in downregulation of albumin expression (Singh *et al.*, 2013). The C/EBP- β was found to exert partial protective pressure from steatosis progression, especially in the studies involving mice provided with methionine-choline deficient diet (MCDD). The absence of this transcription factor was predicted to have direct effects on the decrease in lipid accumulation in the liver (van der Krieken *et al.*, 2015). The positive correlation between the overexpression of C/EBP- β and the onset of hepatic steatosis was also demonstrated in their studies (van der Krieken *et al.*, 2015).

2.4.3 GATA-3

One of the functions of GATA-3 in the thymus is its role in the development of CD4 single-positive (SP) cells. This transcription factor not only has the ability to halt T_h1 cell differentiation, but it can also induce T_h2 cell differentiation (Yagi *et al.*, 2011). The expression of GATA-3 directly affects the IFN- γ synthesis which is vital for the proliferation of T_h2. Besides, the up-regulation of GATA-3 is essential for granting naive CD4 T cells (after they are triggered active via their T cell receptors and IL-4) the ability to synthesize T_h2 cytokines (Yagi *et al.*, 2011).

GATA-3 was found to bind several biological essential regulatory elements namely HSII, NFAT, LCR, STAT5, HSVa and CGRE (Yagi *et al.*, 2011). This supports the role of GATA-3 in mediation of gene expression as well as chromatin remodelling, working in

tandem with several other transcription factors. GATA-3 also plays major role in NK cell homeostasis in the liver. In the absence of GATA-3, the synthesis of IFN- γ will be downregulated or insufficient to counter the infection by the intracellular bacterium *Listeria monocytogenes*. This will eventually result in increased susceptibility to listeriosis and elevated bacterial burden accumulation in organs such as the liver, brain and spleen (Samson *et al.*, 2003). Furthermore, Chou *et al.* (2010) had confirmed the role of GATA-3 in mammary development and breast tumor differentiation in which this role is pivotal as a strong predictor of clinical outcome in human luminal breast cancer.

2.4.4 HNF-1

The HNF-1 is one of the essential regulators of liver genes transcription made up of HNF-1 α and HNF-1 β hetero- and homodimers. This transcription factor plays major role in transactivation of liver-specific albumin transcription (Maire *et al.*, 1989).

The HNF-1 α was found to regulate apolipoprotein gene positively in which elevation of this gene in liver can lead to high risks of stroke as well as coronary artery disease (Wade *et al.*, 1994). Besides, this transcription factor could exert a repression on liver specific genes such as the sucrase-isomaltase gene. The genes that HNF-1 α positively regulates in the liver includes albumin, aldolase B, CRP, Vitellogenin A2, CYP2E1, IGFBP-1, HBV, Apolipoprotein B, Apolipoprotein AII and many more. Interestingly, most of the liver-secreted plasma proteins encompass HNF-1 binding site within their promoter (Maire *et al.*, 1989).

Moreover, one of the unique features that distinguish HNF-1 α from other vertebrate homeodomain-containing proteins is that it can undergo dimerization without the presence of its DNA recognition sequence (Mendel *et al.*, 1991). The dimerization cofactor of HNF-

1 α (DCoH) was found to regulate the assembly of tetrameric complexes that are active during transcription and eventually play major roles in tissue specificity and development of the complex. DCoH is no doubt the major regulator of transcriptional network involving the expression of liver-specific genes as well as liver development (Mendel *et al.*, 1991).

2.4.5 HNF-3 Beta

The HNF-3 β (FOXA2) is one of the three protein members of the HNF-3 gene subfamily apart from HNF-3 α and HNF-3 γ . Together, these three transcription factors are collectively responsible for the regulation of hepatocyte-enriched transcription of genes essential for organ functioning (Samadani & Costa, 1996). The HNF-3 β mutually shares the same function with HNF-3 α in the mediation of gene expression in pancreatic, intestinal and bronchiolar epithelium as well as endoderm-derived hepatocytes (Rausa *et al.*, 1997; Clevidence *et al.*, 1998).

Besides, the HNF-3 β are grouped with HNF-3 α under a huge transcription factor family that are vital in embryonic pattern developmental processes. It was also found that HNF-3 β , HNF-3 α and HNF-3 γ work in tandem with the respective *cis*-acting HNF-3 binding elements in promoters of key liver-specific genes of such as tyrosine amino transferase, phosphoenolpyruvate carboxykinase, albumin and aldolase B gene to interact and regulate gene regulations (Schrem *et al.*, 2002).

2.4.6 NF-1

The site-specific NF-1 transcription factors family is involved actively in regulation of transcription of a huge range of genes in various cell types. The one unique feature that distinguish NF-1 from other transcription factors is that its N-terminal domain is greatly

conserved and depicts no homology to any known classes of DNA binding motifs to date (Blomquist *et al.*, 1999). This transcription factor family are encoded by four different genes in mammals and they are ubiquitous.

The NF-1 transcription factor is one of its kind in terms of its capability in performing functions as both transcription factor as well as viral replication factor. It was also found that it has low affinity for nucleosomal DNA. The role of NF-1 in hormone-mediated transactivation of mouse mammary tumour virus (MMTV) was reviewed by Chikhirzhina *et al.* (2008) in terms of its role in remodelling of nucleosome as well as activation of transcription. The construction and maintenance of chromatin conformation in the MMTV promoter region was found to be regulated by NF-1 during macrophage differentiation. This transcription factor also mediates regulatory elements of lysozyme gene while differentiation of macrophage takes place (Chikhirzhina *et al.*, 2008).

2.5 Significance of non-human primates in biomedical researches

The biomedical research field has recently shed spotlight on the use of non-human primates (NHPs) as model organism of interest in studies, especially in the regulatory sequence comparison between primates and human. The reasons these primates made a good option for studies include the major similarities they share with human in terms of dietary factors and responses to environmental exposures (Huang *et al.*, 2015).

The physiological similarities that humans share with primates in terms of especially development, cognition, neuroanatomy, social complexity, cognitive aging, reproduction as well as pregnancy makes them more feasible model for disease and drug metabolism studies, lung malfunction studies and AIDS research compared to other model organisms (Phillips *et al.*, 2014). The New World monkeys and the Old World monkeys are the ones deemed

more popularly used as primate model compared to prosimians and Great Apes (Phillips *et al.*, 2014). Recently, the completion of genome sequencing of some non-human primates such as the orangutan, rhesus monkey, gorilla, marmoset and bonobo has improved our understanding on human biology that differs humans from their primate counterparts, including the genes that affect gene regulations (Gordon *et al.*, 2016; Prüfer *et al.*, 2012; The Marmoset Genome Sequencing and Analysis Consortium, 2014). There are some ongoing genome sequencing projects involving non-human primates like baboon, owl monkey, black and white colobus, gibbon, drill, mouse lemur, sifaka lemur and sooty mangabey (Baylor College of Medicine-Human Genome Sequence Center, 2016).

2.6 Enhancers in primates

Currently, our understanding on *cis*-regulatory divergence between human and other closely related primates remains limited and speculative albeit the human and chimpanzee genome sequences had been made available years ago. Various *in silico* approaches have been developed to identify non-coding elements which were subjected to accelerated modifications or deletions in the human lineage (Prescott *et al.*, 2015). Furthermore, recent genome-wide comparisons of transcription factor binding and regulatory landscapes between close relative species using full genome sequencing and epigenomic approaches had been proven to be successful in improving our knowledge on *cis*-regulatory evolution and comparative epigenomics in high genome ontology context (Prescott *et al.*, 2015).

Functional epigenomic comparison studies among primates (including humans) had been limited to only lymphoblastoid cell lines (Cain *et al.*, 2011; Shibata *et al.*, 2012; Zhou *et al.*, 2014) and across distant relative species (mammals) in whole organ context (Cotney *et al.*, 2013; Villar *et al.*, 2015) until Cotney *et al.* (2013) analysed comparatively the histone

modification H3K27ac at various developmental stages in embryonic limbs of human, rhesus macaque and mouse. Recently, Boyd *et al.* (2015) started studying the FZD8 enhancer in developing neocortex of human and chimpanzee. Not long after, Prescott *et al.* (2015) made use of iPSC of chimpanzee to compare enhancers in the neural crest cells of human and chimpanzee as well. These studies have provided plenty of candidate enhancers for further evolutionary studies as well as elucidation of the effects of the evolving mechanisms on the morphological variation among higher primates.

The liver organ had been one of the main organs for enhancer studies in human and other mammals due to its important roles in drug responses. Kim *et al.* (2011) had demonstrated in huge details how enhancer variations can influence drug responses across different human ethnics. Besides, the liver organ was also chosen as the subject organ by Villar *et al.* (2015) to study on enhancer evolution across 20 mammalian species. However, so far, no emphasis had been placed on the study of enhancers in the liver among primates other than humans, therefore proboscis monkey had been brought up into the limelight in this research.

2.7 Proboscis monkey and its significance

The proboscis monkey (*Nasalis larvatus*), also known as the long-nosed monkey, is one of the largest monkey species native to Asia. It is grouped under the family Cercopithecidae and it is one of the Old World monkeys exclusively endemic to the Borneo island. It is deemed the most primitive colobine as it possesses special diploid number of $2n=48$ (Chiarelli, 1966; Soma *et al.*, 1974; Stanyon *et al.*, 1992).

This monkey is highly distinguishable from others in terms of their long and fleshy nose which can grow up to a length of four inches in male (Ellis, 1986). Besides, these

monkeys are coated with fur of reddish brown coloration and they have grey limbs and large bellies (Payne *et al.*, 1986). Proboscis monkey can grow up to a head-body length of 76.2 centimetres and body weight of 30 kg for males (Napier & Napier, 1985). Sexual dimorphism can be seen clearly in the females where they exhibit approximately half of head-body length and weight compared to their male counterparts and their noses are much smaller than the males. They feed on seeds, leaves and also unripe fruits (Boonratana, 2002).

The proboscis monkey is one of the best swimmers among other primates and it can be found inhabiting regions such as coastal areas, swamp forests and riverbanks (Sebastian, 2000). They live on the tree all the time and may forage for food on land at some instances. Their harem groups are organized and composed of a dominant male, 2 to 7 females and their offspring (Boonratana, 2002).

One of the reasons the proboscis monkey was selected for this research is that it possesses unique morphological and behavioural features such as skull structure, terrestrial movement and sexual dimorphism as an indication of this species belonging to one of the long-isolated genera within the colobines (Giusto & Margulis, 1981; Groves, 1989; Peng *et al.*, 1993). Besides, the digestive system of the proboscis monkey also differs greatly from other mammals in terms of lysozymes where their lysozymes are more closely related to that of ruminants than mammals due to adaptive convergence (Stewart *et al.*, 1987). In addition, the unique karyotype of the proboscis monkey ($2n=48$) has made this species one of the most feasible candidate among the Old World monkeys for research because its karyotype is considered conserved to that of human (Chiarelli, 1966; Soma *et al.*, 1974; Stanyon *et al.*, 1992). Last but not least, the genome sequence of the proboscis monkey has been sequenced completely (Abdullah *et al.*, 2014), thus allowing genome-wide discovery of enhancers from this species.

All in all, the enhancer landscape discovery in the proboscis monkey genome may provide new insights on how similar or different this primate model from human in terms of genome-wide gene regulatory enhancers. This may also open up windows to the study of disease-related enhancers in proboscis monkey that have human homology. Moreover, the evolutionary status of this primate model can be described preliminarily based on the enhancers identified and thus enabling phylogenetic comparison between primate sequenced so far.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Enhancer identification

3.1.1 Enhancer identification using iEnhancer-2L

The chromosome 18 of proboscis monkey was utilized as input data (as it is fragmented in human chromosome 6 which is highly related to most liver-specific functions) to search for candidate enhancers using iEnhancer-2L (Liu *et al.*, 2015) enhancer prediction tool (this tool is one of most versatile tool to search for functional enhancers across different cells and tissues) (Liu *et al.*, 2015) (Refer to Appendices). This tool was chosen because it can predict enhancers together with their strengths via a 2-layer predictor algorithm: the first layer distinguishes non-enhancers from enhancers whereas the second layer distinguishes strong enhancers from weak ones. The chromosome 18 of proboscis monkey (size: 67,975,966 bp) was downloaded from GenBank (Accession number: CM003007.1) before inputting the sequences into the query section of the iEnhancer-2L online software. The online software was run until the outcomes were obtained. The collected data and results were consolidated to the selection stage as described in the next section. The number of strong, weak and non-enhancers was enumerated and recorded for analytical purposes.

3.1.2 Enhancer identification using conservation from other primates

The genome-wide human liver transporter gene enhancer data was obtained from Kim *et al.* (2011) as a reference to detect similar liver-specific enhancers from the genome of

proboscis monkey. The enhancers located within the range of 200 bp upstream and downstream of the targeted liver transporter genes were all included in this part of enhancer identification, emulating that by Kim *et al.* (2011). Then, using the human liver-specific enhancer identified (as liver enhancers are highly associated to drug responses in host organism) (Kim *et al.*, 2011), these candidate enhancers were subjected to genome multiple alignment search against the rhesus macaque and chimpanzee using ECR browser (Ovcharenko *et al.*, 2004). Only candidate enhancers with above 70% sequence homology (Bauer *et al.*, 1988) across non-human primates like chimpanzee and rhesus macaque were selected for further selection process using MATCH tool (Kel *et al.*, 2003).

3.1.3 Enhancer identification using five different enhancer predictor tools

First, the genomic data (chromosome sequences) of proboscis monkey as well as protein sequences, ESTs and transcriptomes of reference species (i.e. human, mouse, chimpanzee, Sumatran orangutan and rhesus macaque) were utilized to aid in the assembly and mapping of the chromosomes of proboscis monkey (Omar *et al.*, 2017). The mapping of the chromosomes of the proboscis monkey was conducted using MAKER v2.31.9 (for genome annotation) and AUGUSTUS (for gene predictions within the eukaryotic genome) (Omar *et al.*, 2017). Then, the intergenic sequences were identified, and redundant sequences were removed before proceeding to the next stage (Omar *et al.*, 2017).

Next, enhancer prediction was conducted on chromosome 18 using five different enhancer prediction tools namely iEnhancer-2L, CSI-ANN, gkm-SVM, DeepBind and GMFR-CNN with specific parameters respectively (based on CTCF, EP300, H3K27ac, etc.)

(Omar *et al.*, 2017). Then, the predicted candidate enhancers were subjected to bedtools “intersect” and “merge” function to consolidate results from five different enhancer prediction tools to form consensus enhancer sequences as there is no single enhancer mark to date that can be used as marker for enhancer identification. Bedtools “merge” function was used to combine overlapping enhancer regions between different tools whereas bedtools “intersect” function was used to screen for only the overlapping regions between enhancer predicted by different tools. The bedtools “intersect” and “merge” tools were done individually and gradually between two tools, three tools, four tools as well as five tools comparisons.

3.1.4 Selection of enhancers

The outcomes (predicted enhancers) from each prediction method were subjected to MATCH tool (Kel *et al.*, 2003) analysis using five different profiles, namely liver-specific, muscle-specific, immune cells-specific, cell cycle-specific and best selection profile to screen for those with high liver-specific scoring and low in other profile scoring. All profiles were created based on TRANSFAC database. The best selection profile encompasses numerous groups of matrices for the same transcription factor, the criteria of all other tissue-specific profiles are utilized to make decision on which matrix to use, besides, the cut-offs are also chosen to reduce false negative matches in this profile (Kel *et al.*, 2003). To cater to the objectives of this research, the focus was placed on the top-scoring liver-specific profiles enhancers from each group (four from iEnhancer-2L method, five from primate conservation method and four from five-software method) with a minimum of 300 bp in length. The four

enhancers identified from iEnhancer-2L were named as pme001 to pme004; the enhancers identified from primate conservation approach were termed pme101 to pme105 while the enhancers identified from five-software approach were named as pme201 to pme204.

3.2 Primer design and synthesis

A total of 13 pairs of primers were designed in order to isolate the 13 selected candidate enhancers from the genome of the proboscis monkey. The online software tools to design these pairs of primers are the Primer BLAST, Primer3, as well as the OligoCalc (Ye *et al.*, 2012; Untergasser *et al.*, 2012; Kibbe, 2007). First, the templates were inserted into Primer BLAST (Ye *et al.*, 2012) as input to enable for the search of suitable primer pairs. Then, each primer pairs obtained as outcomes were subjected to second stage of verification using Primer3 software (Untergasser *et al.*, 2012) to check for the difference in melting temperatures between the forward and reverse primers to not exceed more than 5°C. The length of the primer was also restricted to around 25-30 bp (including the RE sites and extra AA/AT flanking bp). All candidate primer pairs were screened based on their feasibilities in terms of length and GC content, and the unsuitable primer pairs were not considered for the next steps. Then, OligoCalc (Kibbe, 2007) was used to determine the feasibility of the primer pairs in terms of GC content (35-60%), delta-G, melting temperatures as well as absence of secondary structures, hairpin and self-complementary. The optimized primer sequence for each enhancer was sent to the Apical Scientific Sdn. Bhd. for synthesis.

3.3 Stool DNA extraction using QIAamp® DNA Stool Mini Kit (Qiagen, Germany)

Fresh proboscis monkey faecal samples (a total of 4 samples) were collected from Bako National Park, Sarawak, Malaysia and they were stored in sterile falcon tubes and were kept inside a -80°C freezer. Prior to the stool DNA extraction, a 70°C water bath was prepared, and all buffers were shaken evenly. If there is any precipitate found in Buffer ASL and Buffer AL, they were dissolved by incubating at 70°C. The DNA was extracted from the stool of the proboscis monkey according to the manufacturer's protocol using QIAamp® DNA Stool Mini Kit (Qiagen, Germany) with some modifications on the elution step. The elution step was repeated by applying another 500 µL Buffer AE followed by incubation for 15 min and final centrifugation at 10,000 x g for 1 min. The DNA sample was stored at -20°C for further experimental procedures.

3.4 Gradient PCR temperature optimization

The gradient PCR temperature optimization was conducted for each primer pair to isolate 13 candidate enhancers from the genome of proboscis monkey using T100™ Thermal Cycler (BioRad, USA). All PCR reactions were conducted in total volume of 20 µL with the reagents listed on Table 3.1. These reagents were prepared in master mix solution of about 4 to 5 times reactions before being aliquoted into individual reaction tubes. The mixtures in each tube were spun down using a table-top spinner before subjecting them to thermal cycling parameters as shown in Table 3.2. Via the gradient temperature block, a variety of different temperatures (ranging from 55-70°C) were tested on the primer pair in search of the most optimum temperature for the isolation of each enhancer. The housekeeping gene

selected for verification of species and as positive control for the PCR reaction is the *cytochrome b* gene and the PCR reactions were done as shown in Table 3.3. The primer pair used to amplify the housekeeping gene (*cytochrome b* gene) are universal primers L14724 and H15915R from Irwin *et al.* (1991) [forward primer L14724 (5'-GACTTGAAAAACCACCGTTG-3') and reverse primer H15915R (5'-GGAATTCATCTCTCCGGTTTACAAGAC-3')].

Table 3.1: PCR reaction components for enhancer isolation and amplification from the proboscis monkey genome.

Component	Final concentration	Volume (1X)	Volume (4X)
10X <i>EasyTaq</i> Buffer (with Mg ²⁺)	1X	2.0 µL	8.0 µL
2.5 mM dNTPs	0.2 mM	1.6 µL	6.4 µL
10 µM enhancer forward primer	0.2 µM	0.4 µL	1.6 µL
10 µM enhancer reverse primer	0.2 µM	0.4 µL	1.6 µL
<i>EasyTaq</i> DNA polymerase	2.5 units	0.2 µL	0.8 µL
Nuclease-free water	N/A	14.6 µL	58.4 µL
Genomic DNA	As required	0.8 µL	3.2 µL
Final volume		20.0 µL	80.0 µL

Table 3.2: Thermal cycling parameters used for PCR reactions.

Stage	Temperature (°C)		Duration
	Enhancer	<i>Cytochrome b</i>	
Pre-denaturation	94.0	94.0	2 min
Denaturation	94.0	94.0	30 s
Annealing	55-70	50.5	30 s
Extension	72.0	72.0	40 s
Final Extension	72.0	72.0	5 min

} 35 cycles

Table 3.3: PCR reaction components for *cytochrome b* gene amplification from the proboscis monkey genome.

Component	Final concentration	Volume (1X)
10X <i>EasyTaq</i> Buffer (with Mg ²⁺)	1X	2.0 µL
2.5 mM dNTPs	0.2 mM	1.6 µL
10 µM <i>cytochrome b</i> forward primer	0.2 µM	0.4 µL
10 µM <i>cytochrome b</i> reverse primer	0.2 µM	0.4 µL
<i>EasyTaq</i> DNA polymerase	2.5 units	0.2 µL
Nuclease-free water	N/A	14.6 µL
Genomic DNA	As required	0.8 µL
Final volume		20.0 µL

3.5 PCR product purification or gel extraction

If there is only a single band observed from the PCR product gel electrophoresis, the PCR product was subjected to QIAquick® PCR Purification Kit (Qiagen, Germany) to clean up and remove the PCR residues according to manufacturer's protocol with some modifications on the elution step. To elute DNA, 20 µL of Buffer EB was added to the center of the QIAquick membrane and was stood for 10 min at room temperature before centrifuging at 10,000 x g for 1 min. The elution step was repeated by applying another 10 µL Buffer EB, incubating for 10 minutes and then centrifuging 10,000 x g for 1 min.

3.6 Cloning of enhancers into pGEM®-T Easy Vector

3.6.1 Chemically competent cell preparation

A volume of 10 µL of *Escherichia coli* XL1-Blue cells was aliquoted from the frozen stock stored at -80°C. Then, the cells were subcultured in universal bottle containing 10 mL LB Broth and the bottle was incubated overnight at 37°C with mechanical shaking at 100 x

g. On the following day, 1 mL of overnight cell culture was inoculated into 20 mL fresh LB Broth and then it was incubated at 37°C with mechanical agitation at 100 x g for 1 hr until the UV spectrophotometer reading of the cell culture solution at OD 600 (600 nm wavelength) reaches the range between 0.4 to 0.5. Next, the cell culture solution was transferred into 18-20 ice-cold 1.5 mL microcentrifuge tubes with 1 mL in each tube. Each tube was centrifuged at 5,000 x g at 4°C for 10 min. The media from each tube was decanted completely from the tube without disrupting the cell pellets trapped at the bottom of the tube before returning it on ice again.

Next, the cell pellets from each tube were resuspended using 160 µL ice-cold 0.1 M CaCl₂ solution and the tubes were incubated on ice for 10 min. Then, all tubes were centrifuged for 10 min at 5,000 x g at 4°C. The media from each tube was once again decanted completely from the cell pellets and was placed on ice again. The resuspension of cell pellets was done with 80 µL ice-cold 0.1 M CaCl₂ solution followed by 80 µL ice-cold 40% glycerol to cryopreserve competent cells for long term storage at -80°C.

3.6.2 Ligation of enhancer into pGEM®-T Easy Vector

Each of the purified enhancer PCR product was subjected to ligation process into pGEM®-T Easy Vector (Promega, USA) (Refer to Appendices). Prior to ligation, all the reagents listed in Table 3.4 were centrifuged briefly to gather the contents at the tips of the tube. The components listed in Table 3.4 were added into a 1.5 mL microcentrifuge tube. The mixtures added into the tube were mixed thoroughly before incubating the tube at 4°C overnight to maximize the yield of transformants.

Table 3.4: List of ligation components for enhancer PCR products.

Component	Standard Reaction	Positive Control
2X Rapid Ligation Buffer, T4 DNA ligase	5.0 μ L	5.0 μ L
pGEM [®] -T Easy Vector	1.0 μ L	1.0 μ L
Enhancer PCR product	3.0 μ L	-
Control Insert DNA	-	2.0 μ L
T4 DNA Ligase	1.0 μ L	1.0 μ L
Nuclease-free water	-	1.0 μ L
Final volume	10.0 μ L	10.0 μ L

3.6.3 Bacterial transformation into XLI-Blue *E. coli* competent cells

Before the transformation takes place, the LAIX (LB agar + Ampicillin + IPTG + X-Gal) plates were first prepared in the laminar flow hood with HEPA filter according to the steps as mentioned below. The previously prepared LB agar was melted using a microwave oven and it was cooled to around 50°C before adding in with 100 μ g/mL antibiotic Ampicillin. The antibiotic supplemented LB agar was then poured and spread onto petri dishes before the agar solidifies. After the agar on the petri dishes had solidified completely (it takes around 10 to 20 min), 100 μ L of 50 mM IPTG solution was spread onto the surface of the agar and was left to dry with the aid of spreader. Then, the addition of 20 μ L of 20 mg/mL X-Gal onto the agar surface was also followed by the spreading and drying process.

The transformation procedures began with the addition of 5 μ L of ligation mixture into ice-cold 1.5 mL microcentrifuge tube with 80 μ L of freshly thawed XLI-Blue *E. coli* competent cells. The tube was left on ice for 15 min. While incubating the cells on ice, a water bath at 42°C was prepared. After 15 min of incubation, the tube containing both the competent cells and ligation mixture was heat-shocked in the water bath for 50 sec before returning the tube immediately to ice for 2 min. A volume of 900 μ L ice-cold LB broth

medium was added to each tube of transformants and the tubes were then incubated at 37°C in an incubator orbital shaker with constant agitation at 100 x g for 90 min. The spreading of 80 µL cell suspension was conducted onto LB agar plate using spreader. The surface-dried plates were incubated at 37°C overnight in an incubator with air-flow.

3.6.4 Blue/white primary colony screening and secondary screening

Blue/white colony screening observation was done on the overnight cultured bacterial agar plates to distinguish the desired transformants (white colonies) from the unsuccessful ones (blue colonies). The blue and white colonies were enumerated, and the transformation efficiency was determined. The secondary screening was conducted similarly on LAIX plates with grids drawn on the bottom of plates for each colony picked from the primary plates to eliminate false positive colonies from the primary screening. The plates were then incubated overnight at 37°C overnight in an incubator with air-flow.

3.6.5 Colony PCR

The colony PCR serves as a molecular screening for recombinant plasmids containing the enhancer insert. The PCR master mix solution was prepared for 5 to 6 reactions each time round as listed in Table 3.5. The PCR tubes were added with components from Table 3.5 with the first 4 tubes dipped with four different white colonies and the last tube with one blue colony as negative control set. The PCR was run according to the optimized annealing temperatures of each individual enhancers as well as PCR parameters as described in Table 3.2.

Table 3.5: List of colony PCR reaction mixtures for enhancer insert.

Component	Final concentration	Volume (1X)	Volume (5X)
10X <i>EasyTaq</i> Buffer (with Mg ²⁺)	1X	2.0 µL	10.0 µL
2.5 mM dNTPs	0.2 mM	1.6 µL	8.0 µL
10 µM enhancer forward primer	0.2 µM	0.4 µL	2.0 µL
10 µM enhancer reverse primer	0.2 µM	0.4 µL	2.0 µL
<i>EasyTaq</i> DNA polymerase	2.5 units	0.2 µL	1.0 µL
Nuclease-free water	N/A	15.4 µL	77.0 µL
Final volume		20.0 µL	100.0 µL

3.6.6 Plasmid Miniprep

Overnight bacterial culture was subjected to plasmid miniprep using QIA[®] Spin Miniprep Kit (Qiagen, Germany) according to manufacturer's protocol with some modifications on the elution step. The elution step was done by adding 30 µL Buffer EB to the middle of the spin column and incubating the tube at room temperature for 10 min. Then, the tube assembly was centrifuged at 10,000 x g for 1 min. The elution step was repeated by applying another 20 µL elution buffer, incubating for 10 minutes and final centrifugation 10,000 x g for 1 min.

3.7 Subcloning of enhancer insert into pGL4.23 [*luc2/minP*] vector

3.7.1 Restriction digestion

The restriction digestion was conducted using the reagents listed in Table 2.6. The restriction enzymes used are the *Bam*HI and *Sal*I to release the enhancer insert from pGEM[®]-T Easy Vector. First, all the reagents in Table 3.6 were added into 1.5 mL microcentrifuge tube before incubating the tube at 37°C for 16 hr for complete digestion reactions. One

digestion tube was prepared for digestion of each pGEM[®]-T Easy Vector containing each enhancer respectively. Another tube was prepared to digest the pGL4.23 [*luc2/minP*] vector (Promega, USA) (Refer to Appendix A) using the same restriction enzymes. The tubes were then incubated in 65°C for 20 min for enzyme inactivation. The digestion solutions containing the cut pGEM[®]-T Easy Vector and enhancer insert were cooled to room temperature before loading the entire mixture into one well of 1% agarose. Agarose gel electrophoresis was done to size separate the pGEM[®]-T Easy Vector from the enhancer insert. The enhancer insert was cut from the gel and gel purified using the protocol as mentioned above whereas the digestion mixture containing the cut the pGL4.23 [*luc2/minP*] vector was PCR purified using the procedures as stated before.

Table 3.6: List of restriction digestion reagents.

Component	Volume (1X)
10X buffer O	2.0 µL
Plasmid DNA (0.5-1.0 µg/µL)	5.0 µL
<i>Bam</i> HI restriction enzyme	1.0 µL
<i>Sal</i> I restriction enzyme	1.0 µL
Nuclease-free water	11.0 µL
Final volume	20.0 µL

3.7.2 Ligation of enhancer insert into pGL4.23 [*luc2/minP*] vector

The ligation reaction was done according to the procedures described for pGEM[®]-T Easy Vector. The reagents listed in Table 3.7 were added into 1.5 mL microcentrifuge tube and the tube was incubated at 4°C overnight for maximum transformants yield.

Table 3.7: List of ligation mixtures for enhancer insert.

Component	Standard Reaction	Positive Control
2X Rapid Ligation Buffer, T4 DNA ligase	5.0 µL	5.0 µL
pGL4.23 [<i>luc2</i> /minP] vector	1.0 µL	1.0 µL
Enhancer insert	3.0 µL	-
Control Insert DNA	-	2.0 µL
T4 DNA Ligase	1.0 µL	1.0 µL
Nuclease-free water	-	1.0 µL
Final volume	10.0 µL	10.0 µL

3.7.3 Transformation, primary colony screening, secondary screening and colony PCR

The transformation of pGL4.23 [*luc2*/minP] vector into XLI-Blue *E. coli* competent cells, primary colony screening, secondary screening and colony PCR were conducted according to the protocol as stated above with some modifications. For the primary colony screening, no IPTG and X-Gal were added (no blue/white screening was done) as pGL4.23 [*luc2*/minP] vector does not induce colouring onto colonies (without β -galactosidase gene). The number of white colonies was counted, and the transformation efficiency was determined. Secondary screening was also done with LB agar and Ampicillin (without IPTG and X-Gal) to distinguish false positives from positives one. The colony PCR was completed with testing of white colonies only to identify colonies containing the desired recombinant plasmids. Then, plasmid miniprep was done according to protocols as described above to extract the plasmid from the positive white colonies for each enhancer.

3.7.4 Sequencing of plasmids

A volume of 10 µL pGL4.23[*luc2*/minP] plasmid containing enhancer insert respectively was delivered to Apical Scientific Sdn Bhd for direct single pass DNA

sequencing. The electropherogram files were examined and analysed using CHROMAS (<https://technelysium.com.au/wp/chromas/>). The sequence errors found in sequencing data were corrected accordingly.

3.8 Cell culture maintenance

3.8.1 Culture medium preparation

The cell culture media was prepared in the laminar flow hood with HEPA filter. First, the heat inactivated fetal bovine serum (FBS) and PenStrep (10,000 U/mL Penicillin, 10 mg/mL Streptomycin) were thawed completely to room temperature inside a water bath at 37°C. A volume of 50 mL FBS was added to 500 mL MEM medium (Gibco, USA) using a syringe filter. Then, 5 mL of PenStrep was added to the medium using a syringe filter. The medium mixture was shaken gently and evenly. The medium was aliquoted into 15 mL centrifuge tube before storing the tubes at 4°C. The medium was used within 3 weeks from the date of preparation before the degradation of nutrients in the medium occurs.

3.8.2 Primary cell culture

The cell line used in this study is HepG2 human liver carcinoma cell line and A549 human lung carcinoma (Riken, Japan). The A549 cell line acts as a control cell line to HepG2 cell line to compare the tissue specificities of enhancer activities in two different tissues. Upon arrival of cell vials in dry ice via oversea delivery, the cell vial was thawed by gentle agitation in 37°C water bath for around 2 min. The O-ring cap and vial cap were kept out of

water as soon as the vial contents are thawed to avoid contaminations. The vial was removed from water bath as soon as the vial contents are thawed, and the vial was cleaned with 70% ethanol to decontaminate the surface before introducing it into the laminar flow hood.

All operations beyond this step were carried out in stringent aseptic conditions. The vial contents were transferred to 15 mL centrifuge tube containing 9 mL complete culture MEM medium. Then, the centrifuge tube was subjected to centrifugation at 5,000 x g for 3 min. The supernatant was discarded, and the cell pellets were resuspended with 9 mL growth medium using pipette. To avoid excessive alkalinity of the culture medium during cell recovery, the growth medium was placed in the 37°C CO₂ incubator for 15 minutes to allow the pH to be maintained between 7.0 to 7.6. Next, the cell suspension was then dispensed into a sterile 75 cm³ T-75 culture flask. The cell culture was incubated at 37°C in a 5% CO₂ incubator (50,000 ppm CO₂) for 3 to 4 days until it is ready to be passaged.

3.8.3 Subculture of cell line

When the cell culture is ready to be passaged (normally around 80-90% confluency), the T-75 culture flask was removed from the incubator and introduced into the laminar flow hood gently. The used culture medium was removed and discarded entirely from the flask without scratching the inner surface of the flask where the cells are attached. A volume of 5 mL phosphate buffer saline (PBS) was pipetted gently into the flask and the flask was rocked gently back and forth before all the PBS was discarded. Then, 3 mL of trypsin was added into the flask and the flask was rocked gently to allow the solution to coat the cell sheet completely.

The culture flask was incubated at 37°C in 5% CO₂ incubator and it was observed every 2 to 3-min interval until the cell suspension can be visibly detected. It is important to avoid over-digestion of cells by trypsin to prevent the loss of viable cells for the next culture. The culture flask was tapped gently to dislodge any cells still adhering to the flask. A volume of 5 mL MEM culture medium was added into the flask and the medium was resuspended repeatedly to break up the cell clumps. The cell suspension was then transferred to a sterile 15 mL centrifuge tube and the tube was subjected to centrifugation at 5,000 x g for 3 min.

The supernatant was discarded before the cell pellets were resuspended thoroughly in 8 mL of MEM culture medium with 10% FBS. The cell suspension was mixed gently to dislodge all cell clumps before dispensing 1 mL of cell suspension each into 5 cryogenic vials respectively with 10% DMSO as cryoprotectant. Next, 12 mL of MEM culture medium was added to the remaining 3 mL cell suspension to top up to 15 mL of cell mixture. After mixing the cell mixture gently, 5 mL of the cell mixture was dispensed into 3 different sterile 25 cm³ T-25 culture flasks respectively. The cell culture was incubated at 37°C in a 5% CO₂ incubator for 3 to 4 days until it was ready to be passaged.

3.9 Transient transfection

3.9.1 Cell passaging

When the cell culture was ready to be passaged (normally around 70-90% confluency), the T-25 culture flask was removed from the incubator and introduced into the laminar flow hood gently. The used culture medium was removed and discarded entirely from the flask without scratching the inner surface of the flask where the cells are attached. A volume of 5

mL phosphate buffer saline (PBS) was pipetted gently into the flask and the flask was rocked gently back and forth before all the PBS was discarded. Then, 3 mL of trypsin was added into the flask and the flask was rocked gently to allow the solution to coat the cell sheet completely.

The culture flask was incubated at 37°C in 5% CO₂ incubator and it was observed every 2 to 3-min interval until the cells can be visible detected. It is important to avoid over-digestion of cells by trypsin to prevent the loss of viable cells for the next culture. The culture flask was tapped gently to dislodge any cells still adhering to the flask. A volume of 5 mL MEM culture medium was added into the flask and the medium was resuspended repeatedly to break up the cell clumps. The cell suspension was then transferred to a sterile 15 mL centrifuge tube and the tube was subjected to centrifugation at 5,000 x g for 3 min. The supernatant was discarded before the cell pellets were resuspended thoroughly in 1 mL of MEM culture medium with 10% FBS. The cell suspension was mixed gently to dislodge all cell clumps.

3.9.2 Cell counting

Cell counting is essential to ensure the cell number for transfection is optimal where by adjusting the cell number (via serial dilutions) to the desired amount as a standard. Prior to cell counting, a haemocytometer and trypan blue dye were prepared. The cover slip and the haemocytometer were cleaned with 70% ethanol and wiped dry using Kim Wipes. An equal volume of trypan blue dye was added to the vial containing 25 µL of cell culture. The cell suspension and dye were mixed gently and thoroughly. The mixture was then pipetted

slowly onto the haemocytometer covered with cover slip via the small opening at the edge between the cover slip and haemocytometer. The mixture was spread across the grid area of the haemocytometer. The haemocytometer was then placed onto the microscope and the number of cells in each grid was enumerated using a cell counter. Only the number of cells found on the shaded grids (Figure 3.1) were counted and the average reading was taken. The seeding volume was then determined by using the formula as shown below:

Seeding density =

[average number of cells counted for 5 shaded columns] x [dilution factor] x [suspension volume] x [multiplying factor for haemocytometer (10^4)]

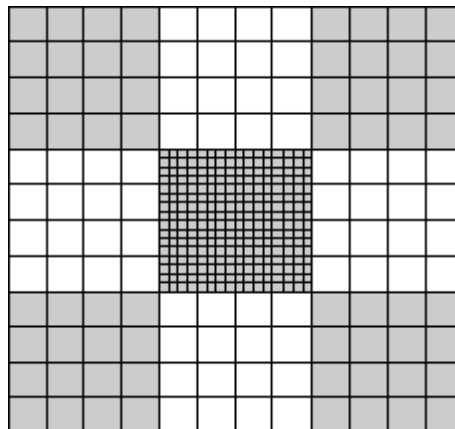


Figure 3.1: Gridlines for haemocytometer, 1 cm^3 represents 1 mm^3 .

The HepG2 and A549 cell suspension was seeded into 24-well plate according to the desired seeding volume (7500 to 10,000 cells per well). The 24-well culture plate was incubated at 37°C in a 5% CO_2 incubator overnight or until the cells are fully attached to the flask.

3.9.3 Transient transfection using Lipofectamine 2000

The transient transfection of plasmid into HepG2 and A549 cell line was conducted using Lipofectamine 2000 (Thermo Fisher Scientific, USA). For each experimental plasmid, a total of nine replicates (three technical replicates for each of the three biological replicates) were conducted alongside with positive (pGL3-control plasmid) and negative control (empty pGL4.23[*luc2*/minP] plasmid) sets. First, the 24-well culture plate was examined whether all cells seeded were attached as a single layer and the cell confluency range was from 70 to 80% under the light microscope. For each enhancer-plasmid set, 9 wells were used: 3 wells were for the experimental plasmid (pGL4.23[*luc2*/minP] containing SV40 promoter and enhancer insert) (3 technical replicates), another 3 wells were designated for positive control plasmid (pGL3-control containing both the SV40 promoter and SV40 enhancer) whereas the last 3 wells were for the negative control plasmid (pGL4.23 [*luc2*/minP] containing SV40 promoter only). Each set mentioned above was co-transfected with pRL-TK plasmid to reduce the background luminescence. The transient transfection reagents were prepared in the proportions of 200 μ L OptiMEM: 25 μ L Lipofectamine : 25 μ L plasmid for each well. The mixture was then incubated for 15 min at room temperature. Then, the mixture was transferred into each well of the 24-well culture plate before incubating the plate overnight at 37°C in a 5% CO₂ incubator. The next day, the culture medium was discarded from each well and the wells were washed with 100 μ L PBS before the new culture medium (250 μ L) was used to replenish the used culture medium in each well.

3.10 Dual-luciferase reporter assay

The luciferase luminescence was conducted using the Dual-Luciferase[®] Reporter Assay (Promega, USA). Prior to measurement of luminescence, LARII solution was prepared by resuspending the lyophilized Luciferase Assay Substrate in 10 mL of supplied Luciferase Assay Buffer II. Then, 500 μ L was aliquoted into each tube (a total of 20 tubes) and the tubes were stored at -80°C for a shelf life of one year. For each reaction, the passive lysis buffer must be freshly prepared. The 1X passive lysis buffer was prepared by adding 4 volumes of sterile double distilled water into 5X passive lysis buffer. For 9 wells reaction, 180 μ L of 5x passive lysis buffer was diluted with 720 μ L of sterile double distilled water. The used culture medium was discarded from each well before a volume of 100 μ L of 1X passive lysis buffer was added into each well and the plate was rocked gently for 15 to 20 min at room temperature.

Meanwhile, the Tecan Infinite M200 Pro luminometer was switched on and was allowed to warm up for 15 min. The instructions for the luminometer was set: wait time for 2 sec, shaking for 5 sec, premeasurement for 2 sec and measurement for 10 sec. For 9 wells reaction, a volume of 9 μ L Stop & Glo substrate was added to 450 μ L of 50X Stop & Glo buffer to produce Stop & Glo solution in a 1.5 mL microcentrifuge tube wrapped in aluminium foil. The 10 μ L cell lysate was transferred from the well of 24-well culture plate to a new well in 96-well culture plate on a well by well basis. Then, 50 μ L of LARII solution was added into each well and the mixture in each well was pipetted gently for 2 to 3 times. The culture plate was placed inside the luminometer for luciferase luminescence measurement. The first measurement reading was recorded for all 9 wells. Then, 50 μ L of

Stop & Glo solution was added into each well and mixed briefly. The second measurement reading for each well was recorded. The luciferase ratio was calculated by dividing the first reading by the second reading. The relative luciferase ratio was calculated as below:

$$\frac{\text{Experimental control set ratio}}{\text{Negative control set ratio}}$$

3.11 Motif distribution analysis

The motif distribution of all 13 candidate enhancers was inspected using MEME tool (Bailey & Elkan, 1994) *in silico* with the parameter of 15 motifs with any repetitions. The 15 motifs identified were labelled from Motif 1 to Motif 15. The possible ontology of some selected motifs was determined via GOMo (Buske *et al.*, 2010) to discover the occurrence of these motifs in relation with the resultant enhancer activities.

3.12 Transcription factor binding site analysis

The 13 candidate enhancers were subjected to transcription factor binding site analysis (TFBS) using the MATCH tool (Kel *et al.*, 2003). Different aspects of TFBS analysis such as the overall TFBS abundance, TFBS composition and TFBS distribution pattern were examined accordingly. The overall TFBS abundance was determined via inserting the enhancer sequences individually as input and the calculated overall TFBS frequency per nucleotide was recorded in a table. The TFBS composition was also determined via MATCH tool using the predefined liver-specific profile (from regulatory regions of liver-enriched transporter genes) from the TRANSFAC[®] database. The liver-specific transcription factor

binding sites (TFBS) detected from the enhancers are AP-1, HNF-1, GATA-3, C/EBP- β , NF-1, HNF-3 β , COUP-TF/HNF-4, USF, CHOP-C/EBP- α , YY1, HNF-4 α 1, TATA, CREB and NF-kappaB. Lastly, the TFBS distribution pattern from MATCH tool outcome was used to plot a histogram to depict the TFBS distribution at every 100 bp interval.

3.13 Deletion analysis

To determine the regions that contributed highly to the enhancer activities in HepG2 cell line (and A549 cell line as control cell line), enhancer pme001, pme101 and pme103 (enhancer pme001 being the highest in activity, pme103 being the second highest and pme101 being one of the lowest in activity) were deleted of several fragments (some TFBS-enriched regions were removed) from the enhancer sequences. Several internal primers were designed to induce deletions of some TFBS-enriched regions. These primers were then utilized to amplify the desired regions from the original template via PCR reactions.

For primers with suffix of “a”, “b” and “c”, the PCR reactions were conducted using the original forward or reverse primer pairing with the newly designed internal forward or reverse primer with flanking RE site (Table 4.11). For primers with suffix “D”, the newly designed forward primer was paired up with the original reverse primer and in a separate tube the newly designed reverse primer was paired up with the original forward primer in the first round of PCR reaction (Table 4.11). Then, for the second round of PCR reaction, both the PCR products from the previous reaction were pooled into one tube and diluted with dilution factor of 100 using double distilled water. This PCR product mixture was then used as template for the second PCR reaction with the use of the original forward and reverse

primer designed for the enhancer. Next, these enhancer fragments were inserted into pGL4.23[*luc2*/minP] plasmid, cloned and subsequently transfected into HepG2 and A549 cell line respectively for enhancer activity detection via the dual-luciferase assays using the protocols as mentioned above.

CHAPTER 4

RESULTS

4.1 Enhancer identification

4.1.1 Enhancer identification using iEnhancer-2L

The chromosome 18 of proboscis monkey was utilized as input data to search for candidate enhancers using iEnhancer-2L (Liu *et al.*, 2015) enhancer prediction tool. This tool was chosen because it can predict enhancers via a 2-layer predictor algorithm: the first layer distinguishes non-enhancers from enhancers while the second layer distinguishes strong enhancers from weak ones. A total of 9,960,241 strong enhancers were predicted from the chromosome 18 of proboscis monkey (Table 4.1). These strong enhancers were predicted to have strong peaks at various histone marks such as H3K27ac, H3K4me1 and H3K4me3.

Table 4.1: The outcome of iEnhancer-2L depicting the number of strong, weak and non-enhancers from chromosome 18 of *N. larvatus*.

Strong enhancers	Weak enhancers	Non-enhancers
9,960,241	11,855,707	50,983,878

4.1.2 Enhancer identification using sequence conservation from other primates

The genome-wide human liver transporter gene enhancer data was obtained from Kim *et al.* (2011) as a reference to detect similar enhancers from the genome of proboscis monkey. The human enhancers used in Kim *et al.* (2011) acted as a reference for the search for sequence conserved counterpart in the species like chimpanzee and rhesus macaque in hope to isolate similar enhancers from the proboscis monkey. Only candidate enhancers with

above 70% sequence homology aligned across non-human primates like chimpanzee and rhesus macaque were accepted (Figure 4.1) and further selection process using liver-specific profile is done using MATCH (Kel *et al.*, 2003). In Figure 4.1, the shaded area represents the query coverage of the selected human enhancer obtained from Kim *et al.* (2011) with evolutionary conserved region (ECR) sequence conservation of more than 70% as criterion as shown on top left box in the figure, for enhancer pme001.

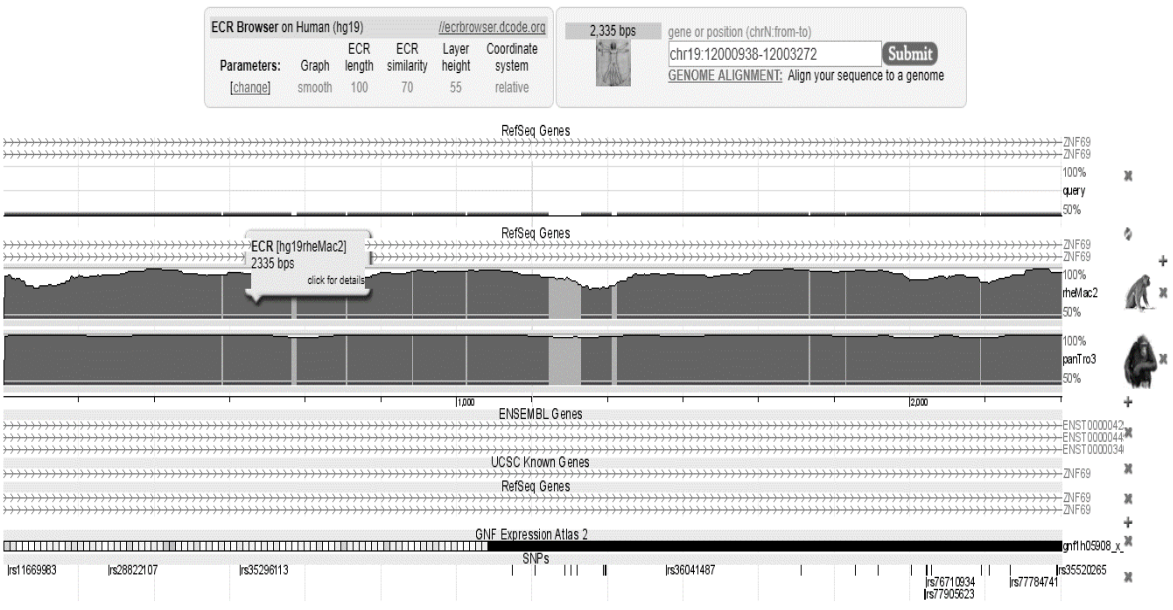


Figure 4.1: The enhancer alignment between human, chimpanzee and rhesus macaque for enhancer pme001. The shaded region indicates conservation.

4.1.3 Enhancer identification using 5 different software packages

The enhancer prediction and identification were conducted on chromosome 18 of proboscis monkey using 5 different software (namely CSI-ANN, gkm-SVM, DeepBind, iEnhancer-2L and GMFR-CNN) with distinctive predictive parameters (based on different enhancer marks) before subjecting these outcomes to bedtools “intersect” and “merge” functions for further analysis.

The “merge” tool of bedtools had yielded a total of 6434 predicted enhancers from the combination of merging 2, 3, 4 and 5 software respectively (Table 4.2). The highest number of predicted enhancers for 2 and 3 software after merging are the same which is 2494 enhancers whereas the lowest amount of 267 enhancers were merged from 5 software. The number of predicted enhancers resulted from merging 2 software using bedtools is 1179 in total.

Table 4.2: Bedtools “merge” results depicting number of enhancers predicted. The 2 software indicates all possible two software combinations from the five selected tools and similarly for 3 software and 4 software.

Number of software outcomes merged	Number of predicted enhancers
2 software	1179
3 software	2494
4 software	2494
5 software	267
Total	6434

The bedtools “intersect” outcomes for 2, 3, 4 and 5 software packages were summarized in Table 4.3, Table 4.4, Table 4.5 and Table 4.6 respectively. The intersect outcomes of software CSI-ANN and DeepBind was found to be lowest in number, that is 1707 enhancers (Table 4.3). The highest intersect outcomes came from the intersection of GMFR-CNN and gkm-SVM software with number of enhancers predicted rocketed to 57695 enhancers in total (Table 4.3). The bedtools “intersect” outcomes had depicted the lowest enhancers predicted (a total of 1707 enhancers) detected in between the intersection of 3 software, namely gkm-SVM, DeepBind and CSI-ANN (Table 4.4). The intersection of

software DeepBind, GMFR-CNN and gkm-SVM was found to top among all 3 software tools combinations, recorded to a total of 57695 enhancers (Table 4.4).

Table 4.3: Bedtools “intersect” outcomes for 2 software depicting number of enhancers predicted. Software abbreviations: A=gkm-SVM; B=GMFR-CNN; C=CSI-ANN; D=DeepBind; E=iEnhancer-2L. Redundant data was removed, for example if the set (not number) of predicted enhancers for AB and BA is the exactly same, only one of them is shown.

	Software combinations	Number of predicted enhancers
1	AB	57695
2	AC	3513
3	AD	13872
4	CB	1707
5	CD	3861
6	CE	3394
7	DB	28874
8	DE	9079
9	EA	21208
10	EB	22341

Predicted enhancers totalling at 1707 in quantity were discovered between intersection of 4 software: DeepBind, CSI-ANN, iEnhancer-2L as well as gkm-SVM. A sum of 28874 enhancers for the 4 software intersections (namely gkm-SVM, DeepBind, CSI-ANN and GMFR-CNN) had made the highest intersection outcomes among all other combinations.

Table 4.4: Bedtools “intersect” outcomes for 3 software depicting number of enhancers predicted. Software abbreviations: A=gkm-SVM; B=GMFR-CNN; C=CSI-ANN; D=DeepBind; E=iEnhancer-2L. Redundant data was removed, for example if the set (not number) of predicted enhancers for ABC and BAC is the exactly same, only one of them is shown.

	Software combinations	Number of predicted enhancers
1	ABC	3513
2	ABE	21208
3	ACD	13872
4	ACE	21208
5	ADB	57695
6	ADC	3513
7	ADE	21208
8	AEB	57695
9	AEC	3513
10	BAE	22341
11	BCA	57695
12	BCD	28874
13	BCE	22341
14	BEC	3861
15	CAB	3861
16	CAD	1707
17	CBE	3394
18	CDA	3513
19	CDB	3861
20	DAB	28874
21	DBA	13872
22	DBC	1707
23	DCE	9079
24	DEA	13872
25	DEB	28874
26	DEC	1707
27	EAC	3394
28	EBC	3394
29	EDA	21208
30	EDC	3394

Table 4.5: Bedtools “intersect” outcomes for 4 software depicting number of enhancers predicted. Software abbreviations: A=gkm-SVM; B=GMFR-CNN; C=CSI-ANN; D=DeepBind; E=iEnhancer-2L. Redundant data was removed, for example if the set (not number) of predicted enhancers for ABCD and BDCA is exactly the same, only one of them is shown.

	Software combinations	Number of predicted enhancers
1	ACBE	21208
2	ACDE	21208
3	ACEB	57695
4	BCAE	22341
5	BCDE	22341
6	BDAE	22341
7	BDEA	57695
8	CABD	1707
9	CABE	3394
10	CAED	1707
11	CBEA	3513
12	CBED	1707
13	CDAE	3394
14	CDBA	3513
15	CDBE	3394
16	CADB	3861
17	CDEA	3513
18	CDEB	3861
19	CEAB	3861
20	DCBA	13872
21	CBEA	3513
22	CBED	1707
23	CDAE	3394
24	CDBA	3513
25	CDBE	3394
26	CADB	3861
27	CDEA	3513
28	CDEB	3861
29	CEAB	3861
30	DCBA	13872

Table 4.6: Bedtools “intersect” outcomes for 5 software depicting number of enhancers predicted. Software abbreviations: A=gkm-SVM; B=GMFR-CNN; C=CSI-ANN; D=DeepBind; E=iEnhancer-2L. Redundant data was removed, for example if the set (not number) of predicted enhancers for ABCDE and BDACE is exactly the same, only one of them is shown.

	Software Combinations	Number of predicted enhancers
1	CDABE	3394
2	CBEDA	3513
3	DCABE	9079
4	DABEC	1707
5	ABECD	13872
6	AECDB	57695
7	BCAED	28874
8	BAEDC	3861
9	EDCAB	22341
10	EBDCA	21208

4.2 Selection of enhancers and primer design

The outcomes (predicted enhancers) obtained from each prediction method mentioned above were subjected to MATCH tool (Kel *et al.*, 2003) analysis using five different profiles, namely liver-specific, muscle-specific, immune cells-specific, cell cycle-specific and best selection profile (Table 4.7).

A total of 13 gene-specific primers were designed to target and isolate enhancers from the genomic DNA of proboscis monkey (Table 4.8). The length of primers was in the range between 24 and 28 base pair and their GC content was designed to range between 39.3% and 56.5%. *Bam*HI and *Sal*II restriction enzyme sites were added to 5’ end of each forward and reverse primer respectively for downstream cloning processes.

Table 4.7: The transcription factor binding site (TFBS) analysis outcome of individual candidate enhancers. TFBS per bp is the number of TFBS found within the enhancer sequences divided by the total length of enhancers.

Enhancer	Length (bp)	Liver-specific		Muscle-specific		Immune cells- specific		Cell cycle-specific		Best selection	
		TFBS	TFBS per bp	TFBS	TFBS per bp	TFBS	TFBS per bp	TFBS	TFBS per bp	TFBS	TFBS per bp
pme001	795	42	0.052830	12	0.015094	55	0.069182	40	0.050314	23	0.028931
pme002	654	22	0.033639	17	0.025994	75	0.114679	30	0.045872	11	0.016820
pme003	685	38	0.055474	19	0.027737	77	0.112409	28	0.040876	17	0.024818
pme004	741	26	0.035088	17	0.022942	81	0.109312	31	0.041835	12	0.016194
pme101	913	87	0.095290	13	0.014239	87	0.095290	35	0.038335	32	0.035049
pme102	587	43	0.073254	4	0.006814	50	0.085179	12	0.020443	15	0.025554
pme103	1145	75	0.065502	18	0.015721	114	0.099563	31	0.027074	35	0.030568
pme104	306	27	0.088235	8	0.026144	31	0.101307	16	0.052288	13	0.042484
pme105	573	57	0.099476	13	0.022688	76	0.132635	27	0.047120	26	0.045375
pme201	308	19	0.061688	7	0.022727	33	0.107143	14	0.045455	9	0.029221
pme202	573	29	0.050611	12	0.020942	47	0.082024	33	0.057592	15	0.026178
pme203	475	45	0.094737	11	0.023158	40	0.084211	12	0.025263	25	0.052632
pme204	941	65	0.069075	19	0.020191	95	0.100956	47	0.049947	32	0.034006

Table 4.8: Gene-specific primers designed to isolate enhancers and their characteristics. Restriction enzyme sites (underlined) where *Bam*HI site was added to 5' end of each forward primer whereas *Sal*I site was added to 5' end of each reverse primer. Letters in bold indicate additional base pair for optimal restriction digestion.

Enhancer	Sequences (5'-3')	Length (bp)	GC Content (%)
pme001	F: A <u>A</u> GGATCCTGGGCCTCGTAGTTAAA R: <u>ATGTCGACACACACACTTCTACGGT</u>	25 25	48.0 48.0
pme002	F: A <u>T</u> GGATCCGACTGTGCTTTTCCCCTG R: A <u>AGTCGACTTTTGTGTTGATGCTGTTG</u>	26 28	53.8 39.3
pme003	F: A <u>A</u> GGATCCTTTGACTCCATGTCTCAC R: A <u>AGTCGACCAAGACAATGGGGAAAAT</u>	26 26	46.2 42.3
pme004	F: A <u>T</u> GGATCCAGCACTGGGACTGATA R: <u>ATGTCGACAGTGGGGACTTTTGTGTTGTT</u>	24 27	50.0 44.4
pme101	F: A <u>T</u> GGATCCGGCAGGAGAATTGCTTGAA R: <u>ATGTCGACAAAATTAGCTGGGCCTGGT</u>	27 27	48.1 48.1
pme102	F: A <u>T</u> GGATCCTACCCAAATAGTGCTTGCTG R: A <u>AGTCGACGCCACATTTCAAGTGCTCA</u>	28 27	46.4 48.1
pme103	F: A <u>A</u> GGATCCTTTCCAATCTGACCAGGTG R: <u>ATGTCGACTACCCTGAAACTTTGCTGA</u>	27 27	48.1 44.4
pme104	F: A <u>T</u> GGATCCGCTGGCTCTCAGTTTCCTG R: A <u>AGTCGACAGGCATGAGCCACTACAT</u>	26 26	53.8 50.0
pme105	F: A <u>T</u> GGATCCCTTCACAACCAACGTTTCAT R: <u>ATGTCGACTATGGAAGGAGCCTTTGG</u>	27 26	44.4 50.0
pme201	F: A <u>A</u> GGATCCATGTGTTCATGCGTGTGTA R: A <u>AGTCGACTGACACGTCACATACGAAAA</u>	26 28	46.2 42.9
pme202	F: A <u>A</u> GGATCCGGGTAGACAGCAAGGACA R: A <u>AGTCGACGGCAAACCTGCTTCAGGAT</u>	26 26	53.8 50.0
pme203	F: A <u>A</u> GGATCCACATTTGGCAGACATAGT R: A <u>AGTCGACGAGGTATGTGTCCAAAGCAA</u>	26 28	42.3 46.4
pme204	F: A <u>A</u> GGATCCCAATCACCTCTCACCA R: A <u>AGTCGACTCCCAAATCTTTTCCTTGG</u>	24 26	50.0 46.2

4.3 Genomic DNA extraction

The genomic DNA of proboscis monkey from stool samples were extracted and subsequently run on 1% agarose gel. Figure 4.2 depicts a band with size approximated at above 10 kb in all four lanes (four technical replicates were done). This indicates the presence of genomic DNA and its feasibility for further downstream procedures was further justified. The

amplification of *cytochrome b* gene via PCR using universal primers by Irwin *et al.* (1991) [forward primer L14724 (5'-GACTTGAAAAACCAACCGTTG-3') and reverse primer H15915R (5'-GGAATTCATCTCTCCGGTTTACAAGAC-3')] was conducted at 50.5°C annealing temperature and had yielded a band between 1000 and 2000 bp which tallies with the expected size of around 1500 bp (Figure 4.3) . This has further proven the successful isolation of proboscis monkey genome following the sequencing and BLASTn analysis with 99% similarity to that mitochondrion genome sequences of proboscis monkey found in the GenBank database with accession number: KM889667.1 (Figure 4.4).



Figure 4.2: Genomic DNA extraction from stool of Proboscis monkey. Lane L: TransGen 1 kb DNA ladder; Lane 1-4: Genomic DNA extracted from the same stool sample (four technical replicates).

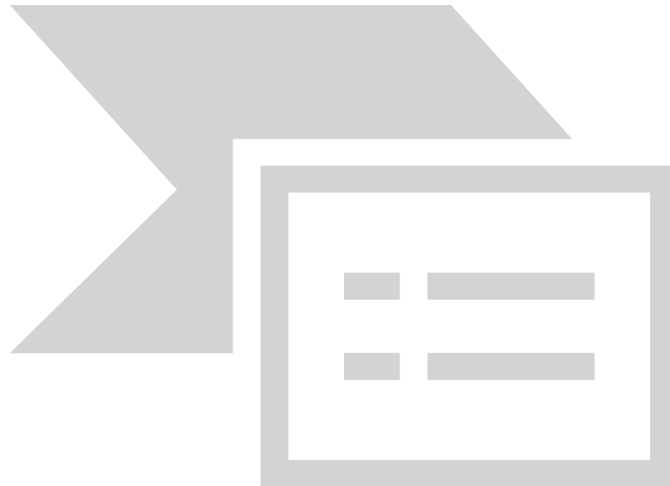


Figure 4.3: Amplification of housekeeping gene, cytochrome b gene. Lane L: TransGen 1 kb DNA ladder; Lane 1: *cytochrome b* gene (~1500 bp) amplified at 50.5°C annealing temperature.



Figure 4.4: Part of the pairwise aligned sequences from BLASTn analysis of cytochrome b gene of proboscis monkey.

4.4 Amplification of enhancers

All 13 gene-specific primers were utilized for gradient PCR optimization respectively (Figure 4.5 and Figure 4.6) to determine the optimum temperature profile (including the optimum annealing temperature) for each enhancer (Table 4.9). The annealing temperatures of these enhancers were found to range from 45.2°C up to 63.0°C. A discrete single band was found in all amplifications with the largest amplicon band at around 1200 bp and the smallest amplicon at around 300 bp which are consistent with the previously predicted sizes for these enhancers.



Figure 4.5: PCR amplification of enhancers. Lane L: TransGen 100 bp DNA ladder; Lane 1-4: pme001-pme004; Lane 5-7: pme101-pme103.



Figure 4.6: PCR amplification of enhancers. Lane L: TransGen 100 bp DNA ladder; Lane 1 and 2: pme104 and pme105; Lane 3-6: pme201-pme204.

Table 4.9: The optimum annealing temperatures for each enhancer primer pair.

Enhancer	Optimum Annealing Temperature (°C)
pme001	51.7
pme002	62.7
pme003	58.2
pme004	63.0
pme101	57.0
pme102	51.6
pme103	51.2
pme104	45.2
pme105	50.8
pme201	50.0
pme202	60.5
pme203	47.8
pme204	47.8

4.5 Cloning of enhancers

The candidate enhancers were first cloned into pGEM[®]-T Easy vector (Promega, USA) and sequenced accordingly. Then, these recombinant pGEM[®]-T Easy vector constructs with enhancer inserts were double digested by both endonucleases *SalI* and *BamHI* and then each enhancer insert was subcloned into pGL 4.23 [*luc2/minP*] (Promega, USA). These enhancer inserts were inserted at downstream of minimal promoter and luciferase gene in the plasmid.

4.6 Motif distribution

The motif distribution of enhancers was inspected using MEME tool (Bailey & Elkan, 1994) with the parameter of 15 motifs with any repetitions (Figure 4.7), emulating the methodology by Madrid-Espinoza *et al.* (2018). The enhancer pme104, pme105, pme201 and pme202 are the ones with few motifs found within their sequences. The enhancer pme004 is most abundance in motifs with 16 motifs found within the 741 bp long sequences. The most occurring motif found across these enhancers is Motif 10 with occurrence of 19 times while the enhancer pme102 has the highest number of Motif 10. The enhancer pme101 has the highest record of Motif 2 at the forward strand.

Interestingly, Motif 5 is only found in enhancer pme001-pme004 cluster, most probably suggesting a functional role. Similarly, Motif 10 can be found in enhancer pme101-105 and in pme003 and pme203, this motif may not be cluster-specific, but it may be involved in some functional roles in the cluster. The enhancer pme103 has much lesser motifs compared to enhancer pme002 and pme004 despite of its longer sequences compared to both the latter, moreover enhancer pme103 has extra Motif 4. Besides, enhancer pme101, pme001 and pme204

all have two Motif 8 but the orientations of the two motifs within the sequence differs: for pme001, there is one motif on each forward and reverse strand; for pme101, both motifs are on forward strand; for pme204, both the motifs are on the reverse strand.

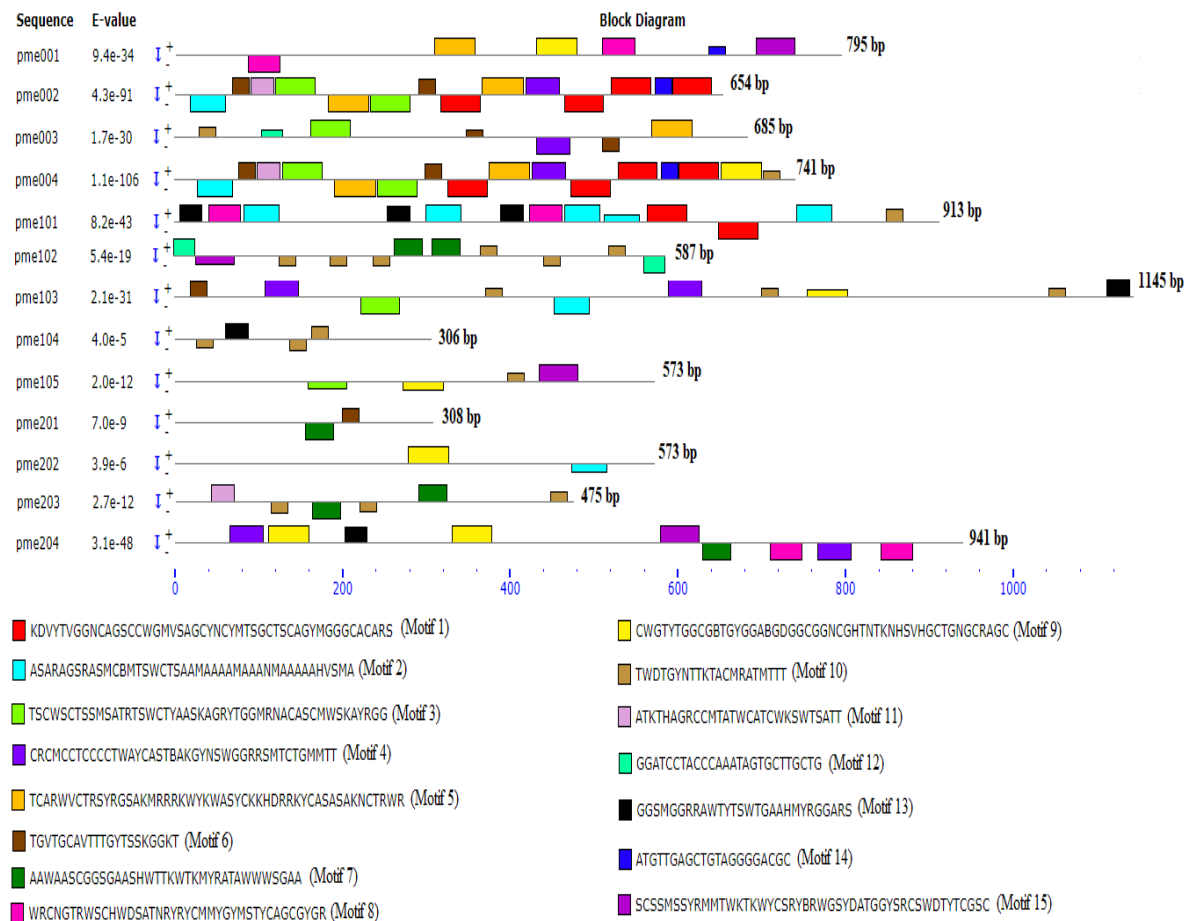


Figure 4.7: The motif distribution of each enhancer with the parameter of 15 motifs.

4.7 Transcription factor binding site (TFBS) analysis

All 13 enhancers were subjected to transcription factor binding site (TFBS) analysis using MATCH tool (Kel *et al.*, 2003). Different aspects of TFBS analysis such as overall TFBS abundance, TFBS composition and TFBS distribution pattern were investigated.

4.7.1 Overall TFBS abundance

The overall TFBS frequency per nucleotide for each enhancer were predicted using MATCH tool (Kel *et al.*, 2003). A bar graph was plotted for all 13 enhancers as shown in Figure 4.8. Enhancer pme105 has the highest record of TFBS frequency per nucleotide (close to 0.1). The second highest scoring enhancer are enhancer pme101 and pme203 with values at around 0.095. The lowest scoring duo among all others are the enhancer pme002 and pme004 with scorings around 0.035.

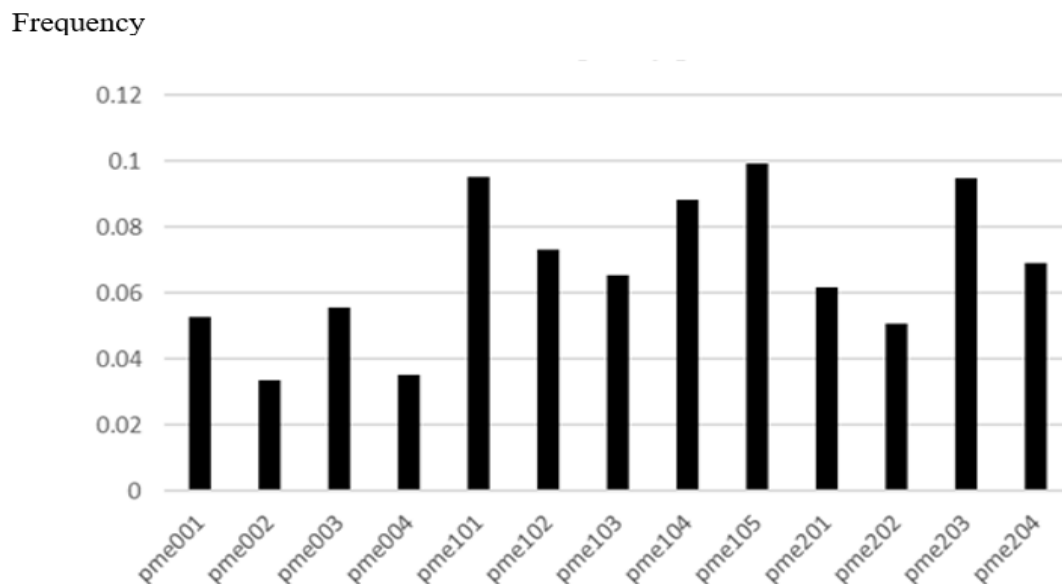


Figure 4.8: The predicted TFBS frequency per nucleotide bar graph across 13 enhancers.

4.7.2 TFBS composition

All 13 chosen enhancers were subjected to MATCH tool (Kel *et al.*, 2003) analysis to determine the transcription factor binding site composition within the enhancer sequences. The profile used to run MATCH tool (Kel *et al.*, 2003) on the enhancers is the predefined liver-specific profile (from regulatory regions of liver-enriched transporter genes) from the

TRANSFAC[®] database. The liver-specific transcription factor binding sites (TFBS) detected from the enhancers are AP-1, HNF-1, GATA-3, C/EBP- β , NF-1, HNF-3 β , COUP-TF/HNF-4, USF, CHOP-C/EBP- α , YY1, HNF-4 α 1, TATA, CREB and NF-kappaB (Figure 4.9 & Table 4.10). Enhancer pme103 has the highest amount of C/EBP- β TFBS (a total of 32 TFBS). Besides, enhancer pme101 was also predicted to contain the highest amount of HNF-3 β TFBS with a total of 29 binding sites. Notably, both enhancer pme203 and pme204 have the highest amount of HNF-1 and NF-1 TFBS respectively, totalling to 14 for each TFBS. Enhancer pme001 has 10 AP-1 TFBS, making it the most AP-1-enriched enhancers among all.

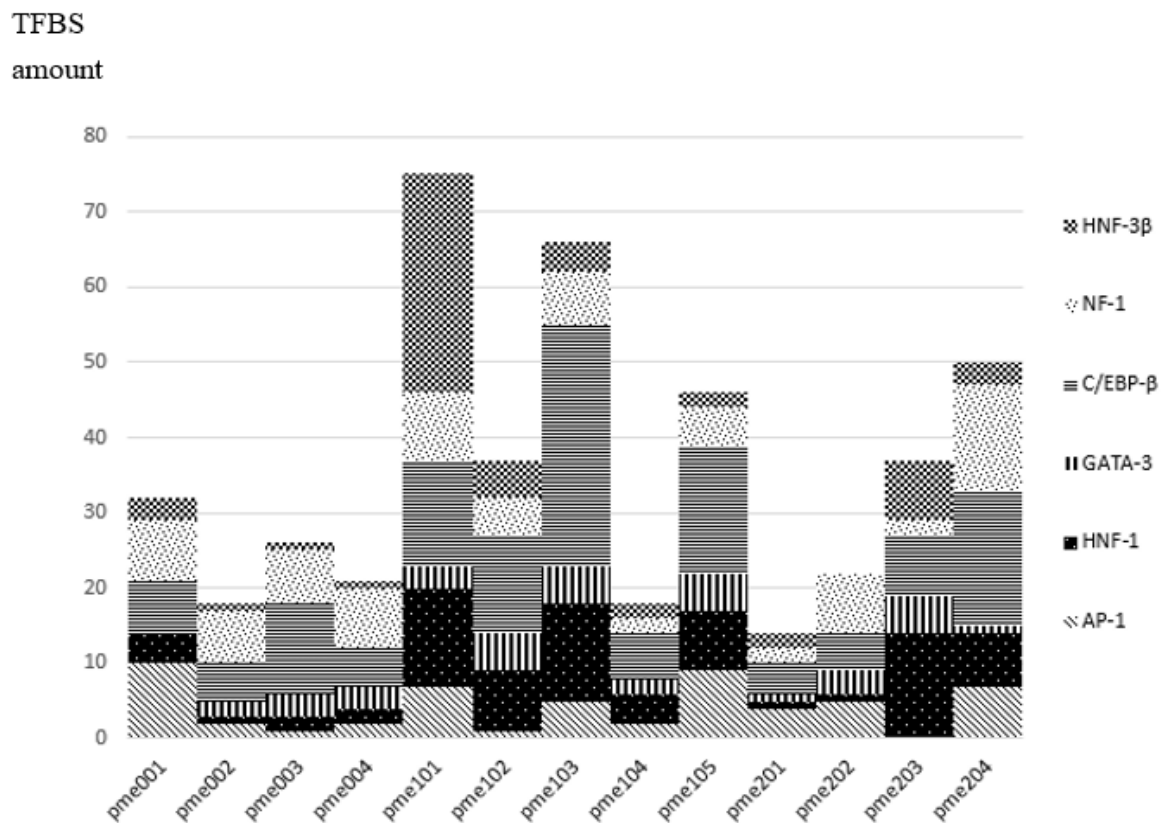


Figure 4.9: The breakdown of predicted TFBS amount across 13 enhancers.

Table 4.10: The enumeration of all transcription factors predicted within candidate enhancer sequences.

pme	AP-1	HNF-1	GATA- 3	C/EBP- β	NF-1	HNF-3 β	COUP- TF/ HNF-4	US F	CHOP- C/EBP- α	YY1	HNF- 4 α 1	TATA	CRE B	NF- kappa B
001	10	4	0	7	8	3	1	4	4	1	0	0	0	0
002	2	1	2	5	7	1	1	2	0	1	0	0	0	0
003	1	2	3	12	7	1	1	7	3	0	1	0	0	0
004	2	2	3	5	8	1	1	2	1	1	0	0	0	0
101	7	13	3	14	9	29	1	4	4	0	1	0	0	1
102	1	8	5	13	5	5	0	0	4	0	0	2	0	0
103	5	13	5	32	7	4	2	4	1	0	0	2	0	0
104	2	4	2	6	2	2	0	4	3	0	0	2	0	0
105	9	8	5	17	5	2	3	2	4	1	0	0	0	0
201	4	1	1	4	2	2	1	2	0	0	0	1	1	0
202	5	1	3	5	8	0	0	4	0	1	0	1	1	0
203	0	14	5	8	2	8	0	4	0	0	0	4	0	0
204	7	7	1	18	14	3	0	6	8	1	0	0	0	0

4.7.3 TFBS distribution pattern

The TFBS distribution data of all 13 enhancers were obtained from MATCH tool (Kel *et al.*, 2003). Based on the data, a histogram was plotted to display the TFBS expected to be found in the specific 100 bp interval regions (to focus on more detailed view of TFBS patterns in the 100 bp resolution) within the enhancers (Figure 4.10). There are some signature patterns found exclusively on certain enhancers. For instance, only enhancer pme001 has four AP-1 adjacent pairs at four different 100 bp intervals. The enhancer pme103 has two sets of five adjacent C/EBP- β TFBS as well as two triplets of HNF-1 TFBS. Moreover, enhancer pme101 has two huge clusters of HNF-3 β TFBS and this may have impact on its activity in cell and tissue. It is also interesting to look at the signature TFBS patterns within the three main clusters of enhancers. At a glance, the most abundance TFBS patterns found within the pme001-pme004 cluster is the pairs of C/EBP- β and NF-1 TFBS which may serve as driver in enhancer activities for this cluster. The driver TFBS found in the cluster pme101-pme105 are the C/EBP- β and HNF-1 TFBS whereas the cluster pme201-pme204 has shown no shared TFBS patterns among them.

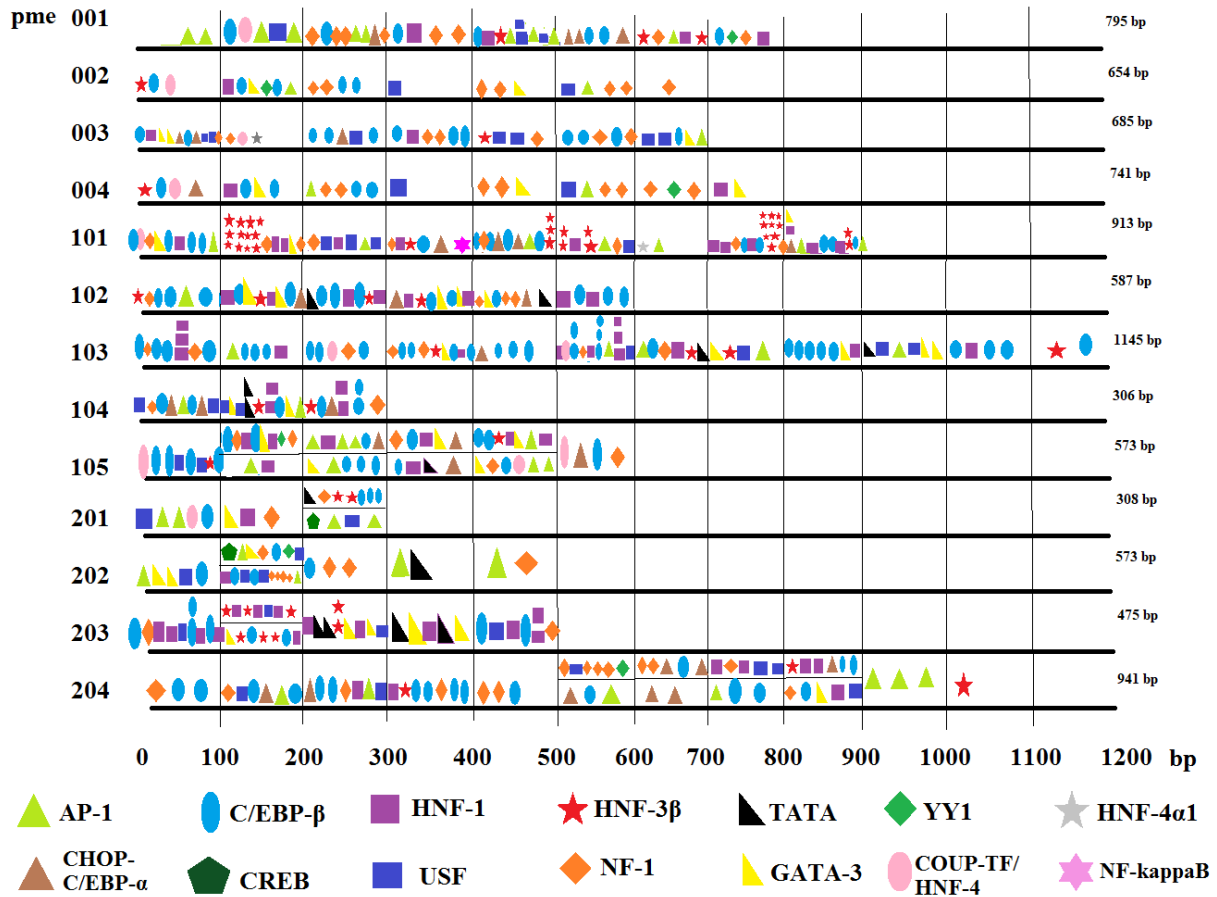


Figure 4.10: The TFBS distribution per 100 bp interval of all 13 enhancers.

4.8 Luciferase reporter assay

The enhancer activity reporter assay was conducted across 13 chosen enhancers by inserting them into an expression plasmid containing luciferase gene and subsequently transfecting the respective recombinant plasmids into HepG2 cell line for luminescence detection. The results of the relative enhancer luciferase assay and statistical data (one-way ANOVA, post-hoc Tukey's Test, $p < 0.05$) were depicted in Figure 4.11. The highest enhancer activity was observed in enhancer pme001 and it is the only enhancer that has higher activity than the positive control with the SV40 viral enhancer insert. The enhancer pme104 has the lowest activity among all in HepG2 cell line. It is also interesting to see that the cluster where enhancer pme001-pme004 reside in has overall higher activities compared to the other two

clusters (with pme003 as exception). The second cluster where pme101-pme105 reside in displayed a balanced population of enhancers in terms of activities in which two enhancers (pme102 and pme103) has higher activities and three other enhancers (pme101, pme104 and pme105) showed lower activities. The third cluster which encompasses pme201-pme204 depicted no big differences in terms of their enhancer activities.

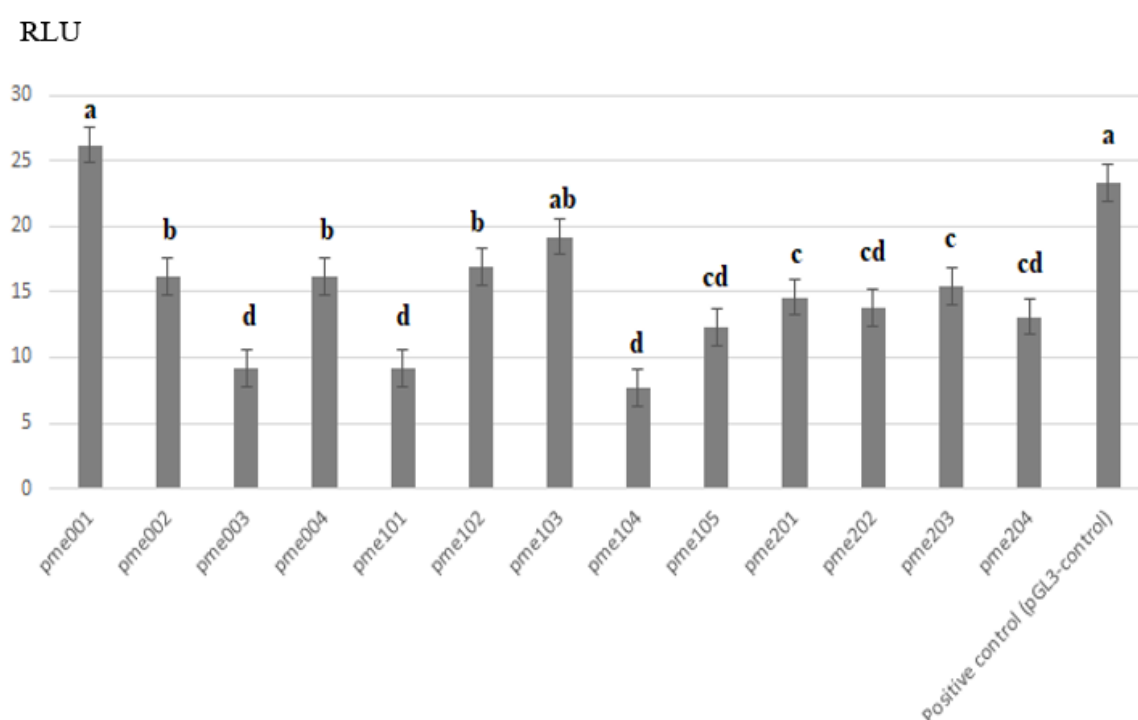


Figure 4.11: The measured normalized relative enhancer luciferase activity (RLU) (relative to non-insert control, NIC=1) across 13 enhancers. Statistical analysis was conducted using one-way ANOVA and post-hoc Tukey's Test with $p < 0.05$. Significantly different data are represented by different alphabets.

4.9 Deletion analysis

Deletion analysis were conducted using the designed internal primers (Table 4.11) to investigate the effects of removal of certain TFBS-enriched regions from the enhancer sequences on the resultant enhancer activities. The candidate enhancers investigated were

enhancer pme001, pme101 and pme103 with their enhancer fragments and deletion illustration as shown in Figure 4.13, Figure 4.14 and Figure 4.15 respectively. These enhancer fragments were isolated via PCR approaches and these DNAs were size-separated via agarose gel electrophoresis as shown in Figure 4.12. Then, these enhancer fragments were subjected to cloning into *E. coli* via the aid of the pGL4.23 [*luc*/minP] plasmid as the biological delivery vehicle.

The yielded recombinant plasmids were then subjected to transient transfection into HepG2 cell line for enhancer activity detection and the outcomes were depicted in Figure 4.13, Figure 4.14 and Figure 4.15 respectively. In general, the deletion analysis had yielded the information about the contribution of significant TFBS-enriched regions towards enhancer activity. The deletion of AP-1 TFBS-enriched regions had resulted in reduction in enhancer activity in enhancer fragments of enhancer pme001 in HepG2 cell line. However, the deletion of HNF-3 β -enriched regions from the enhancer pme101 had resulted in increment of enhancer activity in HepG2 cell line. The deletion of selected regions within the enhancer pme103 had also resulted in decline in enhancer activity in HepG2 cell line. In A549 cell line, all enhancer fragments generally have higher enhancer activity as compared to their parent enhancer.

Table 4.11: The list of internal primers used for enhancer fragments amplification via PCR. Restriction enzyme sites (underlined) where *Bam*HI site was added to 5' end of each forward primer whereas *Sal*I site was added to 5' end of each reverse primer. Letters in bold indicate additional base pair for optimal restriction digestion.

Enhancer Fragment	Sequences (5'-3')	Length (bp)	GC Content (%)
pme001a	F: A <u>AGGATCCT</u> GGGCCTCGTAGTTAAA	25	48
	R: A <u>AGTCGAC</u> AGCGCAGTGCTTCTTCG	25	56
pme001b	F: A <u>AGGATCCT</u> GGGCCTCGTAGTTAAA	25	48
	R: A <u>AGTCGAC</u> CAGCACAATGTCGCGAA	25	52
pme001c	F: A <u>AGGATCCT</u> GGGCCTCGTAGTTAAA	25	48
	R: A <u>AGTCGAC</u> CATCCTGGCTGATTTTT	25	44
pme001D	F: CGCGACATTGTGCTGTGCTGTATCTATGCT	30	50
	R: AGCATAGATACAGCACAGCACAATGTCGCG	30	50
pme101D1	F: AGCGAGACTTTGACTATCTAGAGCGTTGTG	30	47
	R: CACAACGCTCTAGATAGTCAAAGTCTCGCT	30	47
pme101D2	F: ATGCTCTAACTGTAGGAACTGGCACACTGC	30	50
	R: GCAGTGTGCCAGTTCCTACAGTTAGAGCAT	30	50
pme101D3	F1: AGCGAGACTTTGACTATCTAGAGCGTTGTG	30	47
	R1: CACAACGCTCTAGATAGTCAAAGTCTCGCT	30	47
	F2: ATGCTCTAACTGTAGGAACTGGCACACTGC	30	50
	R2: GCAGTGTGCCAGTTCCTACAGTTAGAGCAT	30	50
pme103a	F: A <u>AGGATCCT</u> CACACCTACCCTTTTCG	25	52
	R: A <u>TGTCGAC</u> TACCCTGAACTTTGCTGA	27	44
pme103D1	F: GGCTTCCCTTCCTACGTTACCAGCTATGCT	30	53
	R: AGCATAGCTGGTAACGTAGGAAGGGAAGCC	30	53
pme103D2	F: AATCGTAAATCCTAAAAAGTGTCTTTTAGT	30	27
	R: ACTAAAAGACACTTTTTTAGGATTTACGATT	30	27

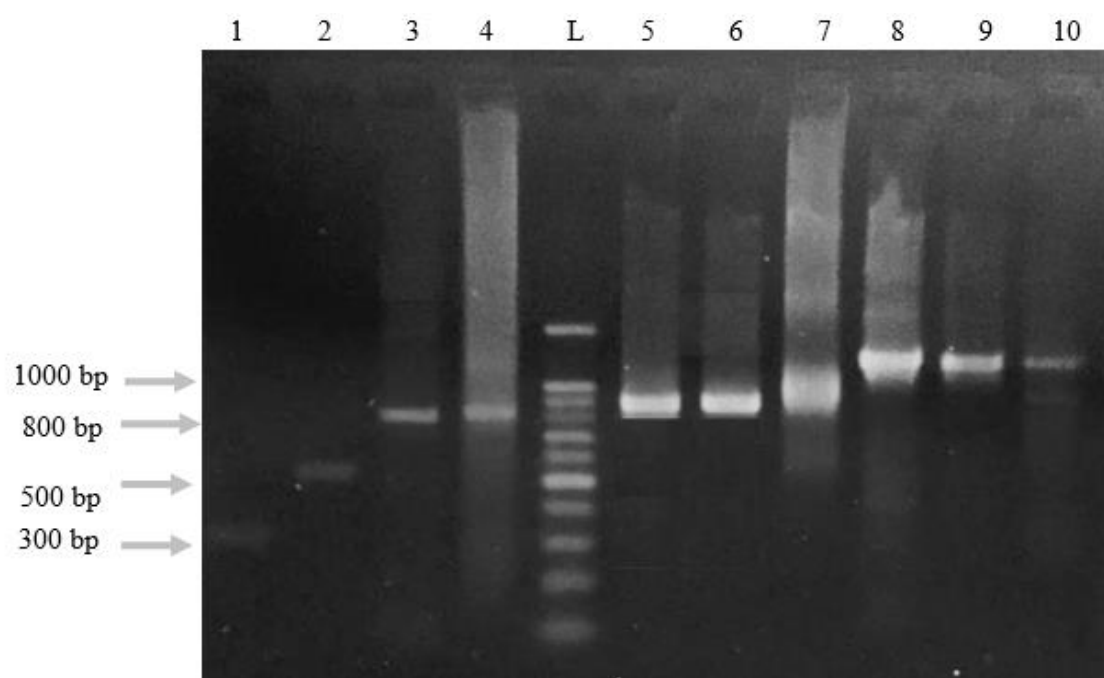


Figure 4.12: PCR amplification of enhancer fragments. Lane L: TransGen 100 bp DNA ladder; Lane 1: pme001a; Lane 2: pme001b; Lane 3: pme001c; Lane 4: pme001D; Lane 5: pme101D1; Lane 6: pme101D2; Lane 7: pme101D3; Lane 8: pme103a; Lane 9: pme103D1; Lane 10: pme103D2.

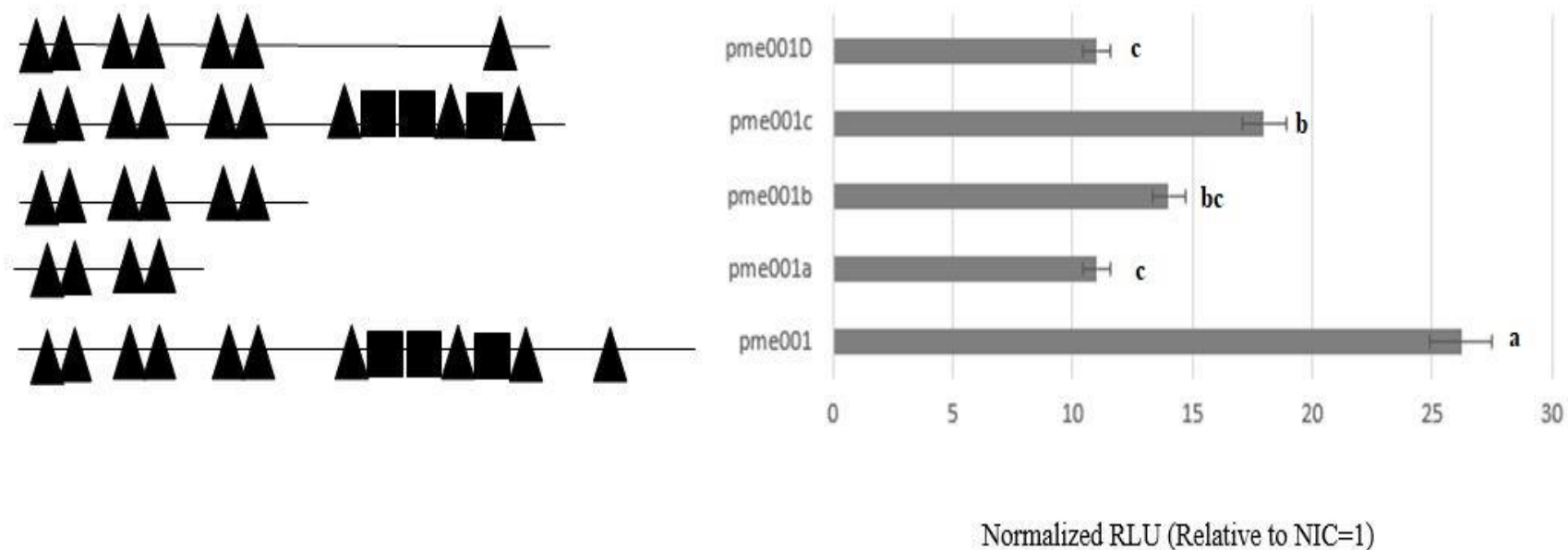


Figure 4.13: Overview of selected TFBS distribution and deletion analysis of enhancer pme001 in pGL4.23 [*luc*/minP], activities as measured in HepG2 cell line. Deletion constructs are depicted on the left. AP-1 TFBSs are represented by black triangles; USF TFBSs are represented by black boxes. Enhancer activities are depicted on the right with values indicating normalized firefly luciferase: *Renilla* luciferase activities relative to no-insert control (NIC). Results are mean $\pm$ S.E. (n=3). Mean values were subjected to One-Way ANOVA and Tukey's test, $p < 0.05$. Mean values represented by different alphabets are significantly different.

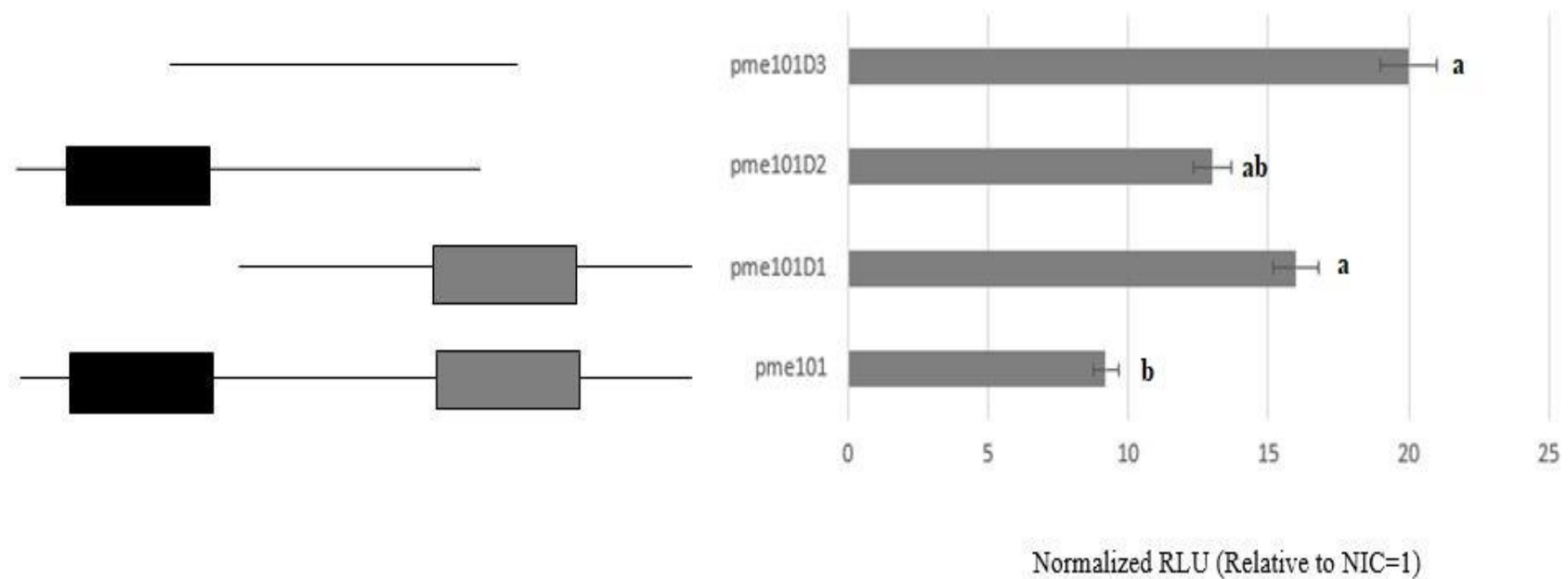


Figure 4.14: Overview of selected TFBS distribution and deletion analysis of enhancer pme101 in pGL4.23 [*luc*/minP], activities as measured in HepG2 cell line. Deletion constructs are depicted on the left. The first HNF-3 β TFBS-enriched region (12 HNF-3 β TFBSs) is represented by black rectangles; The second HNF-3 β TFBS-enriched region (8 HNF-3 β TFBSs) is represented by grey rectangles. Enhancer activities are depicted on the right with values indicating normalized firefly luciferase: *Renilla* luciferase activities relative to no-insert control (NIC). Results are mean $\pm$ S.E. (n=3). Mean values were subjected to One-Way ANOVA and Tukey's test, $p < 0.05$. Mean values represented by different alphabets are significantly different.

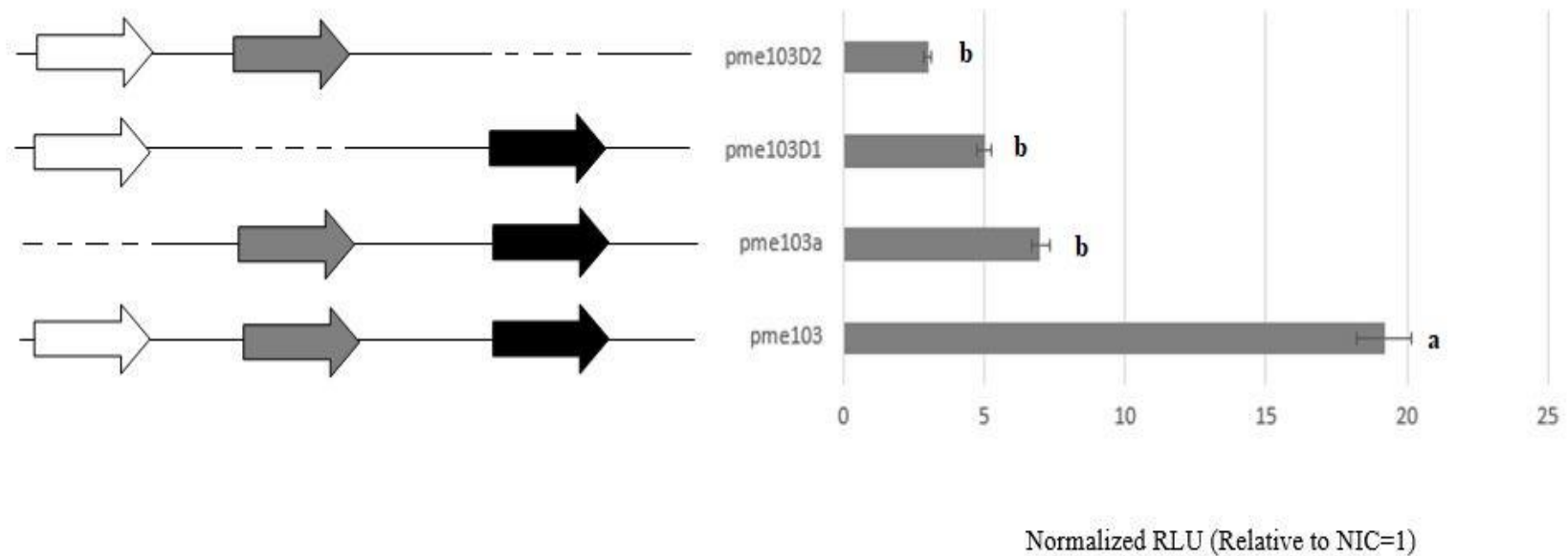


Figure 4.15: Overview of selected TFBS distribution and deletion analysis of enhancer pme103 in pGL4.23 [*luc/minP*], activities as measured in HepG2 cell line. Deletion constructs are depicted on the left. The 1-100 bp region is represented by white arrows; The 500-599 bp are represented by grey arrows; The 800-899 bp are represented by black arrows; Deleted regions are represented by dotted lines. Enhancer activities are depicted on the right with values indicating normalized firefly luciferase: *Renilla* luciferase activities relative to no-insert control (NIC). Results are mean $\pm$ S.E. (n=3). Mean values were subjected to One-Way ANOVA and Tukey's test, $p < 0.05$. Mean values represented by different alphabets are significantly different.

4.10 Comparison of enhancer activity across two cell lines

The activity of the 13 liver-specific candidate enhancers were also tested on A549 cell lines using the same protocol as optimized on HepG2 cell line sets (Figure 4.16). All the tested enhancers had shown a decline in enhancer activity in A549 cell line as compared to HepG2 cell line. Besides, the enhancer fragments resulted from the deletion analysis were also tested for their activities in A549 cell lines and their resultant activities were compared directly alongside with that of their parent enhancer sequences (Figure 4.17). In general, the enhancer fragments have shown higher activities as compared to their parent enhancers in A549 cell line (except for enhancer fragments pme001a, pme001b and pme001c).

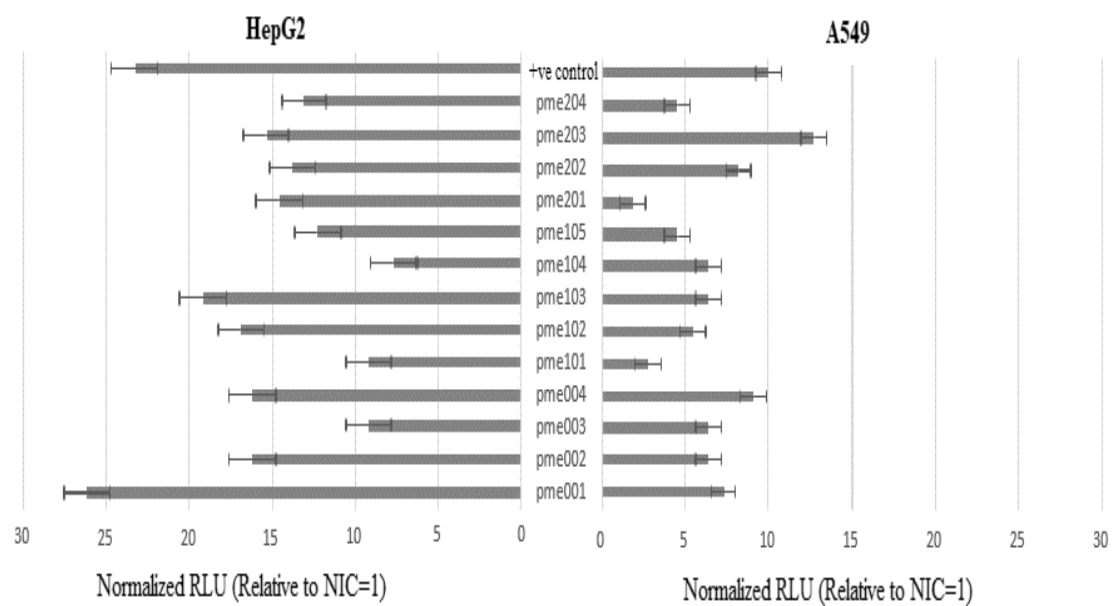


Figure 4.16: The normalized relative luciferase unit (relative to NIC=1) of enhancers in both HepG2 and A549 cell lines.

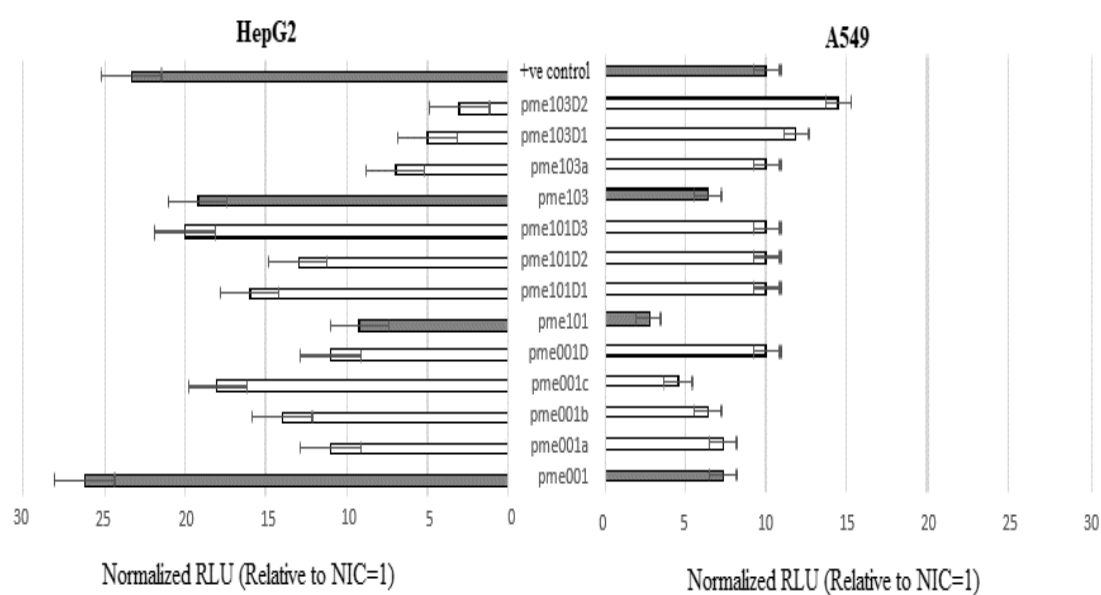


Figure 4.17: The normalized relative luciferase unit (relative to NIC=1) of enhancer fragments in both HepG2 and A549 cell line. The filled bar represents original parent enhancer whereas the empty bar represents deleted fragment.

CHAPTER 5

DISCUSSION

5.1 *In silico* enhancer prediction and enhancer activity

The *in silico* enhancer prediction had been long introduced into the enhancer identification field of study decades ago, hoping that it can serve as a highly effective approach not to replace but to complement experimental enhancer prediction approaches (wet lab approaches). To date, numerous enhancer predictor tools (such as iEnhancer-2L, gkm-SVM, GMFR-CNN, CSI-ANN, as well as DeepBind (Firpi *et al.*, 2010; Ghandi *et al.*, 2014; Alipanahi *et al.*, 2015; Liu *et al.*, 2016; Wong *et al.*, 2016)) have emerged to facilitate the accurate prediction of enhancers from various genomes of species of interest. However, there is still not a single enhancer mark or golden standard been successfully established yet, not to mention an enhancer predictor tool that can cater to our needs to unravel the enhancer landscape thoroughly. In this study, three different *in silico* enhancer prediction methods were utilized, namely iEnhancer-2L approach, primate conservation approach as well as the 5-software approach.

The iEnhancer-2L approach had successfully yielded massive prediction of 9,960,241 strong enhancers, 11,855,707 weak enhancers as well as 50,983,878 non-enhancers from the chromosome 18 of proboscis monkey via employment of the 2-layer predictor algorithm. Looking at the overall results of enhancer activities, the enhancer with the highest activity across all 13 enhancers (pme001) comes from this approach. Besides, this approach had also successfully predicted another two moderately-high-activity enhancers (pme002 and pme004) and one low-activity enhancer (pme003). This may suggest that this approach is indeed useful and capable in predicting strong liver-specific enhancers

from genome of proboscis monkey.

The primate conservation approach uses the sequence conservation and previous literature (Kim *et al.*, 2011) as starting points to determine primate-specific enhancers from genome-wide based on regulatory regions of liver-enriched genes as described by Kim *et al.* (2011). In this approach, the enhancers identified varies greatly in terms of their activities. The high activity enhancers in this approach are the enhancer pme102 and pme103. For the other three low activity enhancers (pme101, pme104 and pme105), their activities are about two-fold lower compared to pme102 and pme103. A balanced distribution curve can be seen across these five enhancers in this approach. Thus, the prediction ability of this approach in terms of finding strong liver-specific enhancers is moderate, suggesting a non-bias way to explore a wide distribution of a population of both strong and moderately strong enhancers.

The five-software approach (iEnhancer-2L, gkm-SVM, GMFR-CNN, CSI-ANN, as well as DeepBind (Firpi *et al.*, 2010; Ghandi *et al.*, 2014; Alipanahi *et al.*, 2015; Liu *et al.*, 2016; Wong *et al.*, 2016) is deemed to be one of the best approaches to capture the entire enhancer landscape via the consolidation of the power of different enhancer predictor tools in hope that when their outcomes are consolidated, they work hand in hand and in tandem to provide the best predictions. In this study, the utilization of five distinctive enhancer prediction tools in predicting enhancers from the proboscis monkey genome has proven to be successful in identifying moderately strong enhancers that have high functional specificities in the liver organ (enhancer pme201-pme204). This approach is deemed more feasible as compared to other methods because it considers multiple enhancer features such as histone modifications, TFBS, EP300 binding and sequence conservation. The bedtools (Quinlan & Hall, 2010) is one of the powerful tools that can consolidate and combine all outcomes from different tools and provide consensus enhancer sequences relative to the

parameters tested. The bedtools had been proven essential in terms of its efficiency and performance to accurately combine outcomes of various enhancer predictor tools. In this approach, the enhancers predicted do not have high activities as compared to the other two approaches, but their level of activities are not significantly different among each other. This phenomenon can be explained by the consolidation of various enhancer predictor tools utilizing various enhancer features that indirectly lead to opting out of some strong enhancers from the merging of outcomes from different tools (Omar *et al.*, 2017). It is important to note that this approach was not aimed to search mainly for strong enhancers, but for enhancers that have the most enhancer features (Omar *et al.*, 2017). Thus, this approach is feasible for the identification of highly versatile enhancers such as developmental enhancers that can function for long term and have housekeeping or multitasking functions.

5.2 Enhancer activity

The enhancer activity in this study was determined via the reporter assay approach where all 13 enhancers were subjected to recombinant plasmid insertion followed by transient transfection into HepG2 cell line. The resultant activity recorded was normalized against negative control and statistical analyses were conducted. The enhancer pme001 was found to have the highest activity among all and it has much higher expression compared to positive control (plasmid containing SV40 viral enhancer). The next high-activity enhancers come from the first and second approach (two from each). Surprisingly, the lowest activity enhancers also come from the first two approach (one from the first approach and two from second approach).

Looking onto the approach basis, the first approach provides an overview of mostly very strong enhancers with one exception (pme003). The possible reason to explain this

phenomenon is that iEnhancer-2L has indeed the capability to accurately identify very strong enhancers via their two-layer approach where the first layer distinguishes enhancers from non-enhancers whereas the second layer distinguishes strong enhancer from weak enhancers (Liu *et al.*, 2015). On the other hand, the second approach has only 40% of very strong enhancers compared to that 75% of the first approach. The second approach provides an interesting variety of very strong and moderately strong enhancers, making this approach a balanced landscape of enhancer (even though the population size in this study is small). The third approach provides a population of enhancers with almost equal level of expression across the four members. Due to small population size in this study, the above assumptions are yet to be strengthened with larger population size testing in future.

Interestingly, some strong correlations can be deduced from relating aspects such as motif distribution, TFBS patterns and TFBS abundance to enhancer activity (Kerppola & Curran, 1993; Slattery *et al.*, 2014; Grossman *et al.*, 2017).

5.3 Motif distribution and enhancer activity

The motif distribution of each candidate enhancer was constructed using MEME tool (Bailey & Elkan, 1994) with the parameter of 15 motifs with any repetitions. The motif distributions of some of the enhancers identified are found to have correlations to the resultant expression level of the enhancers. For instance, enhancer pme101, pme001 and pme204 all have two Motif 8 but the orientations of the two motifs within the sequence differs: for pme001, there is one motif on each forward and reverse strand; for pme101, both motifs are on forward strand; for pme204, both the motifs are on the reverse strand. Only the enhancer pme001 has Motif 8 on alternative strands and this may have contributed to its highest activity among all. The function of Motif 8 is unknown and yet to be identified.

Interestingly, the enhancer pme103 has much higher expression level than enhancers pme002 and pme004 but it has fewer motifs compared to enhancer pme002 and pme004, moreover enhancer pme103 has extra Motif 4. According to the outcomes of gene ontology conducted using GOMo (Buske *et al.*, 2010), Motif 4 was found to be involved in biological processes in human based on the top five specific predictions done. The biological process that Motif 4 predicted to play part in are response to amphetamine, response to cocaine, positive regulation to MAPKKK cascade, cell fate specification as well as grooming behaviour (Takekawa *et al.*, 2005).

Besides, some motifs were found to be signature within the enhancer clusters. For instance, Motif 5 is only found in enhancer pme001-pme004 cluster. However, there is no gene ontology records found for this motif from GOMo (Buske *et al.*, 2010) in human. Similarly, Motif 10 can be found in enhancer pme101-105 and in pme003 and pme203, this motif may not be cluster-specific, but it may be involved in some functional roles in the cluster. Based on GOMo (Buske *et al.*, 2010), Motif 10 is involved in two gene ontology categories: molecular function and biological process in human. Specifically, Motif 10 is involved actively in olfactory receptor activity (molecular function) as well as sensory perception of smell (biological process).

5.4 Transcription factor binding site (TFBS) and enhancer activity

The correlation between transcription factor binding site (TFBS) and enhancer activity is one aspect not to be missed when investigating the contributing factors towards strong enhancer activities. Among the main TFBS aspects that are believed to be influential towards enhancer activities are the overall TFBS abundance, TFBS composition and TFBS distribution patterns (Kerppola & Curran, 1993; Slattery *et al.*, 2014; Chen *et al.*, 2015).

5.4.1 Overall TFBS abundance and enhancer activity

The overall TFBS abundance (TFBS frequency per nucleotide) of all 13 enhancers is not strongly reflected on their expression level measured experimentally in HepG2 cell line. For instance, enhancer pme101 was predicted to be one of the enhancers having the highest TFBS frequency per nucleotide ratio among all, however it turned out to have one of the lowest expression levels compared to its counterparts. Besides, enhancer pme002 and pme004 was predicted to have the lowest TFBS frequency per nucleotide ratio but their experimentally determined expression level turned out to be among the top 5 strong enhancers among the 13 enhancers identified in this study. According to Chen *et al.* (2015), TFBS-clustered regions contribute more to enhancer activities and cellular functions than individually scattered TFBS within the enhancer.

On the cluster basis, the first cluster depicts negative correlation between predicted TFBS frequency per nucleotide ratio and enhancer activity (except for enhancer pme001). The highest TFBS frequency per bp ratio in enhancer pme003 was found to be lowest in activity after determined experimentally. The low predicted TFBS frequency per bp ratio in enhancer pme002 and pme004 had resulted in their high enhancer activities in HepG2 cell line. Only enhancer pme001 has moderate TFBS frequency per nucleotide ratio and experimentally it turned out to express highly in cell line.

The second cluster of enhancers (pme101-pme105) also showed the same negative correlation pattern as the first cluster. The enhancer pme103 with the lowest TFBS frequency per nucleotide ratio was determined to have the highest activity within the cluster. The moderately high TFBS abundance predicted for the other members of this cluster had given low activity level in general.

The third cluster of enhancers (pme201-pme204) depicted a positive correlation except for enhancer pme203. The positive relationship between TFBS abundance and enhancer activity can be seen in enhancer pme201, pme202 and pme204 where both their TFBS abundance and activities are considered moderately high. The enhancer pme203 however followed the negative correlation as seen in the other two clusters.

The negative correlation shown in most of the enhancers may indicate that compaction of TFBS within the enhancer sequences may not contribute to the elevation of enhancer activities. To a certain extent, although the TFBS of an enhancer is abundant, but there is possible that the binding site of TFs in some regions are overlapped, thus binding of one TF onto a region that serves as binding sites for various TFs may impede binding of other TFs capable of binding to that region. Therefore, TFBS abundance itself is not sufficient to be utilized to deduce the contributing factor to enhancer activity in this case (Hu *et al.*, 2007; Gotea *et al.*, 2010; Lusk & Eisen, 2010). In the case of C/EBP- β , the occurrence and frequency C/EBP- β binding sites are generally high and random throughout the genome and for the C/EBP- β to be triggered to activate the opening of chromatin, several closely spaced C/EBP- β binding sites are required (Pachetka *et al.*, 2008). Furthermore, the ratio may not serve as sensitive indicator because there are various TFBS components within the enhancer and each can function individually and mutually to bring about functional role of enhancers (Pachetka *et al.*, 2008; Taher *et al.*, 2011). According to Chen *et al.* (2015), TFBS-clustered regions contribute more to enhancer activities and cellular functions than individually scattered TFBS within the enhancer. Therefore, it is important to look into the arrangement and position of TFBS clusters rather than focusing onto individual TFBS abundance alone.

5.4.2 TFBS composition and enhancer activity

The TFBS composition of enhancers highlighted in this study are AP-1, HNF-1, GATA-3, C/EBP- β , NF-1, HNF-3 β , COUP-TF/HNF-4, USF, CHOP-C/EBP- α , YY1, HNF-4 α 1, TATA, CREB and NF-kappaB. These TFBS are known to regulate liver-specific genes, especially those function as transporter in the liver organ (Kim *et al.*, 2011). In general, the TFBS composition of enhancer studied have a greater correlation towards enhancer activity as compared to overall TFBS abundance.

Based on Figure 4.9 and Table 4.9, the enhancer pme001 has the highest amount of AP-1 TFBS within its sequence among all 13 enhancers and considering that it has the highest enhancer activity as well, the AP-1 may have been the major contributing factor towards this phenomenon. In a study that involves systematic dissection of genomic features determining transcription factor binding and enhancer function, Grossman *et al.* (2017) deduced that mutating the AP-1 and ATF (activating transcription protein) within enhancer sequences had the most significant effect among all and 87% of the decreased enhancer expressions are at least two-fold. The C/EBP- β TFBS was found to be the most abundant in enhancer pme103 and this TFBS is also considered as one of the main drivers in enhancer activity of enhancer pme103 as this enhancer came out second in terms of expression level. The C/EBP- β is well known for the vital roles it plays in the liver, namely metabolism, adipose tissue differentiation and inflammation (van der Krieken *et al.*, 2015). Notably, the enhancer pme101 is ranked the top in terms of HNF-3 β TFBS with a total of 29 binding sites, however the activity of enhancer pme101 was discovered to be the second lowest in activity, suggesting the role of this TF in repressing enhancer activity. However, the HNF-3 β is well known for its role as a strong contributor towards human prothrombin gene enhancer and liver-specific enhancers (Ceelie *et al.*, 2003; Kim *et al.*, 2011).

The first cluster containing enhancer pme001-pme004 are enriched in AP-1 and C/EBP- β TFBS. The presence of these strong transcription factors within the members of this cluster is one of reason for the high number of high-activity enhancers arising from this cluster. The second cluster of enhancers with enhancer members pme101-pme105 are enriched with HNF-1 and C/EBP- β TFBS. The third cluster of enhancers (pme201-pme204) are like the second cluster which are also enriched in HNF-1 and C/EBP- β TFBS. The lack of AP-1 TFBS as main driver in the cluster may have suggested on why the second and third cluster has rather fewer strong enhancers compared to the first cluster (Grossman *et al.*, 2017).

5.4.3 TFBS distribution pattern and enhancer activity

The TFBS distribution pattern has substantial effects on enhancer activity in terms of the occurrence of pairs or sets of TFBS combinations that elevates the enhancer activity more than the effects of individual TFBS. For instance, the occurrence of TFBS pairs of the same type or even the cluster of several TFBS at a single region may exert positional influences onto the resultant enhancer activities. There are several combinations di-, tri- and multiple TFBS sets that are signature towards an enhancer or groups of enhancers in the cluster.

To relate some special TFBS pattern characteristics to enhancer activity outcomes, the first one to note is the enhancer pme001 which had the highest expression among all, it is the one and only enhancer with not only the highest AP-1 TFBS but also with four AP-1 adjacent pairs at four different 100 bp intervals. As transcription factor AP-1 is known for its role in improving enhancer functioning in liver cells, these AP-1 pairs of TFBS are probably the reason for the activity of enhancer pme001 to rocket high above the others

(Kerppola & Curran, 1993). The transcription factor AP-1 was known to have various vital functions across many cells and tissues. To focus on the roles of AP-1 in the liver, AP-1 works with another growth-induced transcription factor (STAT3) as well as tissue specific transcription factor (HNF-1) in an interaction that leads to an adaptive stimulus to injury of liver organ (Leu *et al.*, 2001). They are also the essential transcription factors that play major role in the amplification of hepatic genes expressions in which these increased gene expressions contribute to liver organ regeneration via homeostatic stimuli (Leu *et al.*, 2001). Although Nguyen *et al.* (2016) had proven that AP-1 is biased towards activation of enhancer activity more than promoter via high-throughput functional comparison, there is no evidence yet for the effects of AP-1 adjacent pairs in the improvement of enhancer activity. From this study, it is found that enhancer pme001 (with 10 AP-1 TFBS) outperformed enhancer pme105 (with 9 AP-1 TFBS) by two-fold in terms of expression level, thus it is clear that transcription factor AP-1 TFBS itself does not guarantee high enhancer activity unless it is found in pairs located adjacently to one another. It was proven that pair of AP-1 is capable of DNA bending to further increase transcription level (Kerppola & Curran, 1993).

The second highest expressing enhancer, the pme103, is also packed with unique TFBS feature combinations that are hard to get unnoticed. Besides having the highest C/EBP- β TFBS, the enhancer pme103 also has two sets of five adjacent C/EBP- β TFBS as well as two triplets of HNF-1TFBS, making it equipped with all the potential to express strongly in HepG2 cell line in this study. The C/EBP- β is essential for multiple myeloma as it orchestrate all key TFs to concertedly work to maintain survivability, antiapoptotic responses as well as proliferation (Pal *et al.*, 2009). In the same study, they had proven that C/EBP- β deficiency in mice directly leads to impairment of B lymphocytes (Pal *et al.*, 2009). There are many lines of evidence that the transcription factor C/EBP- β plays profound roles

in the liver organ, namely inflammation, high density lipoprotein synthesis, hepatic steatosis, lipid accumulation, triacylglycerol breakdown, insulin and glucose metabolism, endoplasmic reticulum stress as well as adipose tissue differentiation (van der Krieken *et al.*, 2015). The transcription factor HNF-1 is another strong TF not to be ignored as it regulates various genes such as the albumin, aldolase B, CRP Vitellogenin A2, CYP2E1, IGFBP-1, HBV, Apolipoprotein B and Apolipoprotein AII gene in the liver (Maire *et al.*, 1989). The enhancer pme103 is undeniably a strong enhancer for having the composition of strong TFBS distribution pattern like the C/EBP- β and HNF-1 sets.

Moreover, enhancer pme101 has two huge clusters of HNF-3 β TFBS and this may have impact on its activity in cell and tissue, especially it is one of the enhancers with low expression levels. The transcription factor HNF-3 β works similarly with HNF-3 α to elevate liver-specific gene expression leading to regulation of endoderm-derived, intestinal, pancreatic and bronchiolar epithelium hepatocytes in the liver (Rausa *et al.*, 1997; Clevidence *et al.*, 1998). Also, HNF-3 β , HNF-3 α and HNF-3 γ interacts with some other HNF-3 binding elements in the promoters of liver-specific genes such as tyrosine amino transferase, phosphoenolpyruvate carboxykinase, albumin and aldolase B gene to bring about transcriptional elevations in the liver (Schrem *et al.*, 2002). Besides, Ceelie *et al.* (2003) had also conducted mutagenesis studies to functionally characterize TFBS in the human prothrombin gene enhancer and they found out that the binding of HNF-3 β is pivotal to enhancer activity as loss of binding of this TF leads to decrease in enhancer activity. Despite the numerous evidence that supports HNF-3 β as an activator for enhancers (Rausa *et al.*, 1997; Clevidence *et al.*, 1998; Schrem *et al.*, 2002; Kim *et al.*, 2011), the effects of deletion of a large HNF-3 β TFBS region onto enhancer activity is still unknown. Besides, the binding behaviour of HNF-3 β onto this region remains a mystery, whether only one

HNF-3 β can bind onto the HNF-3 β -rich TFBS site at one time or multiple HNF-3 β can bind cooperatively and share the same TFBS site. Therefore, it is possible that the HNF-3 β TFBS-enriched regions found within the enhancer pme101 are accountable for the low resultant expression level in HepG2 cell line. In other words, it is either the multiple HNF-3 β binding or only a single HNF-3 β binding onto the HNF-3 β -enriched regions that had led to the low enhancer activity in enhancer pme101.

It is also interesting to look at the signature TFBS patterns within the three main clusters of enhancers. At a glance, the most abundance TFBS patterns found within the pme001-pme004 cluster is the pairs of C/EBP- β and NF-1 TFBS which may serve as driver in enhancer activities for this cluster. The driver TFBS found in the cluster pme101-pme105 are the C/EBP- β and HNF-1 TFBS whereas the cluster pme201-pme204 has shown no obvious uniform TFBS patterns among them.

5.5 Deletion analysis

The deletion analysis has been conducted on selected enhancers to determine the major contributors towards their enhancer activities. The enhancers selected for deletion study are enhancer pme001, pme101 and pme103. Based on the results obtained, some contributors towards the enhancer activity have been identified with most of them possessing enhancing effects towards enhancer while the others have repressing effects onto enhancers.

In the case of the deletion analysis of enhancer pme001, four enhancer fragments were amplified from the parent enhancer sequences to further study the effects of absence of vital TFBS such as AP-1 and USF. It has been noticed that enhancer pme001 is unique in terms of its possession of many pairs of adjacently-located AP-1 TFBS among the 13 enhancers studied. Overall, the adjacent AP-1 TFBS pair was determined to be one of major

contributors towards enhancer activity in enhancer pme001 which is consistent with the results generated by Grossman *et al.* (2017) on the effects of different combination pairs of AP-1 onto enhancer activities. The activity of the enhancer fragment pme001a and pme001D are the same and they both scored the lowest value among all the other fragments. Looking into the deletion structure of these fragments, the enhancer fragment pme001a only contains the first two adjacent AP-1 TFBS pair sets whereas the enhancer fragment pme001D has all three adjacent AP-1 TFBS pair sets and a tailing AP-1 TFBS. On the other hand, enhancer fragment pme001b has all three adjacent AP-1 TFBS pair sets (similar to pme001D but without the tailing AP-1 TFBS) and had scored a much higher expression level in HepG2 cell line compared to pme001D. The deletion done onto enhancer fragment pme001D has not introduced any new TFBS sites in that region. Thus, it is possible that the binding affinity of AP-1 towards enhancer pme001 was affected with the removal of that region, suggesting the role of this region in orchestrating binding affinities of transcription factors to activate enhancer activity (Slattery *et al.*, 2014).

For the deletion analysis done onto enhancer pme101, the deletion of the HNF-3 β -enriched regions had resulted in the increment of enhancer activity up to more than two-fold. The deletion of the first HNF-3 β -enriched regions containing 12 HNF-3 β TFBS in enhancer fragment pme101D1 had yielded enhancer activity around 1.7-fold more than the parent enhancer pme101 whereas the deletion of the second HNF-3 β -enriched region containing 8 HNF-3 β TFBS in enhancer fragment pme101D2 had yielded a 1.4-fold times higher in activity from the parent enhancer pme101. In the case of enhancer fragment pme101D3, the deletion of the both abovementioned HNF-3 β -enriched regions had produced enhancer activity of more than two-fold higher than that of the parent enhancer pme101. Although there are many lines of evidence that had supported HNF-3 β as a strong enhancement

towards liver activities (Rausa *et al.*, 1997; Clevidence *et al.*, 1998; Schrem *et al.*, 2002; Kim *et al.*, 2011), interestingly, there is one study conducted by Li *et al.* (2002) that concluded that HNF-3 can indeed work in tandem with NKX2.1 (transcription factor that activates promoter of thyroid specific genes such as thyrotropin receptor, thyroperoxidase and thyroglobulin) (Provenzano *et al.*, 2016) to introduce repression effects onto the transcription of *Sp-B* (surfactant protein B) gene. Therefore, it is safe to postulate that the clusters of HNF-3 β TFBS present in enhancer pme101 may have contributed towards repression of enhancer activity for enhancer pme101 in HepG2 cell line in this study.

The deletion analysis done on enhancer pme103 has revealed the major players in enhancer activity, namely the C/EBP- β and HNF-1 transcription factors. The deletion of the first C/EBP- β and HNF-1-enriched region (three C/EBP- β and three HNF-1 TFBS in 1-100 bp region) had resulted in reduction of enhancer activity up to 2.7-fold from the parent enhancer in the case of enhancer fragment pme103a. A much higher decline (3.8-fold) in enhancer activity was observed in enhancer fragment pme103D1 with the second C/EBP- β and HNF-1-enriched region (five C/EBP- β and three HNF-1 TFBS in 500-599 bp region) being deleted. The most significant decrease (6.4-fold) in enhancer activity from the parent enhancer was seen in enhancer fragment pme103D2 with the adjacently located five C/EBP- β -enriched region (800-899 bp region) being deleted. This result from this study is consistent with that conducted by Plachetka *et al.* (2008) where they deduced that the opening of chromatin at *mim-1* enhancer can be activated by C/EBP- β when several binding sites are present within a close distance. Thus, it can be deduced that adjacently located five C/EBP- β -enriched region has the most substantial contributions towards enhancer activity of enhancer pme103 in HepG2 cell line from this study.

5.6 Comparison of enhancer activity across two cell lines

The enhancer activities were tested on two cell lines in this study, namely HepG2 and A549. In A549 cell line, all the 13 liver-specific enhancers tested in this study had shown lower enhancer activity compared to that in HepG2 cell line. As for the enhancer fragments, some of them had shown higher expression level in A549 than in HepG2 cell line.

In the case of enhancer pme001 and its enhancer fragments, most enhancer fragments have lower or equal expression than the parent enhancer except for enhancer fragment pme001D in A549 cell line. The difference between the parent enhancer and enhancer fragment pme001D is the absence of AP-1 and USF-enriched region (three AP-1 and three USF TFBS), thus suggesting the role of this region in maintaining liver-specific status of the enhancer pme001.

The expressions of all enhancer fragments of enhancer pme101 were at the same level when they are being expressed in A549 cell line. The absence of either one of the HNF-3 β -enriched regions or both regions had caused the expression level of these enhancer fragments to elevate higher than that of the parent enhancer pme101 in A549 cell line. Similarly, these HNF-3 β -enriched regions may have vital roles in repressing the activity of enhancer pme101 even in cell lines other than liver, but their repressive actions are not as strong and as quantitatively-dependent as in liver cell line (HepG2).

The enhancer fragments of enhancer pme103 had depicted a completely distinctive phenomenon in A549 cell line than that in HepG2 cell line. In HepG2 cell line, the enhancer fragment activity trend was seen to be decreasing from pme103, pme103a, pme1013D1 to pme103D2; however, the increasing trend was seen in the similar order in A549 cell line. This phenomenon may suggest that all three deleted TFBS-enriched regions (in the case of

pme103a, pme103D1 and pme103D2) are important regulators to make sure enhancer pme103 only functions optimally and specifically in the liver tissues.

These results are consistent to that discovered by Duan and Simpson-Haidaris (2006) on IL-6-receptor-gp100-Stat3 signalling role in Interleukin-6 activation of γ FBG promoter. In this study, it was found that this signalling is highly liver tissue-specific since the promoter activity was shown to be higher in A549 cell line as compared to that in HepG2 cell line (Duan & Simpson-Haidaris, 2006).

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusions

The objectives of this study had been achieved with the completion of the identification, isolation and activity determination of enhancers in the chromosome 18 of the proboscis monkey. The identification of enhancers from the proboscis monkey had yielded 9,960,241 strong candidate enhancers. However, this number is only limited to chromosome 18 only. The isolation of a total of 13 enhancers in this study has been successful via the various experimental approaches applied throughout the study.

The activity determination of these selected enhancers had also yielded satisfactory results with contribution towards a better understanding of how enhancer activities are closely related to their TFBS and motifs within their sequences. For instances, the presence of adjacently located TFBS pairs found uniquely in some of the enhancers were suspected to hold key to enhancer activities in HepG2 cell line and these TFBS-enriched regions are subjected to a more detailed investigations via the deletion analysis approaches.

Deletion studies had also been conducted on selected enhancers to finally deduce the functional contributors towards enhancer activities. The TFBS-enriched regions such as the AP-1 adjacent pairs, AP-1 and USF-enriched regions as well as the five C/EBP- β -enriched regions are found to have enhancing effects onto enhancer activities in HepG2 cell line whereas HNF-3 β -enriched regions are found to have suppressing effects towards enhancer activities in HepG2 cell line. Moreover, by comparing the expression levels of these enhancer fragments across two different cell lines, more information has been yielded on whether these deleted TFBS-enriched regions function in a tissue-specific manner. From this

study, some TFBS-enriched regions were found to be involved in liver-specific expression either in a quantitatively-dependent manner (enhancer fragments of pme103), quantitatively-independent manner (enhancer fragments of pme101) or even based on individual signature TFBS-enriched region (enhancer fragment pme001D).

6.2 Recommendations

One of the recommendations of this study is to further improve the enhancer identification methods with the involvement of more enhancer predictor tools with high accuracies. This helps in further consolidate more tools with different prediction parameters covering more important enhancer features and markers. Besides, more enhancers identified in this study should also be tested on their activities at least *in vitro*. This helps to identify more vital TFBS-enriched regions that are essential for enhancer functioning.

Next, the enhancer activity can also be tested on other different cell lines to determine and analyse on their tissue specificities. This may provide a more comprehensive view of enhancer-regulated expression in different tissues and further characterize these enhancers. On the other hand, more deletion analysis can be conducted onto other enhancers from different chromosomes or from those identified in this study to provide a more comprehensive enhancer landscape of the proboscis monkey. Further *in vivo* studies such as RNAi analysis using mice as model are also essential for determination of enhancer activities in the living organism context. High-throughput assays are also recommended for a more rapid enhancer identification process, especially on a larger scale (on whole chromosome basis).

All in all, this study had conducted preliminary isolation of enhancers from the proboscis monkey genome and further tested on the expression levels of the 13 selected

enhancers and their fragments in both HepG2 and A549 cell lines. It is hoped that the results yielded from this study can be useful in bridging the knowledge gap in enhancer landscapes across primates in the near future.

REFERENCES

- Abdullah, M. T., Mat Daud, M. H. R., Nur Aida, M. T., Idris, A., Croft, L., Saidin, A., ... Hercus, R. (2014). *Draft genome sequence of Nasalis larvatus enables comparative analysis of 10 simian genomes*. Unpublished manuscript, Department of Zoology, Universiti Malaysia Sarawak, Sarawak, Malaysia.
- Alipanahi, B., Delong, A., Weirauch, M., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831-839.
- Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123), 1074-1077.
- Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36.
- Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1), 299-308.
- Bauer, S. R., Kudo, A., & Melchers, F. (1988). Structure and pre-B lymphocyte restricted expression of the *VpreB* gene in humans and conservation of its structure in other mammalian species. *The EMBO Journal*, 7(1), 111-116.
- Baylor College of Medicine-Human Genome Sequence Center. (2016). *Non-human primates*. Retrieved August 8, 2017 from <https://www.hgsc.bcm.edu/non-human-primates>
- BioSpectrum Bureau. (November 12, 2010). *Malaysia commences genome sequencing and analysis project*. Retrieved August 12, 2017 from <https://www.mgrc.com.my/file/media/BioSpectrum121101.pdf>
- Blomquist, P., Belikov, S., & Wrangé, Ö. (1999). Increased nuclear factor 1 binding to its nucleosomal site mediated by sequence-dependent DNA structure. *Nucleic Acids Research*, 27(2), 517-525.

- Boonratana, R. (2002). Social organization of proboscis monkeys (*Nasalis larvatus*) in the lower Kinabatangan, Sabah, Malaysia. *The Malayan Nature Journal*, 56(1), 57-75.
- Bulger, M., & Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes & Development*, 13, 2465-2477.
- Bulger, M., & Groudine, M. (2010). Enhancers: The abundance and function of regulatory sequences beyond promoters. *Developmental Biology*, 339(2), 250-257.
- Buske, F. A., Bodén, M., Bauer, D. C., & Bailey, T. L. (2010). Assigning roles to DNA regulatory motifs using comparative genomics, *Bioinformatics*, 26(7), 860-866.
- Cain, C. E., Blekman, R., Marioni, J. C., & Gilad, Y. (2011). Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*, 187, 1225-1234.
- Cao, Q., & Yip, K.Y. (2016). A survey on computational methods for enhancer and enhancer target predictions. In K. Wong (Ed.), *Computational biology and bioinformatics: Gene regulation* (pp. 3-27). New York, NY: CRC Press.
- Carlberg, C., & Molnár, F. (2014). *Mechanisms of gene regulation*. Dordrecht, Netherlands: Springer Netherlands.
- Ceelie, H., Spaargaren-van Riel, C. C., De Jong, M., Bertina, R. M., & Vos, H. L. (2003). Functional characterization of transcription factor binding sites for HNF1-alpha, HNF3-beta (FOXA2), HNF4-alpha, Sp1 and Sp3 in the human prothrombin gene enhancer, *Journal of Thrombosis and Haemostasis*, 1, 1688-1698.
- Chen, H., Li, H., Liu, F., Zheng, X., Wang, S., Bo, X., & Shu, W. (2015). An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Scientific Reports*, 5(8465), 1-13.
- Chiarelli, B. (1966). The chromosome complement of *Nasalis larvatus* (Wurm 1781), *Experientia*, 22(12), 797-799.

- Chikhirzhina, G. I., Al'-Shekhatat, R. I., & Chikhirzhina, E. V. (2008). Transcription factors of the nuclear factor 1 (NF1) family: Role in chromatin remodeling. *Molecular Biology (Mosk)*, 42(3), 388-404.
- Chou, J., Provot, S., & Werb, Z. (2010). GATA3 in development and cancer differentiation: Cells GATA have it! *Journal of Cellular Physiology*, 222(1), 42-49.
- Clevidence, D. E., Zhou, H., Lau, L. F., & Costa, R. H. (1998). The -4 kilobase promoter region of the winged helix transcription factor HNF-3 α gene elicits transgene expression in mouse embryonic hepatic and intestinal diverticula. *The International Journal of Developmental Biology*, 42, 741-746.
- Corradin, O., & Scacheri, P. C. (2014). Enhancer variants: Evaluating functions in common disease. *Genome Medicine*, 6(10), 85-85.
- Cotney, J., Leng, J., Yin, J., Reilly, S. K., DeMare, L. E., Emera, D., ... Noonan, J. P. (2013). The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, 154, 185-196.
- 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. *PNAS*, 107(50), 21931-21936.
- De Vooght, K. M., van Wijk, R., & van Solinge W. W. (2009). Management of gene promoter mutations in molecular diagnostics. *Clinical Chemistry*, 55, 698-708.
- Duan, H. Q., & Simpson-Haidaris, J. (2006). Cell-type differential induction of the human γ -fibrinogen promoter by interleukin-6. *The Journal of Biological Chemistry*, 281(18), 12451-12457.
- Ellis, D. (1986). Proboscis monkey and aquatic ape. *The Sarawak Museum Journal*, 36(57), 251-262.
- Epstein, D. J. (2009). Cis-regulatory mutations in human disease. *Briefings in Functional Genomics and Proteomics*, 8(4), 310-316.

- Firpi, H., Ucar, D., & Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, 26(13), 1579-1586.
- Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Computational Biology*, 10(7), e1003711.
- Giusto, J., & Margulis, L. (1981). Karyotypic fission theory and the evolution of old world monkeys and apes. *Biosystems*, 13(4), 267-302.
- González-Rubio, S., Linares, C. I., Aguilar-Melero, P., Rodríguez-Perálvarez, M., Montero-Álvarez, J. L., de la Mata, M., & Ferrin, G. (2016). AP-1 inhibition by SR 11302 protects human hepatoma HepG2 cells from bile acid-induced cytotoxicity by restoring the NOS-3 expression. *PLoS ONE*, 11(8), e0160525.
- Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., ... Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, 352(6281), doi:10.1126/science.aae0344.
- Gotea, V., Visel, A., Westland, J. M., Nobrega, M. A., Pennacchio, L. A., & Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20, 565-577.
- Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., ... Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *PNAS*, doi:10.1073/pnas.1621150114, E1291-E1300.
- Groves, C. P. (1989). *A theory of human and primate evolution*. Oxford, UK: Clarendon Press.
- He, B., Chen, C., Teng, L., & Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21), E2191-E2199.
- Herendeen, D. R., Kassavetis, G. A., & Geiduschek, E. P. (1992). A transcriptional enhancer whose function imposes a requirement that proteins track along DNA. *Science*, 256(5061), 1298-1303.
- Hu, Z., Hu, B., & Collins, J. F. (2007). Prediction of synergistic transcription factors by function conservation. *Genome Biology*, 8, R257.

- Huang, Y. S., Ramensky, V., Service, S. K., Jasinska, A. J., Jung, Y., Choi, O. W., ... Freimer, N. B. (2015). Sequencing strategies and characterization of 721 vervet monkey genomes for future genetic analyses of medically relevant traits. *BMC Biology*, 13(41), 1-10.
- Irwin, D. M., Kocher, T. D., & Wilson, A. C. (1991). Evolution of the cytochrome-b gene in mammals. *Journal of Molecular Evolution*, 32, 128-144.
- Jakubowski, H. (2016). *Biochemistry*. Retrieved June 15, 2017, from <https://employees.csbsju.edu/hjakubowski/classes/ch331/bind/olbindtranscription.html>
- Jochum, W., Passequé, E., & Wagner, E. F. (2001). AP-1 in mouse development and tumorigenesis. *Oncogene*, 20, 2401-2412.
- Kel, A. E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Morgoulis, O. V., & Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31, 3576-3579.
- Kerppola, T. K., & Curran, T. (1993). Selective DNA bending by a variety of bZIP proteins. *Mol Cell Biol*, 13(9), 5479-5489.
- Kibbe, W. A. (2007). OligoCalc: An online oligonucleotide properties calculator. *Nucleic Acids Research*, 35, W43-46.
- Kim, M. J., Skewes-Cox, P., Fukushima, H., Hesselson, S., Yee, S. W., Ramsey, L. B., ... Ahituv, N. (2011). Functional characterization of liver enhancers that regulate drug-associated transporters. *Clinical Pharmacology & Therapeutics*, 89(4), 571-578.
- King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(2184), 107-188.
- Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2015). Progress and challenges in bioinformatics approaches for enhancer identification. *Briefings in Bioinformatics*, doi:10.1093/bib/bbv101, 1-13.
- Kolesky, S. E., Mohamed Ouhammouch, & Geiduschek, E. P. (2002). The mechanism of transcriptional activation by the topologically DNA-linked sliding clamp of bacteriophage T4. *Journal of Molecular Biology*, 321, 767-784.

- Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., & Ppantonis, A. (2012). Enhancers and silencers: An integrated and simple model for their function. *Epigenetics & Chromatin*, 5(1), 1-8.
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, 33(7), 1870-1874.
- Kvon, E. Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics*, doi: 10.1016/j.ygeno.2015.06.007.
- Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24, 1595-1602.
- Lancôt, C., Cheutin, T., Cremer, M., Cavalli, G., & Cremer, T. (2007). Dynamic genome architecture in the nuclear space: Regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8, 104-115.
- Leu, J. I., Crissey, M. A., Leu, J. P., & Ciliberto, G. (2001). Interleukin-6-induced STAT3 and AP-1 amplify hepatocyte nuclear factor 1-mediated transactivation of hepatic genes, an adaptive response to liver injury. *Molecular and Cellular Biology*, 21(2), 414-424.
- Lelli, K. M., Slattery, M., & Mann, R. S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annual. Review of Genetics*, 46, 43-68.
- Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424, 147-151.
- Li, C., Zhu, N. L., Tan, R. C., Ballard, P. L., Derynck, R., & Monoo, P. (2002). Transforming growth factor- β inhibits pulmonary surfactant protein B gene transcription through SMAD3 interactions with NKX2.1 and HNF-3 transcription factors. *The Journal of Biological Chemistry*, 277(41), 38399-38408.
- Lim, L. W. K., Chung, H. H., Chong, Y. L., & Lee, N. K. (2018). A survey of recently emerged genome-wide computational enhancer predictor tools. *Computational Biology and Chemistry*, 74, 132-141.

- Liu, B., Fang, L., Long, R., Lan, X., & Chou, K. (2016). iEnhancer-2L: A two layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, 32(3), 362-369.
- Lusk, R. W., & Eisen, M. B. (2010). Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. *PLoS Genetics*, 6, e1000829.
- Madrid-Espinoza, J., Brunel-Saldias, N., Guerra, F. P., Gutiérrez, A., & Del Pozo, A. (2018). Genome-wide identification and transcriptional regulation of aquaporin genes in bread wheat (*Triticum aestivum* L.) under water stress. *Genes*, 9(10), 497-499.
- Maire, P., Wuarin, J., & Schibler, U. (1989). The role of cis-acting promoter elements in tissue-specific albumin gene expression. *Science (Wash DC)*, 244, 343-346.
- Mello, C. C., Kramer, J. M., Stinchcomb, D., & Ambros, V. (1991). Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO Journal*, 10, 3959-3970.
- Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Vrielink, J. A. F., ... Agami, R. (2014). eRNAs are required for p53-dependent enhancer activity and gene transcription. *Molecular Cell*, 49, 524-535.
- Mendel D. B., Hansen, L. P., Graves, M. K., Conley, P. B., & Crabtree, G. R. (1991). HNF-1 α and HNF-1 β (vHNF-1) share dimerization and homeo domains, but not activation domains, and form heterodimers *in vitro*. *Genes & Development*, 5, 1042-1056.
- Miguel-Escalada, I., Pasquali, L., & Ferrer, J. (2015). Transcriptional enhancers: Functional insights and role in human disease. *Current Opinion in Genetics and Development*, 33, 71-76.
- Napier, J. R. & Napier, P. H. (1985). *The natural history of the primates*. Cambridge, Massachusetts: MIT Press.
- National Institute of Health. (2017). *Genomics of gene regulation project overview*. Retrieved January 4, 2017, from <https://www.genome.gov/27561317/genomics-of-gene-regulation/>

- Natoli, G., & Andrau, J. C. (2012). Noncoding transcription at enhancers: General principles and functional models. *Annual Review of Genetics*, 46, 1-19.
- Nguyen, T. A., Jones, R. D., Snively, A. R., Pfenning, A. R., Kircher, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, 26, 1023-1033.
- Ovcharenko, I., Nobrega, M. A., Loots, G. G., & Stubbs, L. (2004). ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Research*, 32, W280-286.
- Omar, N., Wong, Y. S., Chong, Y. L., Abdullah, M. T., & Lee, N. K. (2017). Enhancer prediction in proboscis monkey genome: A comparative study. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(1), 175-179.
- Pal, R., Janz, M., Galson, D. L., Gries, M., Li, S., Jöhrens, K., ... Lentzsch, S. (2009). C/EBP β regulates transcription factors critical for proliferation and survival of multiple myeloma cells. *Blood*, 114(18), 3890-3898.
- Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27, 1173-1175.
- Payne, J., Francis, C. M., & Phillips, K. (1985). *A field guide to the mammals of Borneo*. Malaysia, Kuala Lumpur: World Wildlife Fund Malaysia & The Sabah Society.
- Peng, Y. Z., Pan, R. L., & Jablonski, N. G. (1993). Classification and evolution of Asian colobines. *Folia Primatologica*, 60, 106-117.
- Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., ... Rubin, E. M. (2006). *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature*, 444, 499-502.

- Petrascheck, M., Escher, D., Mahmoudi, T., Verrijzer, C. P., Schaffner, W., & Barberis, A. (2005). DNA looping induced by a transcriptional enhancer *in vivo*. *Nucleic Acids Research*, 33(12), 3743-3750.
- Phillips, K. A., Bales, K. L., Capitanio, J. P., Conley, A., Czoty, P. W., 't Hart, B. A., ... Voytko, M. L. (2014). Why primate models matter. *American Journal of Primatology*, 76(9), doi:10.1002/ajp.22281.
- Plachetka, A., Chayka, O., Wilczek, C., Melnik, S., Bonifer, C., & Klempnauer, K. H. (2008). C/EBP β induces chromatin opening at a cell-type-specific enhancer. *Molecular and Cellular Biology*, 28(6), 2102-2112.
- Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., ... Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, 163(1), 68-83.
- Provenzano, C., Zamboni, M., Veneziano, L., Mantuano, E., Garavaglia, B., Zorzi, G., ... Civitareale, D. (2016). Functional characterization of two novel mutations in TTF-1/NJX2.1 homeodomain in patients with benign hereditary chorea. *Journal of Neurological Sciences*, 360(1), 78-83.
- Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., ... Eichler, E. E. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486, 527-531.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6): 841-842.
- Rausa, F., Samadani, U., Ye, H., Fletcher, C. F., Jenkins, N. A., Copeland, N. G., & Costa R. H. (1997). The cut-homeodomain transcriptional activator HNF-6 is coexpressed with its target gene HNF-3 β in the developing murine liver and pancreas. *Developmental Biology*, 192, 228-246.
- Sebastian, A. C. (2000). Proboscis monkeys in Danau Sentarum National Park. *Borneo Research Bulletin*, 31, 359-371.

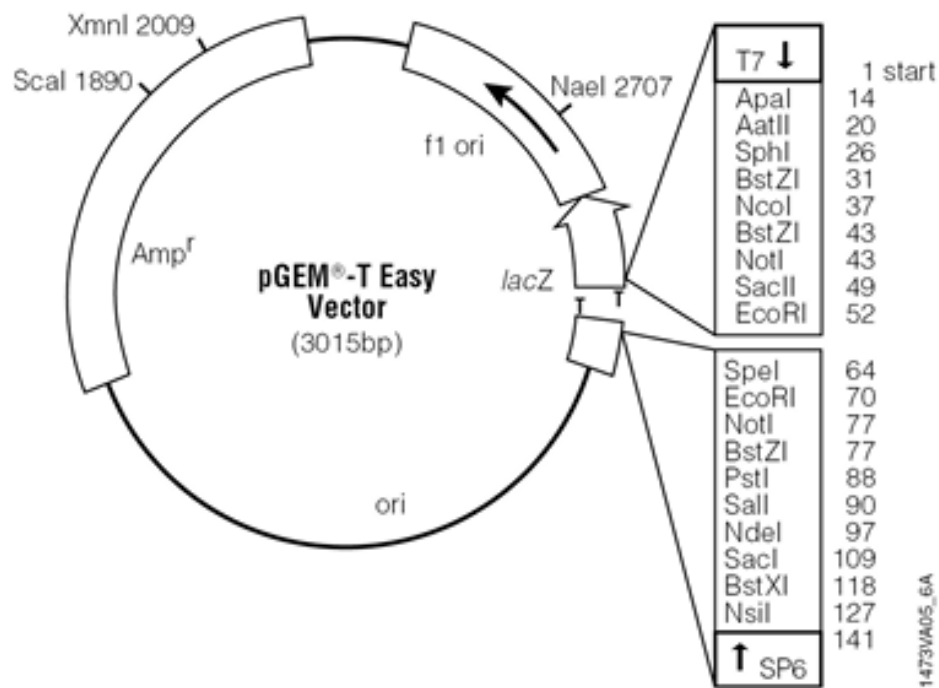
- Samadani, U., & Costa, R. H. (1996). The transcriptional activator hepatocyte nuclear factor 6 regulates liver gene expression. *Molecular and Cellular Biology*, 16, 6273-6284.
- Samson, S. I., Richard, O., Tavian, M., Ranson, T., Vosshenrich, C. A. J., Colucci, F., ... Di Santo, J. P. (2003). GATA-3 promotes maturation, IFN- γ production, and liver-specific homing of NK cells. *Immunity*, 19, 701-711.
- Schrem, H., Klempnauer, J., & Borlak, J. (2002). Liver-enriched transcription factor in liver function and development. Part I: The hepatocyte nuclear factor network and liver-specific gene expression. *Pharmacological Reviews*, 54, 129-158.
- Schwarzer, W., & Spitz, F. (2014). The architecture of gene expression: Integrating dispersed cis-regulatory modules into coherent regulatory domains. *Current Opinion in Genetics & Development*, 27, 74-82.
- Shaulin, E., & Karin, M. (2001). AP-1 in cell proliferation and survival. *Oncogene*, 20, 2390-2400.
- Shaulin, E., & Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nature Cell Biology*, 4, E131-E136.
- Shibata, Y., Sheffield, N. C., Fedrigo, O., Babbitt, C. C., Wortham, M., Tewari, A. K., ... Iyer, V. R. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genetics*, 8, e1002789.
- Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4), 272-286.
- Singh, R., Sharma, M. C., Sarkar, C., Singh, M., & Chauhan, S. S. (2014). Transcription factor C/EBP- β mediates downregulation of dipeptidyl-peptidase III expression by interleukin-6 in human glioblastoma cells. *FEBS Journal*, 281, 1629-1641.
- Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9), 381-399.
- Soma, H., Bernischke, K., & Robinson, K. P. (1974). The chromosomes of proboscis monkey (*Nasalis larvatus*). *Chromosome Information Service*, 17, 24-29.

- Stanyon, R., Camperio-Ciani, A., Sineo, L., & Morescalchi, M. A. (1992). The G-based chromosomes of the proboscis monkey (*Nasalis larvatus*) compared with the macaque (*Macaca mulatta*), *Journal of Contemporary Anthropology*, 15, 101-104.
- Stewart, C. B., Schilling, J. W., & Wilson, A. C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330(6146), 401-404.
- Strohmeier, R., Shelton, J., Loughheed, C., & Breitkopf, T. (2014). CCAAT-enhancer binding protein- β expression and elevation in Alzheimer's disease and microglial cell cultures. *PLoS*, 9(1), 1-11.
- Sun, C., Southard, C., Witonsky, D. B., Olopade, O. I., & Di Rienzo, A. (2010). Allelic imbalance (AI) identifies novel tissue-specific cis-regulatory variation for human UGT2B15. *Human Mutation*, 31(1), 99-107.
- Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S., & Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Research*, 21(1), 1139-1149.
- Takekawa, M., Tatebayashi, K., & Saito, H. (2005). Conserved docking site is essential for activation of mammalian MAP kinase kinases by specific MAP kinase kinase kinases. *Molecular Cell*, 18(1), 295-306.
- The Marmoset Genome Sequencing and Analysis Consortium. (2014). The common marmoset genome provides insight into primate biology and evolution. *Nature Genetics*, 46(8), 850-857.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3: New capabilities and interfaces. *Nucleic Acids Research*, 40(15), e115.
- Van der Krieken, S. E., Popeijus, H. E., Mensink, R. P., & Plat, J. (2015). CAAT/enhancer binding protein β in relation to ER stress, inflammation, and metabolic disturbances. *BioMed Research International*, 2015(2015), 1-13.
- Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., ... Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(1), 554-566.

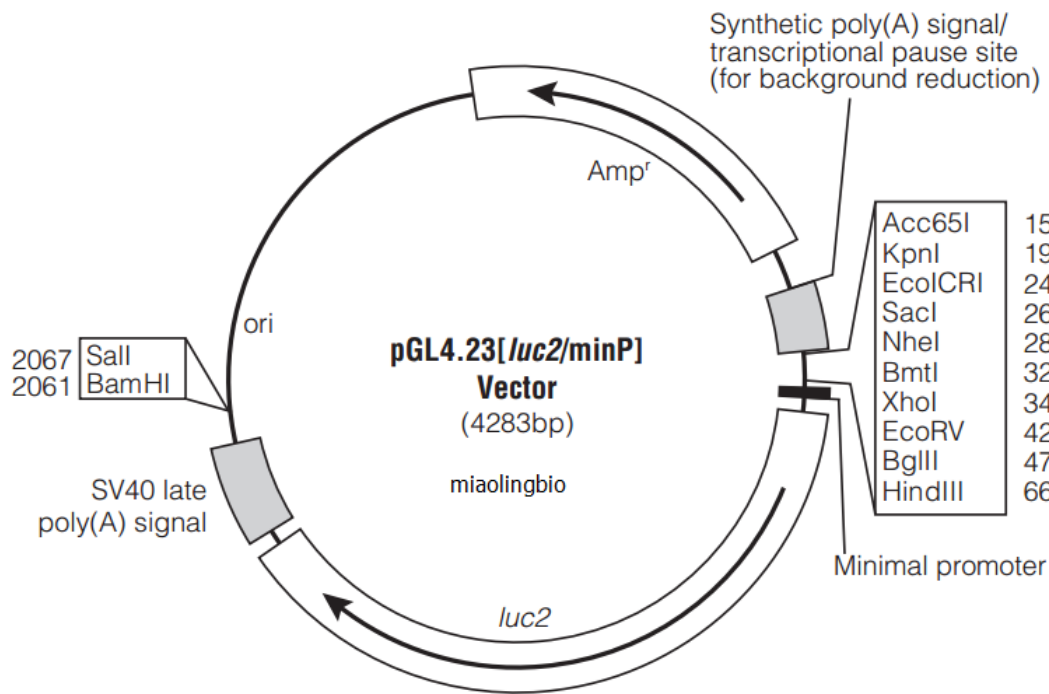
- Visel, A., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., & Pennacchio, L. A. (2009a). Functional autonomy of distant-acting human enhancers. *Genomics*, 93(1), 509-513.
- Visel, V., Rubin, E. M., & Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, 461(7261), 1-19.
- Wade, D. P., Lindahl, G. E., & Lawn, R. M. (1994). Apolipoprotein(a) gene transcription is regulated by liver-enriched trans-acting factor hepatocyte nuclear factor 1 α . *The Journal of Biological Chemistry*, 269(31), 19757-19765.
- Wong, Y. S., Lee, N. K., & Omar, N. (2016). GMFR-CNN: An integration of gaped motif feature representation and deep learning approach for enhancer prediction. *Association for Computing Machinery*, doi:10.1145/3029375.3029380, 41-47.
- Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., & Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, 20(9), 1377-1419.
- Yagi, R., Zhu, J., & Paul, W. E. (2011). An updated view on transcription factor GATA-3 mediated regulation of T_H1 and T_H2 cell differentiation. *International Immunology*, 23(7), 415-420.
- Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Maden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13(1), 134-134.
- Yip, K Y., Cheng, C., & Gerstein, M. (2013). Machine learning and genome annotation: A match meant to be? *Genome Biology*, 14(5), 205-205.
- Zaidi, S. K., Young, D. W., Montecino, M., Lian, J. B., Stein, J. L., van Wijnen, A. J., & Stein, G. S. (2010). Architectural epigenetics: Mitotic retention of mammalian transcriptional regulatory information. *Molecular and Cellular Biology*, 30(20), 4758-4766.
- Zhang, G., Shi, J., Zhu, S., Lan, Y., Xu, L., Yuan, H., ... Li, X. (2018). DiseaseEnhancer: A resource of human disease-associated enhancer catalog. *Nucleic Acids Research*, 46(1), D78-D84.

- Zhou, X., Cain, C. E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E. R., ... Gilad, Y. (2014). Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biology*, 15(1), 547.
- Zhu, Y., Sun, L., Chen, Z., Whitaker, J. W., Wang, T., & Wang, W. (2013). Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Research*, doi:10.1093/nar/gk826, 1-12.

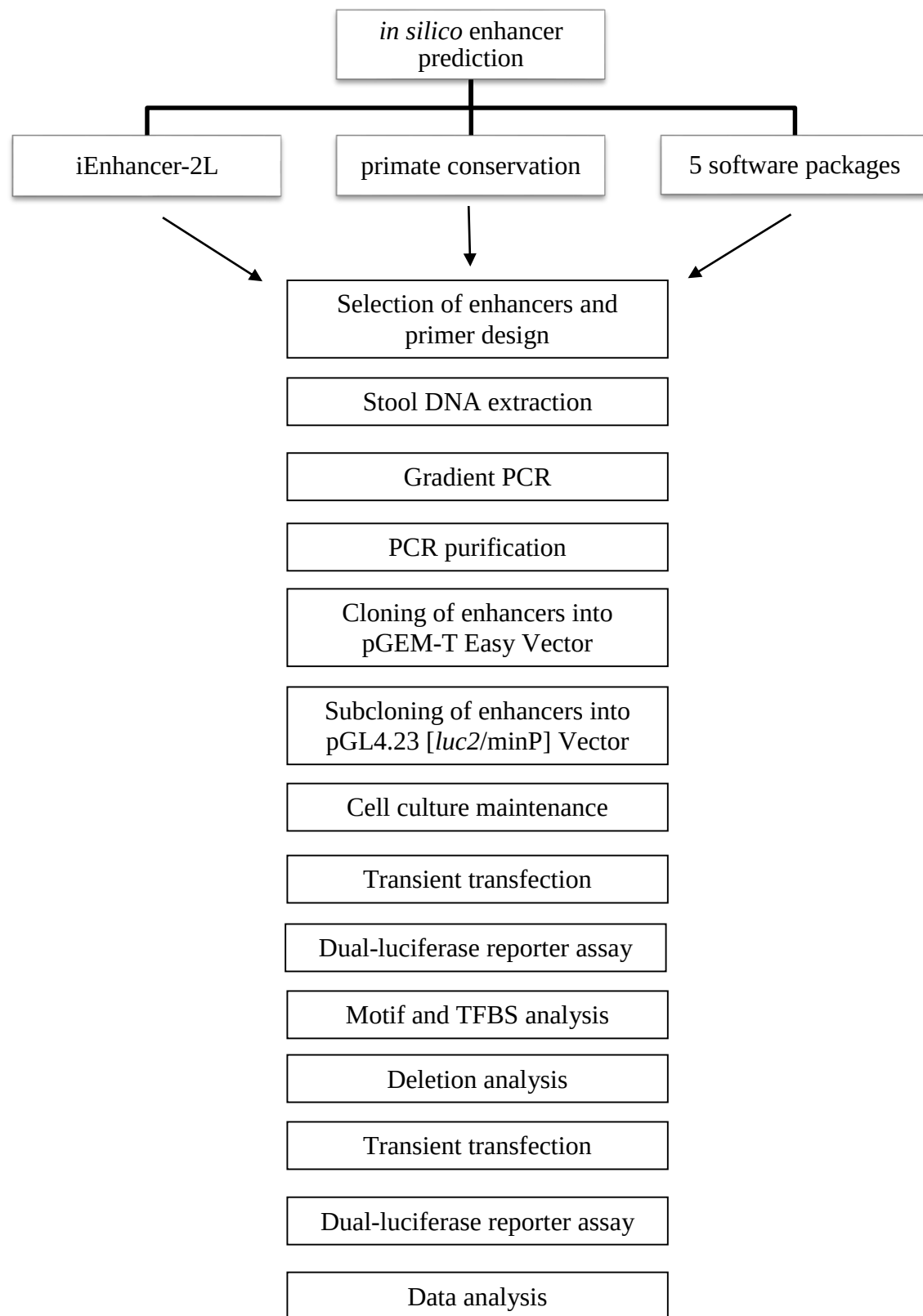
APPENDICES



The plasmid map for pGEM-T Easy Vector.



The plasmid map for pGL4.23[*luc2*/minP] Vector.



The methodology flow chart.