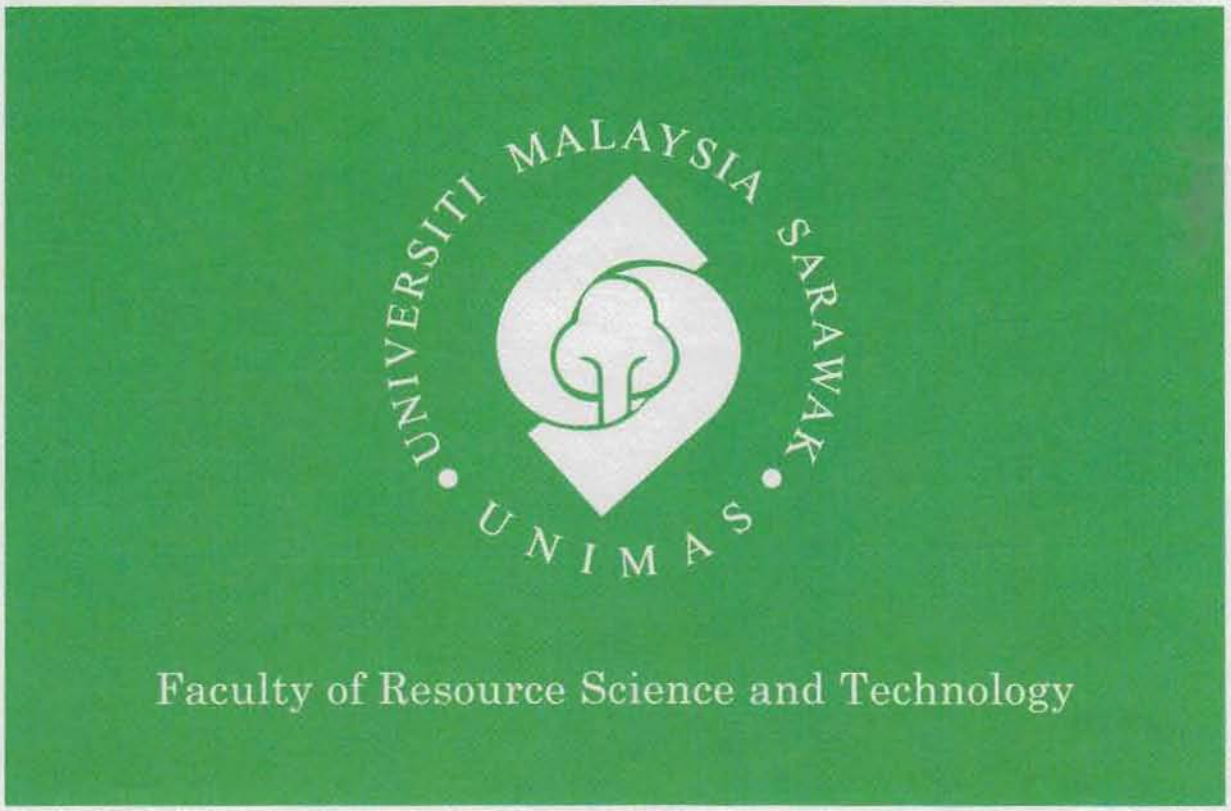

CLONING AND ACTIVITY DETERMINATION OF ENHANCER-LIKE SEQUENCE FROM PROBOSCIS MONKEY *(NASALIS LARVATUS)*

Lee Ying
(47304)

Bachelor of Science with Honours
(Resource Biotechnology)
2017

Cloning and Activity Determination of Enhancer-like Sequence from Proboscis Monkey (*Nasalis larvatus*)

LEE YING (47304)

A Thesis Submitted
in Partial Fulfilment of the Requirement for the Degree of
Bachelor of Science with Honours (Resource Biotechnology)

Supervisor: Dr. Chung Hung Hui

Resource Biotechnology Programme
Department of Molecular Biology

Faculty of Resource Science and Technology
Universiti Malaysia Sarawak
2017

UNIVERSITI MALAYSIA SARAWAK

Grade: _____

Please tick (☑)

Final Year Project Report ☒

Masters ☐

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 12 of June 2017.

Student's Declaration:

I, LEE YING (47304) from FACULTY OF RESOURCE SCIENCE ANDTECHNOLOGY (FRST) hereby declare that the work entitled CLONING AND ACTIVITY DETERMINATION OF ENHANCER-LIKE SEQUENCE FROM PROBOSCIS MONKEY *(Nasalis larvatus)* is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

12 JUNE 2017 — Date submitted

LEE YING (47304) — Name of the student (Matric No.)

Supervisor's Declaration:

I, _____ (SUPERVISOR'S NAME) hereby certify that the work entitled _____(TITLE) was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a * partial/full fulfillment for the conferment of _____ (PLEASE INDICATE THE DEGREE TITLE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: _____ Date: _____

(Name of the supervisor)

I declare that Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*

☑ **OPEN ACCESS**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature _____ Supervisor's signature: _____

12/6/2017 12/6/2017

Current Address:

NO. 618, TAMAN ANTEK AVENUE, JALAN SULTAN ABDULLAH, 36000 TELUK INTAN, PERAK.

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

[The instrument was prepared by The Centre for Academic Information Services]

Table of Contents

Contents	Page
Acknowledgement	I
Declaration	II
Table of Contents	IV
List of Abbreviations	VII
List of Figures	VIII
List of Tables	X
Abstract	1
CHAPTER 1: INTRODUCTION	2
CHAPTER 2: LITERATURE REVIEW	5
2.1 Enhancer	5
2.1.1 Gene expression regulation mechanism by enhancer	7
2.1.2 Transcription factor binding sites (TFBS)	8
2.2 Enhancer identification techniques	9
2.2.1 Enhancer features for selection	9
2.2.2 Genome wide search for enhancer	11
2.2.3 Experimental approach for enhancer recognition	12
2.2.4 Computational approach- rapid location of enhancer	12
2.3 Proboscis monkey (*Nasalis larvatus*)	14
2.3.1 Proboscis monkey as animal model in enhancer research	15
2.3.2 Chromosome number in proboscis monkey	16
CHAPTER 3: MATERIALS AND METHODS	18
3.1 Stool sampling from proboscis monkey	18
3.2 DNA extraction	19

3.3 Agarose gel electrophoresis..... 20

3.4 DNA analysis using UV spectrophotometry 21

3.5 Enhancer identification..... 21

3.6 Primer design 22

3.7 Gradient PCR 22

3.8 PCR purification..... 23

3.9 Gel Extraction..... 24

3.10 Preparation of *Escherichia coli* XL1-Blue competent cell..... 25

3.11 Cloning of PME001 into pGEM®-T Easy Vector 26

3.12 Restriction digestion..... 26

3.13 Cloning of PME001 into pGL3.0 Basic Vector with SV40 promoter 26

3.13.1 Ligation of purified PME001 into vector..... 26

3.13.2 Bacterial transformation of XL1-Blue competent cells..... 27

3.14 Blue/white colony screening..... 28

3.15 Colony PCR..... 28

3.16 Plasmid miniprep 29

3.17 Plasmid sequencing..... 30

3.18 Sequence verification and transcription factor determination..... 30

CHAPTER 4: RESULTS 31

4.1 DNA extraction..... 31

4.1.1 Genomic DNA verification using Cytochrome b gene 32

4.1.2 PCR purification..... 33

4.1.3 Sequencing result and BLAST..... 33

4.2 Enhancer identification- iEnhancer-2L outcomes..... 34

4.3 Primer design and synthesis..... 35

4.4 Temperature optimization for primers by gradient PCR..... 36

4.4.1 Gel extraction and Purification of PME001..... 37

4.5 Blue white screening..... 38

4.6 Colony PCR..... 42

4.7 Plasmid miniprep..... 43

4.8 Restriction digestion of plasmid DNA..... 44

4.9 PME001 sequencing result and its predicted sequence..... 45

4.9.1 PME001 sequencing result and ECR Browser..... 47

4.10 Transcription factor binding site (TFBS) analysis..... 48

CHAPTER 5: DISCUSSION..... 50

5.1 Cytochrome b gene in *Nasalis larvatus* genome..... 50

5.2 Genome wide enhancer search..... 50

5.3 Enhancer identification tool..... 51

5.4 Enhancer sequence analysis..... 53

5.5 Pattern analysis on transcription factor binding motif..... 54

CHAPTER 6: CONCLUSION..... 58

CHAPTER 7: REFERENCES..... 59

CHAPTER 8: APPENDICES..... 65

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
ANN	Artificial Neural Network
ChiP-seq	Chromatin Immunoprecipitation Sequencing
CRM	*cis*-regulatory module
DNA	Deoxyribonucleic acid
DLR	Dual Luciferase Reporter
*eve*S2E	*eve-skipped* stripe 2 enhancer
FBS	Fetal Bovine Serum
Ig	Immunoglobulin
IPTG	Isopropyl-β-D-thiogalactopyranoside
IUCN	International Union for Conservation of Nature
HIV-1	Human Immunodeficiency Virus Strain 1
LAR II	Luciferase Assay Reagent II
LB	Luria broth
NaOAc	Sodium acetate
NHP	Non-human primate
PBS	Phosphate-buffered saline
PLB	Passive lysis buffer
RNA	Ribonucleic acid
SV 40	Simian virus 40
SVM	Support vector machines
TAE	Tris-acetate Ethylenediaminetetraacetic acid
TF	Transcription factor

List of Figures

Figure		Page
2.1	DNA looping mechanism (Adapted from G. M. Cooper, 2000)	7
2.2	A schematic drawing of DNaseI hypersensitive site (Adapted from Chen *et al.*, 2015).....	11
2.3	Proboscis monkey (Nasalis larvatus) – long-nosed monkey (Adapted from www.monkeyworlds.com/proboscis-monkey/).....	15
2.4	G-banded karyotype of *N. larvatus* showing the in situ hybridization results. The proboscis monkey chromosomes are numbered below, and the homology with human chromosomes is on the right. (Adapted from Bigoni *et al.*, 2003).	17
4.1	The gel electrophoresis diagram obtained after DNA extraction. Lane L was loaded with 1kb DNA ladder (TransGen, China) while Lane 1 was loaded with DNA sample.....	31
4.2	The gel electrophoresis diagram of gradient Polymerase Chain Reaction of *Cyt b* gene. Lane L represents 1kb DNA ladder (TransGen, China). Lane 1 indicates negative control (55.6 °C). Lane 2, 3, 4, 5 and 6 represents PCR amplicon yielded in 54.3°C, 52.7°C, 51.3°C, 50.5°C and 50.0°C respectively.....	32
4.3	The gel electrophoresis diagram after gel purification. Lane L was loaded with 1kb ladder (TransGen, China) while Lane 1 was loaded with purified PCR products.....	33
4.4	The pairwise alignment between *Cyt b* gene sequence and *Nasalis larvatus* mitochondrion genome.....	34
4.5	iEnhancer-2L results after input of sequence for enhancer identification.....	35
4.6	The gel electrophoresis diagram of gradient PCR of PME001. Lane L represents 100bp DNA ladder (TransGen, China). Lane 1, 2, 3 and 4 represent PCR amplicon yielded in 52.7 °C, 51.7°C, 54.4°C 50.0°C respectively. Lane 5 indicates negative control (55.8°C).....	37
4.7	Secondary PCR of PME001. Lane L was loaded with 100bp ladder. Lane 1 and 2 was loaded with secondary PCR product of PME001 in annealing temperature of 51.7°C.....	37
4.8	Gel containing PCR amplicon in lane 1 and 2 was excised for subsequent purification process.....	38
4.9	(a) The negative control plate, (b) Experimental LAIX plate, (c) The secondary LAIX plate of pGEM®-T Easy Vector with PME001	40

insert.....

4.10 (a) Primary experimental LA plate, (b) The secondary LAIX plate of pGL3.0 basic vector with SV40 promoter and PME001 insert..... 41

4.11 The gel electrophoresis diagram of colony PCR. Lane L represents 100bp ladder (TransGen, China) and Lane 1 to 5 represent PCR amplicons of 5 single white colonies collected from secondary LA plates respectively while Lane 6 represents negative control (non-template control). The colony PCR was performed in optimized annealing temperature of 51.7°C..... 43

4.12 Plasmid miniprep agarose gel electrophoresis result. Lane L represents 1kb ladder (TransGen, China) while Lane 1 represents the uncut pGEM®-T Easy Vector with PME001 insert..... 44

4.13 Plasmid miniprep agarose gel electrophoresis result. Lane L represents 1kb ladder (TransGen, China) while Lane 1 represents the uncut pGL3.0 Basic Vector with SV40 promoter and PME001 insert..... 44

4.14 Agarose gel electrophoresis result of *BamHI* and *SalI* double restriction enzyme digestion. Lane L_1 and L_2 represent 1kb ladder and 100bp ladder (TransGen, China) respectively. Lane 1 represent digested pGEM®-T Easy Vector plasmid band and PME001 insert band..... 45

4.15 Agarose gel electrophoresis result of *BamHI* and *SalI* double restriction enzyme digestion. Lane L_1 and L_2 represent 1kb ladder and 100bp ladder (TransGen, China) respectively. Lane 1 represent digested pGL3.0 Basic Vector with SV40 promoter plasmid band and PME001 insert band..... 45

4.16 The pairwise alignment analysis of *Nasalis larvatus* PME001 and *Homo sapiens* chromosome 16..... 47

5.1 The flowchart of the two-layer predictor of iEnhancer-2L. Layer-I of sub-predictor will identified for potential enhancer sequence followed by identifying their strength in Layer-II of sub-predictor. Training dataset I and II was applied to train Layer-I and II sub-predictors respectively. (Adapted from *iEnhancer-2L : A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition* by Liu *et al.*, 2016)..... 53

List of Tables

Table		Page
3.1	List of PCR reaction mixture components for *Cyt b* gene and negative control	23
3.2	List of PCR reaction mixture components for PME001	23
3.3	Thermal cycling parameters for 35 PCR cycles of *Cyt b* and PME001	23
3.4	List of restriction digestion components	26
3.5	List of ligation reaction components	27
3.6	List of colony PCR reaction mixture for PME001	29
4.1	The characteristic of *Cyt b* gene obtained from BLASTn	34
4.2	The number of distinct types of enhancers predicted by iEnhancer-2L software	35
4.3	The characteristics of predicted enhancer sequence primer pair synthesis	36
4.4	The characteristic of *Nasalis larvatus* PME001 sequence obtained from BLASTn	46
4.5	Transcription factor binding sites that appear in sequence input of PME001 and human region. The lists of outcomes are obtained from MATCH™ tool	49
4.6	The condition of transcription factors (TF) when nucleotide variation occurs within PME001 and human sequence. The underline sequence represents nucleotides which are not conserved	49

List of Appendices

Appendix		Page
A	pGL3-Basic Vector map, multiple cloning size and sequences	65
B	The sequencing results of *cyt b* gene of *N.larvatus*	66
C	The multiple alignment sequence analysis of predicted enhancer sequence from *Nasalis larvatus* with five related primate species, human (*Homo sapiens*), bornean orangutan (*Pongo pygmaeus*), olive baboon (*Papio anubis*), common chimpanzee (*Pan troglodytes*) and western gorilla (*Gorilla gorilla*)	67
D	The sequencing results of pGL3.0 with PME001 insertion.....	70
E	The sequence alignment between PME001 enhancer and its expected enhancer sequence.....	71
F	Pairwise alignment analysis with existing vertebrate genome of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and rhesus macaque (*Macaca mulatta*)vin ECR Browser.....	73
G	Multiple sequence alignment outcome of PME001 with three primate species which are human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and rhesus macaque (*Macaca mulatta*).....	74

Cloning and Activity Determination of Enhancer-like Sequence from Proboscis Monkey (*Nasalis larvatus*)

Lee Ying

Resource Biotechnology Programme
Department of Molecular Biology
Faculty of Resource Science and Technology
Universiti Malaysia Sarawak

ABSTRACT

Enhancer is a distal regulatory element which plays a significant role in activating gene transcription. The hallmark of enhancer is critical, such that its spatiotemporal pattern in gene regulation has encouraged various research interests on this molecule. Enhancer identification has remained the greatest challenge as different enhancer possesses distinct function in different cell types. *Nasalis larvatus* which was indigenous in island of Borneo is a candidate species for us to explore its regulatory gene element due to some similarity that share within its genome and that of human. The main aim of this research is to isolate and clone potential enhancer sequence from *Nasalis larvatus* and subsequently analyse the transcription factor binding sites (TFBS) present in liver. Initially, putative enhancer was determined *in-silico* via the software iEnhancer-2L and strong enhancer of more than 500 bp was selected with at least 90% similarity as compared to other primates. Primers were designed for PCR amplification based on conserved domains from multiple alignments of four primate species. The amplicon was then cloned into pGL 3.0 basic vector modified with SV40 promoter insertion. The size of enhancer after sequencing was 795 bp which is deviated from expected size of 823 bp. Alignment analysis reveals that the enhancer was not conserve that it contains a number of nucleotide variations. The study of TFBS was then conducted by using MATCH™ programme and several liver specific TFBSs such as AP-1, c/EBPβ, CHOP c/EBPα, HNF-1, HNF-3β and NF-1. Finding of synonymous and non-synonymous nucleotide variation within TFBSs have given some novel insights on their relationship with gene expression output. Further enhancer activity determination is necessary to confirm the above status but due to time limitation it is not achievable.

Keywords: Enhancer, i-Enhancer-2L, *Nasalis larvatus*, pGL3.0 Basic Vector, transcription factor

ABSTRAK

Enhancer merupakan unsur kawal atur distal yang penting dalam merangsang transkripsi gen. Ciri-ciri enhancer yang kritikal, misalnya corak spatiotemporal dalam perarturan gen telah menarik minat pelbagai penyelidik untuk mengkaji molekul ini. Proses dalam mengenalpasti enhancer masih bercabar disebabkan fungsinya yang unik dalam jenis sel yang berbeza. Nasalis larvatus telah dicalonkan dalam kajian ini kerana ia bukan sahaja asli di kawasan Borneo malah sebahagian genomnya amat merupai dengan manusia. Tujuan kajian ini adalah untuk mengasingkan dan mengeklon enhancer daripada Nasalis larvatus serta membuat analisis tentang tapak lampiran faktor transkripsi yang wujud dalam sel hati. Pada mulanya, pemilihan enhancer dilaksanakan secara in-silico dengan menggunakan iEnhancer-2L dan kriteria untuk pilihan adalah berdasarkan kekuatan, saiz yang melebihi 500 bp serta mengandungi sekurang-kurangnya 90% persamaan dengan spesies primat yang lain. Sepasang primer telah direka berdasarkan domain terpelihara dalam "multiple sequence alignment" empat species primat untuk penggandaan melalui PCR. Pengganda itu telah diklon ke dalam pGL 3.0 vektor asas sisipan SV40 promoter. Keputusan sequencing menunjukkan saiz enhancer sebagai 795 bp yang amat melencong dengan jangkaan saiz 823 bp. Analisis penjajaran pula mendedahkan bahawa enhancer tidak dipelihara kerana wujudnya sebilangan variasi nukleotida. Program MATCH™ telah digunakan dan sebilangan tapak lampiran faktor transkripsi telah didapati seperti AP-1, c/EBPβ, CHOP c/EBPα, HNF-1, HNF-3β dan NF-1. Pendapatan bahawa variasi nukleotida yang sinomin dan tidak sinonim amat mempengaruhi ekspresi gen. Oleh itu, aktiviti enhancer harus diuji supaya status gen dapat disahkan tetapi ujian itu tidak dicapai disebabkan kekangan masa.

Kata kunci: Enhancer, , i-Enhancer-2L, Nasalis larvatus, pGL 3.0 vektor asas, faktor transkripsi

CHAPTER 1: INTRODUCTION

Enhancer serves as a fundamental regulatory element for gene expression during cell development and differentiation (Shlyueva *et al.*, 2014). It helps to regulate the activity of RNA polymerase at promoter sequence through the formation of DNA looping to increase gene transcription (Cooper, 2000). The hallmark of enhancer is that it can stimulate transcription in spatiotemporal patterns, either upstream or downstream as well as in any distance and orientation (Maston *et al.*, 2006).

Transcription factors (TFs) compromise of wide range of proteins and their complex multiple interactions with particular DNA binding motif on enhancer have provided substantial function in regulating gene expression, either stimulate or repress gene transcription (Kranz *et al.*, 2011; Makeev *et al.*, 2003). An active enhancer is usually occupied by various TFs and thus this fact has been exploited to search for candidate enhancer (Maniatis *et al.*, 1987; Palstra & Grosveld, 2012). TFs are lineage-specific as well as developmental stage specific whereby inappropriate binding may lead to abnormal gene expression (Palstra *et al.*, 2012).

Nowadays, massive amount of genomic data has been made available into public domain such as GenBank database to facilitate data mining by biological researcher using computational techniques (Wang *et al.*, 2013). Enhancer identification is now made use of these genomic data. Various software tools such as i-Enhancer 2L and DEEP were designed with different computational algorithms to determine putative enhancer sequence along entire genome (Shlyueva *et al.*, 2014). The sets of algorithms are created based on several enhancer features such as comparative genomic features, transcription factor binding motifs as well as epigenetic features (Wang *et al.*, 2013). However, current

computational approach alone is not sufficient to validate an enhancer unless it has been experimentally tested to confirm its precision.

Proboscis monkey (*Nasalis larvatus*) which was endemic to island of Borneo is a fascinating organism to be studied despite of its extraordinary morphology. From molecular viewpoint, the genomic structure of proboscis monkey was quite similar to that of human genome (Bigoni *et al.*, 2003). Generally, its chromosome 18 is homologous to human chromosome 6 and has held much of the protein-coding gene information which is extensively useful in medical genetic research (Mungall *et al.*, 2003). Yet, the size proboscis monkey genome was almost identical to human genome size of approximately 3000Mb (Rasko & Downes, 1994). There is still limited approach in study of gene regulatory elements in proboscis monkey, thus make it an interesting topic to be discovered.

Currently, the knowledge gap in terms of gene regulatory element is still existed, especially enhancer due to its flexibility and undefined spatiotemporal pattern in regulating gene transcription (Wang *et al.*, 2013). This research was generally focused on isolation and cloning of enhancer sequence to deal with the problem as mentioned. It is hypothesized that strong enhancer sequence is present in chromosome 18 of *N. larvatus* and it is highly conserved among various primate species. Furthermore, it is expected that the enhancer can be cloned into a luciferase gene vector for subsequent assay application to study enhancer activity and their relationship to gene expression level. Advance study on this sequence is necessary to understand its mutation effect on regulating gene expression of an organism. Perhaps this information is meaningful in disease manifestation (Sakabe *et al.*, 2012).

The objectives throughout the following research are:

1) To isolate enhancer from *Nasalis larvatus* chromosome 18.
2) To clone enhancer into pGL3.0 basic vector with SV 40 promoter insertion.
3) To study the liver specific transcription factor binding sites embedded within enhancer sequence.

CHAPTER 2: LITERATURE REVIEW

2.1 Enhancer

Generally, enhancers are well-known as *cis*-acting DNA sequences which promote gene transcription (Pennacchio *et al.*, 2013). Enhancers were primitively identified by Walter Schaffner in 1981 within a specific region of SV40 virus genome (Cooper, 2000). It contains two 72 bp repeats which facilitate viral replication and transcription in infected host cell (Levine, 2010). On the other hand, the first human enhancer was identified in mammalian B lymphocytes which modulated immunoglobulin (Ig) heavy chain gene expression (Levine, 2010).

Enhancer which is typically measured 200bp to 1kb in length is surrounded by cluster of transcription factors binding sites (TFBSs) that interact cooperatively to enhance gene transcription (Levine, 2010). Enhancer-bound transcription factors are responsible to recruit chromatin remodelling complexes for de-compaction of chromatin fibre so that DNA can access to other proteins easily (Ong & Corces, 2011). As a gene regulatory element, enhancer is also associated with specific repressor binding sites to restrict certain gene overexpression and subsequently to prevent gene mutation (Levine, 2010). Enhancers are also found to carry epigenetic information in the form of specific histone modification for future gene expression and gene expression activation (Ong *et al.*, 2011).

Enhancer grammar refers to spatial arrangement of TFBSs along the entire length of enhancer and this has led to introduction of two contrasting model namely enhanceosome and billboard model (Rubinstein & de Souza, 2013). The former model represents transcription factors that work synergistically in precise orientation for a functional enhancer activity whereas the latter model consist of transcription factor that bind in a more flexible order but still remain some of the enhancer function (Rubinstein *et*

al., 2013). Hence, both models indicate that distinct enhancer elements will play different roles in developmental gene expression and study on its transcription factor complexes can further aid in structure and functional analysis of enhancer sequence.

Enhancer is modular such that a single promoter can be activated by several remote enhancer elements from distinct gene locus (Levine, 2010). Studies have shown that uncoupling process of enhancer from target genes is a key insight to understand evolutionary pattern within biodiversity (Levine, 2010). Although enhancer is evolutionary conserved, accumulation of random mutation will either produced novel enhancer or loss of enhancer that leads to changes over evolutionary pathway (Rubinstein *et al.*, 2013).

Mutation in particular enhancer sequence will alter gene expression which may bring along little or no consequences in other region. A significant example of enhancer changes but its function retain is best demonstrated in *Drosophila* species, whereby the different arrangement of TFBSs within *eve-skipped* (*eve*) stripe 2 enhancer (S2E) can lead to equivalent gene expression pattern upon transfection into *Drosophila melanogaster* embryonic cell line (Rubinstein *et al.*, 2013).

Enhancer evolution generally involves phenotypic diversification within animal diversity due to regulatory gene mutation. For instance, marine stickleback contains bony spines in pelvic region to act as armour from enemy attack whereas freshwater stickleback usually lack of spine due to deletion of pelvic enhancer that regulating the *Pitx1* gene (Chan *et al.*, 2010). On the other hand, acquisition of novel enhancer could also result in innovation of new gene expression pattern; subsequently cause morphological variation between closely-related species. More detailed studies on enhancer mutation that lead to species evolution and adaptation should be conducted to establish genuine knowledge on gene regulatory mechanism.

2.1.1 Gene expression regulating mechanism by enhancer

Throughout the studies of numerous enhancers, it has concluded that this specific element typically regulates transcription in a spatiotemporal-specific pattern which means that they function independently of both the distance and orientation relative to promoter (Maston *et al.*, 2006). Enhancer usually situated distally away from core promoter, about few hundreds kbp upstream or downstream in forward or backward orientation (Shlyueva *et al.*, 2014).

The mechanism involved in regulating gene expression by enhancer over such long physical distance was well-demonstrated in Figure 2.1 as shown below. DNA looping model has contributed to this mechanism whereby the activated enhancer loops to core promoter regions by intervening DNA to regulate release of RNA polymerase II from promoter proximal-pausing site (Cooper, 2000). A rapid gene transcription is thus stimulated. According to Levine (2010), the long range looping could be stabilized by cohesin and non-coding RNAs to facilitate enhancer-promoter interactions while the assembly of mediator complex initiate right after the looping to stimulate transcription.

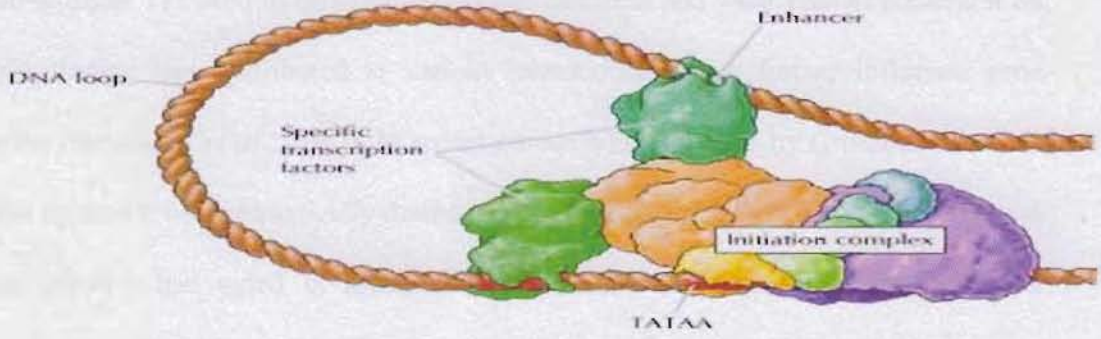

Figure 2.1. DNA looping mechanism whereby the activated distal enhancer binds to the core promoter (TATAA box) by intervening of DNA strands. (Adapted from *The Cell: A Molecular Approach (2nd ed.)* by G. M. Cooper, 2000).

2.1.2 Transcription factor binding sites (TFBS)

Transcription factors (TFs) consists of wide range of proteins which will bind to cognate binding domain on enhancer to orchestrate specific gene expression (Krantz *et al.*, 2011). Various characteristics have influenced TF binding to induce enhancer function, particularly the chromatin state of genomic locus, TF binding affinity and complex multiple interactions among TFs and that on enhancer to stimulate gene expression (Grossman *et al.*, 2017). In past decade, characterization of several TF binding motifs were done *in-vitro* as well as genome-wide mapping of TFBS were done *in-vivo* (Badis *et al.*, 2009; Johnson *et al.*, 2007). Furthermore, experimental approaches have allowed functional characterization of enhancer activity by TF binding (Grossman *et al.*, 2017).

However, scientists have revealed that only a small portion of potential TFBSs are occupied by their respective TFs in eukaryotic genomes and yet their substantial function are varied across cellular context (Spitz & Furlong, 2012). Hence, identifying and characterizing of TF interactions has proven difficult because question arises basically involved whether TFs bind to different functional elements and what kind of constraint on motif positioning has contributed to various interactions which further influence gene expression (Grossman *et al.*, 2017). In recent research as proposed by Grossman *et al.* in 2017, the approach of systematically dissecting the features contributed to TF binding and enhancer activity had aided in identification potential TF interactions for functional characterization as well as revealing sets of grammatical rules governing the activity of gene regulatory elements (Grossman *et al.*, 2017). This approach thus facilitates the understanding how gene expression works, either drives a normal biological processes or leads to aberrant gene expression with disease as an ultimate consequences (Grossman *et al.*, 2017).

TFBSs can actually play a pivotal role in predict enhancer activity whereby in a research done by Dogan *et al.* (2015), it indicates that TFBSs which occupied by key TFs can predict the enhancer activity more precise than histone modification and chromatin accessibility.

2.2 Enhancer identification techniques

Identification of enhancers is challenging and complex, one of the reasons is that the enhancer sequence can disseminate in any gene locus and orientation, thus resulting in more complex and huge searching effort (Pennacchio *et al.*, 2013). Yet in this postgenomic era, identification of enhancers from enormous amount of genomic data is impossible without the aid from computational approaches. Nonetheless, computational methods are only considered to be an alternative to search for putative enhancer sequence. Instead, experimental approach for enhancer identification should not be abandoned to confirm the accurateness of predicted enhancer.

2.2.1 Enhancer features for selection

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has become a widely used approach which allows mapping of genomic location of transcription factor binding and histone modifications in different cell types (Landt *et al.*, 2012). This method has developed more than a decade ago and still applying in today's research in abundant (Landt *et al.*, 2012). ChIP-seq enables genome-wide identification of sequence-specific transcription factor binding sites (TFBSs) as well as other DNA binding proteins to localized gene regulatory element (Raha *et al.*, 2010). According to Wang *et al.* (2013), ChIP-seq uses to measure the binding affinity between transcription factor and DNA sequences indirectly by calculating the number of transcription binding events.

Histone acetyltransferase P300 which is a co-activator belongs to a chromatin feature to indicate putative enhancer sequence (Zhu *et al.*, 2013).

Another way of estimating enhancer position is through the detecting of DNaseI hypersensitive site in which active enhancer are often found within these TFBSs (Liu *et al.*, 2016). Figure 2.2 has shown the DNaseI hypersensitive site for various regulatory elements. DNaseI hypersensitivity assay is initiated by de-compaction of chromatin to expose DNA binding site for transcription factor which in turn lead to increased susceptibility of wide-open DNA site to digestion by enzyme DNaseI (Lu & Richardson, 2004). This kind of assay is not considered as an ideal model for predicting active enhancer because promoter region may also contain DNA hypersensitive site (Liu *et al.*, 2016).

In addition, high-throughput sequencing of epigenomic feature by ChIP-seq was introduced whereby the level of histone modifications is mapped directly to predict enhancer location precisely (Pennacchio *et al.*, 2013). Some specific histone marks such as H3 lysine 4 monomethylation (H3K4me1), H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 acetylation (H3K27ac) are associated within active enhancer (Rubinstein *et al.*, 2013). The level of histone modification correlated to enhancer activity is well-demonstrated by elevated H3K4me1 level and depleted H3K4me3 level (Wang *et al.*, 2013).

Although experimental-based approach of evaluating transcription factor binding affinity and epigenetic characteristic can further indicate enhancer activity but this method is consider to be expensive and time-consuming. It is still essential to develop a more advance model, perhaps with the help of computational algorithms to achieve highly specific enhancer prediction (Wang *et al.*, 2013).

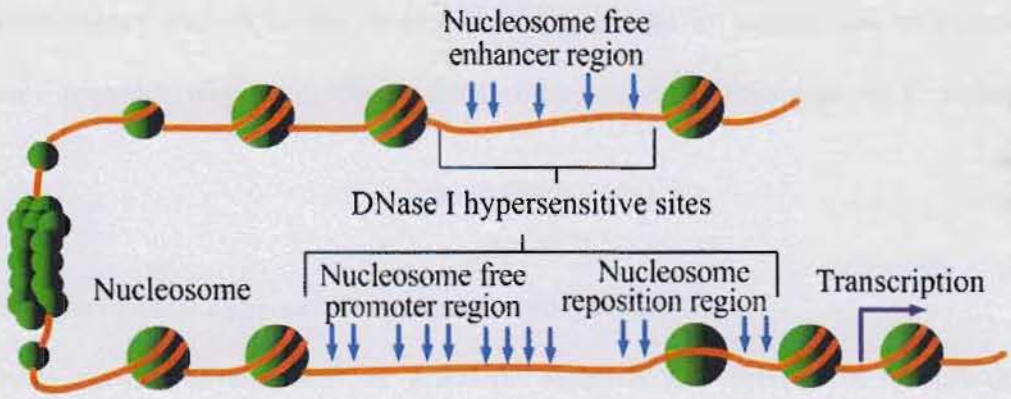

Figure 2.2. A schematic drawing of DNaseI hypersensitive site in open chromatin region for enhancer site recognition. (Adapted from *Molecular Biosystems* by Chen et al., 2015).

2.2.2 Genome wide search for enhancer

Identification of enhancer at genome scale can be achieved by using phylogenetic footing to compare genome sequence of different species by relying on evolutionary conservation of non-coding regions across diverse species (Pennacchio *et al.*, 2006). The idea of comparative genomic analysis was inspired by the evolutionary constraint whereby a functional DNA sequences such as enhancer usually evolve slower than those consist of non-functional sequence, thus facilitating in enhancer identification (Rubinstein *et al.*, 2013). According to Cooper *et al.* (2004), evolutionary conserved regions within human genome are interpreted to enrich with enhancer by comparing with the neutrally mutated sequences between human, rat and mice.

Comparative genomic analysis is suitable to represent the potential of a locus to be an enhancer sequence within genome of particular species (Wang *et al.*, 2013). In genome-wide studies, the putative enhancer sequence was predicted also by identifying their features, particularly the chromatin modification structure by using current technological advances (Shlyueva *et al.*, 2014). The enhancer activity is hardly to determine by just calculating the conservation score (Pennacchio *et al.*, 2006). Nonetheless, the knowledge

on evolutionary pattern is still limited for vast amount of species due to incomplete genome sequence alignment, thus making comparative genomic analysis a challenging task.

2.2.3 Experimental approach for enhancer recognition

A more mature development in scientific research has solved the complexity of determining enhancer sequence within wide genome. Originally, transgenic assay was applied to identify individual enhancer whereby the transgenic organisms or cell lines are assessed for reporter gene expression to evaluate transcriptional activity regulated by enhancer (Rubinstein *et al.*, 2013). This technique involves cloning of candidate enhancer sequences upstream of a minimal promoter fused to a reporter gene (Rubinstein *et al.*, 2013).

By using this method for more sophisticated studies, it has found out that gene was actually controlled by several enhancers simultaneously (Rubinstein *et al.*, 2013). For example, the expression developmental gene such as *Sox2* gene from chicken embryo was regulated by 11 enhancers (Uchikawa *et al.*, 2003). Although transgenic assay can provide such spatiotemporal expression information but it ended up to be laborious and costly, hence results in low throughput for genome-wide analyses (Banerji *et al.*, 1981).

2.2.4 Computational approach- rapid location of enhancer

In the post-genomic era, enormous amount of biological data has been made available through various public domains such as GenBank database to facilitate rapid enhancer identification by computational approaches (Wang *et al.*, 2013). The earlier computational algorithms were basically refer to evolutionary conserved functional DNA regions that evolve slower than non-functional one but the data *per se* is still insufficient to determine