國立臺灣大學基因體與系統生物學學位學程

（合辦單位：中央研究院）

博士論文

Genome and Systems Biology Degree Program
College of Life Science
National Taiwan University and Academia Sinica
Doctoral Dissertation

黑腹果蠅長非編碼RNA特性研究

Characterizing Long Non-coding RNAs in *Drosophila melanogaster*

陳玫如

Mei-Ju Chen

指導教授：陳倩瑜 博士
李文雄 博士

Advisor: Chien-Yu Chen, Ph.D.
Wen-Hsiung Li, Ph.D.

中華民國 105 年 7 月

July, 2016

doi:10.6342/NTU201601384

國立臺灣大學（碩）博士學位論文
口試委員會審定書

黑腹果蠅長非編碼 RNA 特性研究

Characterizing Long Non-coding RNAs in *Drosophila melanogaster*

本論文係陳玫如君（D99B48004）在國立臺灣大學基因體與系統生物學學位學程之博士學位論文，於民國 105 年 7 月 20 日承下列考試委員審查通過及口試及格，特此證明

口試委員：陳倩瑜 李文雄 （簽名）

（指導教授）

蔡懷寬 莊樹諄

吳君泰

系主任、所長 （簽名）

ACKNOWLEDGMENTS

“*The first step in wisdom is to know the things themselves.*” — Carolus Linnaeus (1735)

基因體與系統生物學為一個新穎學門，其所涉範疇廣泛，在我的博士生涯裡亦時逢茫然無據的時刻，感謝陳倩瑜教授與李文雄教授總能一語點領關鍵，引亮學生在龐雜的基因體與系統生物學網路裡該行的道路。

「重要的科學通常概念簡單，而簡單的事情通常困難」，這一語約莫道盡這篇博論成果背後曾有的汗水。而基因體與系統生物學學門之特殊，使得這個領域的研究往往無法僅是一人之功，也時常需要透過不同領域實驗室的合作，感謝吳君泰老師實驗室在黑腹果蠅的專業，讓我們在探討這個課題時能迅速尋得脈絡，也透過在君泰老師實驗室的生物實驗訓練，讓我能在這篇論文中做出關鍵的生物實驗證明支持。而在論文發表的過程中，多次與審查者間的論文修訂往來與數據的補充，也衷心感銘郭建言博士於倩瑜老師實驗室的博士後研究期間不吝對論文英文撰寫的多加指教，亦感謝倩瑜老師實驗室的學弟們所給予的強力後盾支持：昱行、祐榆、東祈、柏均、秉翰、翊安、張平，而終成這簡單而困難的研究，期望奠定黑腹果蠅 lncRNA 研究的堅實基礎。

從進入生物資訊領域而後從事基因體與系統生物學研究，該從大三進入倩瑜老師實驗室進行專題研究算起，迄今已歷十年，倩瑜老師亦師亦友的指導與栽培，以及其對於研究的熱情與態度，再再影響我許多，我想這些影響也將遠及一生，衷心感謝能遇良師。而亦也在老師的實驗室裡遇得一生伴侶──祐榆，一路支持我專心埋首研究，也因專業互補而與我相互磨鍊拋光整個研究成果，感激這樣的幸運，能得一這樣無論研究與生活皆相知的伴侶。

博士生涯終將階段性結束，然研究是一生志業，望尋得適宜舞台，承先啟後，貢獻所學，不負一路栽培。

中文摘要

次世代定序技術(Next-generation sequencing; NGS)開啟 RNA 領域研究的新紀元。過往認為只是轉錄訊號擾動的長非編碼 RNA (long non-coding RNA; lncRNA)，已由許多研究證實其在許多重要生理機制中扮演要角。然而，現今文獻對於重要模式生物—黑腹果蠅(*Drosophila melanogaster*)的 lncRNA 瞭解仍相當有限；究其原因，乃黑腹果蠅 lncRNA 的基礎資訊之稀缺所致。因此，本論文追根溯源，由四個面向對黑腹果蠅 lncRNA 進行系統性探究—(1) **收集與發現**：本論文開發一生物資訊方法，自我們產生的組織特異性 RNA-seq 資料鑑定出為數不少的新 lncRNAs，並與公開資訊可收集之已知 lncRNAs 整合，呈現迄今最新之黑腹果蠅 lncRNA 資料集；(2) **特性註解**：本論文採用大量的 RNA-seq 與 ChIP-seq 資料集(總計 93 組)增進現有 lncRNA 的註解資訊如轉錄方向與染色質特徵之品質，並進而觀察摘要出黑腹果蠅 lncRNA 的一般特性；(3) **基因表現**：本論文以 RT-qPCR 實驗驗證了挑選之 lncRNA 的基因表現，並彰顯 RNA-seq 技術平台用於發現 lncRNA 的結果具有相當的可信度；(4) **轉錄調控**：本論文提出一結合序列特徵探勘之生物資訊方法，系統性分析轉錄因子結合位(Transcription factor binding site; TFBS)於 lncRNA 啟動子出現與否，以及其與 lncRNA 基因轉錄調控的關聯性。結果顯示，當使用核小體佔據與跨物種保留性資訊，於共表現之編碼基因集進行序列探勘，其所得的序列特徵(或稱順式因子；*cis*-element)，多數與已知的 TFBS 相似；此外，這些順式因子可在共表現之編碼基因與 lncRNA 基因的啟動子區域同時觀察得見(較常見於第三期幼蟲至雄蟲階段共表現群集)，顯示出共表現之編碼基因與 lncRNA 基因具有被共同調控的可能性。簡言之，本論文彰顯系統性整合研究的優點，透過基因體與轉錄體資料的整合，大幅加速鑑別 lncRNA 的特性；而所得之觀察結果可作為黑腹果蠅 lncRNA 功能研究的堅實基礎。

關鍵詞：整合性研究、黑腹果蠅、長非編碼 RNA、RNA 定序技術、染色體免疫沉澱定序技術

doi:10.6342/NTU201601384

ABSTRACT

Recent advances in sequencing technology have opened a new era in RNA studies. Novel types of RNAs such as long non-coding RNAs (lncRNAs) have been found to play essential roles in biological processes. However, only limited information is available for lncRNAs in *Drosophila melanogaster*, an important model organism. Thus, this thesis aims at chracterizing fruit fly lncRNAs from four aspects: (1) collection and discovery; (2) annotation; (3) expression; and (4) regulation. I developed a computational approach to discover novel lncRNAs from the newly generated tissue-specific RNA-seq data, and then I combined the discovered lncRNAs with previously published lncRNAs into a curated dataset. Next, numerous RNA-seq and ChIP-seq datasets (93 sets) were used to improve the lncRNA annotation such as transcriptional direction and presence of conventional chromatin signatures. With these efforts, I summerized general characteristics of fruit fly lncRNAs in the thesis. In addition, I used RT-qPCR experiments to validate the expression of some randomly selected lncRNAs and demonstrated that RNA-seq is a reliable platform to discover lncRNAs. Moreover, I proposed a method to incorporate *de novo* motif discoveries to systemically investigate the presence of TFBSs in lncRNA promoters and how it is related to the regulation of lncRNA expression. The result revealed that most of the motifs (*cis*-elements) discovered from the co-expressed coding gene promoters are similar to the annotated TFBSs, where the motif dicscovery procedure considerd the information of nucleosome occupancy and evolutionary conservation. I also found that common *cis*-elements were usually observed in the promoters of the co-expressed coding and lncRNA genes in the development stages from L3 to male adlut. In conclusion, this thesis demostrated that integration of genomic and transcriptomic data can largely facilitate lncRNA discovery and characterization, and provided a solid foundation for studying the functions of lncRNAs in *D. melanogaster*.

Keywords: Integrative research, *Drosophila melanogaster*, Long non-coding RNA, RNA-seq, ChIP-seq

doi:10.6342/NTU201601384

TABLE OF CONTENTS

博士學位論文口試委員會審定書 i

ACKNOWLEDGMENTS ii

中文摘要 iii

ABSTRACT iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES xi

CHAPTER 1 Introduction 1

1.1 Challenges of lncRNA studies in *D. melanogaster* 3

Limited numbers of known lncRNAs in *D. melanogaster* 3

Incomplete annotation of lncRNAs in *D. melanogaster* 4

Reliability of lncRNA expression detected from RNA-seq data 5

Transcriptional regulation of lncRNA expression 5

1.2 Integrative approach for characterizing lncRNAs by utilizing genomics and trnascriptomics data 7

1.3 Thesis structure 10

CHAPTER 2 Related Works 12

2.1 Brief history of long non-coding RNAs studies 12

2.2 Integrative and systemic studies on lncRNAs 13

2.2.1 Related works for characterizing lncRNAs in in *Drosophila melanogaster* 13

2.2.2 Related works for transcriptional regulation of lncRNA expression 14

CHAPTER 3 Collection and Discovery of lncRNAs in *Drosophila melanogaster* ... 17

3.1 Known lncRNAs collected from databases and literatures 17

3.2 Novel lncRNAs identified from brain samples 18

3.3 Up-to-date list of long non-coding RNAs in *D. melanogaster* 19

3.4 Methods for collection and discovery of fruit fly lncRNAs 19

3.4.1 Collection of published lncRNAs 19

3.4.2 RNA-seq data of the fly brain 22

doi:10.6342/NTU201601384

3.4.3 Novel lncRNA discovery 23
CHAPTER 4 Annotation of the curated lncRNAs.. 26
4.1 Improving the annotation of the lncRNAs from Young *et al.* 26
4.2 Utilizing additional RNA-seq datasets to improve the annotation of the 4,599 curated lncRNA transcripts... 28
4.3 General characteristics of the fruit fly lncRNAs.. 31
4.3.1 Location distribution of lncRNAs in Genome 31
4.3.2 Length and structure of lncRNAs.. 34
4.3.3 Evolutionary conservation of lncRNAs 37
4.3.4 Supporting evidences for lncRNA expression in the developmental stages… 39
4.4 Methods for annotation of the curated lncRNAs 46
4.4.1 Improving the annotation of curated lncRNAs 46
4.4.2 Genomic and transcriptomic data for supporting lncRNA expression in the developmental stages 49
CHAPTER 5 Reliability of lncRNA expression 52
5.1 Reliability of the lncRNAs newly discovered identified from brain samples… 52
5.2 Experimental validation of a selected set from the curated lncRNAs by RT-qPCR.. 55
5.3 Details of the RT-qPCR experiments 58
CHAPTER 6 Regulation of lncRNA Expression.. 60
6.1 Hierarchical clustering of co-expressed coding and long non-coding genes……………………………………………………………………....61
6.2 *De novo* motif discovery on the promoters of co-expressed coding genes.. 68
6.2.1 Promoter regions of genes in *D. melanogaster* 68
6.2.2 Parameter tuning for the weights of nucleosome occupancy and evolutionary conservation while conducting *de novo* motif discovery 70
6.2.3 Evaluation of the discovered motifs... 72
6.3 Co-occurrence of TF binding motifs in the promoter regions of co-expressed coding and non-coding genes 74
6.4 Materials and methods for the proposed workflow..................................... 76
6.4.1 Collection of annotated transcription factor binding sites 76

doi:10.6342/NTU201601384

6.4.2 Hierarchical clustering 76

6.4.3 *De novo* motif discovery of cis-elements from co-expressed coding gene promoters 77

6.4.4 Identification of shared *cis*-element in the co-expressed lncRNA promoters 78

CHAPTER 7 Limitations of this work 80

CHAPTER 8 Conclusions and Future Directions 82

REFERENCE: 84

APPENDIX 90

List of Publications 90

Appendix Figures 94

Appendix Tables 95

doi:10.6342/NTU201601384

LIST OF FIGURES

Figure 1. Four challenges for characterizing lncRNAs in *D. melanogaster* ... 2

Figure 2. Characterizing lncRNAs in *D. melanogaster* 11

Figure 3. Procedures for discovering novel lncRNAs from RNA-seq data of the present study. The sequencing read datasets of mRNA and total RNA were respectively mapped to the reference genome sequence using TopHat and Cufflinks. Putative lncRNAs were then discovered by Cuffcompare. Sequencing reads were again mapped to the set of putative lncRNAs to construct the final set of novel lncRNAs.. 24

Figure 4. Distribution of lncRNA types in euchromatin. 33

Figure 5. Distribution of exon numbers in lncRNA and mRNA genes......... 35

Figure 6. Expression profiles at different developmental stages of fruit fly. (a) Averaged RPKM values at different developmental stages for lncRNAs and mRNAs. (b) Numbers of expressed transcripts (RPKM > 1) at different developmental stages for lncRNAs and mRNAs, respectively. 40

Figure 7. Expression profiles of the 2,926 lncRNAs which varied between at least two developmental stages 41

Figure 8. Histogram of members in a co-expressed lncRNA cluster among the 2,926 lncRNAs which varied between at least two developmental stages 43

Figure 9. Expression profiles of the lncRNAs which are co-expressed with at least 9 other lncRNAs (Namely, a co-expressed lncRNA cluster is selected for this figure while it has at least 10 members). 43

Figure 10. Analysis of chromatin signatures (Pol II, H3K36me3 and H3K4me3) in the curated lncRNA genes.. 44

Figure 11. Rules for classifying lncRNAs. Black arrows (transcripts) represent coding genes and colored transcripts are lncRNAs. (a) lncRNAs with intronic overlaps. This group includes lncRNAs (dark green and light green transcripts) located in intronic regions of coding genes (black transcripts). (b) Intergenic lncRNAs. This group includes lncRNAs (red transcripts) located in regions between two coding genes (black transcripts). (c) lncRNAs with exonic

doi:10.6342/NTU201601384

overlaps. This group includes lncRNAs (dark blue and light blue transcripts) overlapping exonic regions of coding genes (the black transcript). 47

Figure 12. Occupied regions for each chromatin signature 51

Figure 13. RT-qPCR experiments for a selected set of lncRNAs in brains. (a) 22 novel lncRNAs discovered in the present study were selected for validation. RpL 32 (a coding gene) and ROX1 (a non-coding gene) were included as positive controls. The horizontal line indicated −delta Ct ≥ 1. The rectangle indicated the five lncRNAs with considerably low expression, and was tested again by the second RT-qPCR experiment shown in (b). (b) The five lncRNAs from the rectangle of (a) were tested again by RT-qPCR with twofold amount of template cDNA. Ten FlyBase lncRNAs were included for comparison. The three FlyBase lncRNAs highlighted by the orange stars were selected because their RPKM values in our brain RNA-seq data was 0. 54

Figure 14. RT-qPCR experiments of a selected set of lncRNAs in male adults. G1: high expression with chromatin signatures (11 lncRNAs); G2: low expression with chromatin signatures (11 lncRNAs); G3: high expression without chromatin signatures (10 lncRNAs); and G4: low expression without chromatin signatures (10 lncRNAs). Three negative controls (un-transcribed region 1, 2, and 3) were all around zero. Stars were used to highlight the lncRNAs that were not from the databases (Orange stars: the selected lncRNAs from Young *et al*. [5]. Blue stars: the lncRNAs from the present study). The horizontal line indicated the cutoff (−delta Ct ≥ 2) used to define a validated lncRNA. Green stars: the transcripts that are now annotated as other types of transcripts by FlyBase, and thus were removed from the list of the curated lncRNAs in the present study. 56

Figure 15. Workflow for identifying shared *cis*-elements of co-expressed coding and long non-coding genes.. 60

Figure 16. Frequency distribution of clusters along with different member numbers in a cluster 62

Figure 17. Expression profiles for the lncRNAs of the 27 co-expressed clusters……….. 65

Figure 18. Second-phase hierarchical clustering for the 27 co-expressed clusters with at least 50 transcripts. The rainbow color bar indicates the groups categorized by the second-phase hierarchical clustering. The gray color bar

doi:10.6342/NTU201601384

represents the number of lncRNAs within a cluster. Details of group and cluster information can be found in Table 13. 66

Figure 19. Second-phase hierarchical clustering for the co-expressed clusters with at least 30 transcripts. The rainbow color bar indicates the groups categorized by the second-phase hierarchical clustering. The gray color bar represents the number of lncRNAs within a cluster. Details of Group 1-7 can be found in Table 13, while Group X represents the clusters that contain members of 30~49 transcripts……. 67

Figure 20. Distribution and Conservation scores (CS) analysis of the 2,059 annotated binding sites collected from REDfly database [90] (170 TFs and 2,048 target genes included). (a) Position distribution. The averages CS of TFBSs located within (−500 to +200 bp) is 0.482; (b) Frequency of TFBSs that have a CS value ≥ 0.482…. 69

Figure 21. Parameter tuning for the weights (*b*, *c*) which are given to nucleosome occupancy and evolutionary conservation. Different colors denote different weights for nucleosome occupancy. The colors, (blue, red, green, orange), indicate $b \in (0, 1, 2, 3)$. Different types of lines represent different weights for evolutionary conservation. The line types, (solid line, thick broken line, broken line), indicate $c \in (1, 2, 3)$. 71

Appendix Figure 1. Parameter tuning for different pattern supports (ratio of pattern-hit promoters/all promoters in the positive set) and different weights used for pattern ranking. Precision is calculated by the ratio of (True Positives/Predicted instances), and presented as percentage. 94

doi:10.6342/NTU201601384

LIST OF TABLES

Table 1. Numbers of fly lncRNAs from different data sources...................... 4

Table 2. Statistics of the public data used for studying transcriptional regulation in yeast, fruit fly and human 6

Table 3. Summary statistics of datasets used in thesis. 8

Table 4. Statistics of exon numbers in lncRNA and mRNA genes from different sources. 27

Table 5. Statistics of transcriptional direction in the lncRNA genes from different sources. 28

Table 6. Types of lncRNA transcripts. 29

Table 7. The number of lncRNAs from three different sources in each of the euchromosomes and heterochromosomes... 32

Table 8. Length of lncRNA transcripts.. 34

Table 9. Coverage of scored bases in UCSC 15-way alignment.................. 38

Table 10. Average conservation scores of each chromosome....................... 38

Table 11. Conservation scores of different sequence groups 39

Table 12. Statistics of clusters with different cutoff of correlation 61

Table 13. Summary of the selected clusters which contain at least 50 transcripts…… 64

Table 14. Summary of *de novo* motif discovery results 73

Table 15. Investigation of similarity between lncRNA and mRNA promoters…… 73

Table 16. Summary of *cis*-elements shared by co-expressed coding and long non-coding genes 75

Appendix Table 1. Primer list of the selected lncRNAs for RT-qPCR experiments…………… 95

Appendix Table 2. Raw Ct values of RT-qPCR experiments for un-transcribed regions and the selected lncRNAs. 97

doi:10.6342/NTU201601384

CHAPTER 1 Introduction

Recent advances in sequencing technology, such as RNA-seq, have opened a new era in RNA studies. Novel types of RNAs such as long non-coding RNAs (lncRNAs) have been discovered by transcriptomic sequencing and some lncRNAs have been found to play essential roles in biological processes such as development and diseases [1, 2]. More and more studies have discovered and investigated lncRNAs in many organisms such as human and mouse. However, only limited information is available for lncRNAs in *Drosophila melanogaster* (fruit fly), an important model organism.

Through considerable literature survey, most of lncRNA studies were found to be conducted in human or mouse, while only a few in *D. melanogaster*. For example, some lncRNAs have been observed to regulate developmental processes in *D. melanogaster*. Two genes, roX1 and roX2 recruit the MSL (male specific lethal) chromatin remodelling complex to genes on the male X chromosome, but not the autosomes or the female X chromosomes, to increase the acetylation of histone H4K16 [3]. This regulation can coordinate the dosage compensation required for male development. While the functionality of some lncRNAs in fruit fly was known, most of the lncRNAs have not yet been functionally characterized. The reason behind is probably owing to the fact that some of fundamental knowledge is currently scarce for fly lncRNAs.

doi:10.6342/NTU201601384

Therefore, characterization of lncRNAs in *D. melanogaster* is an important area of research.

To characterize fly lncRNAs, four essential questions would need to be clarified (Figure 1). First, it remains unclear whether the current set of fly lncRNAs is comprehensive. This question could be answered by collecting know lncRNAs to assess the current state. Nevertheless, discovery of novel lncRNAs from newly generated RNA-seq data also helps to infer whether additional lncRNAs could be found. Second, properties of fly lncRNAs are not well characterized because of incomplete annotation. Integrating multiple data sources from fly genomics and transcriptomics could improve annotations of lncRNAs. Third, the reliability of novel lncRNAs discovered from RNA-seq need to be assessed. Quantitative reverse transcription polymerase chain reaction (RT-qPCR) is the gold standard to validate the expression of the discovered

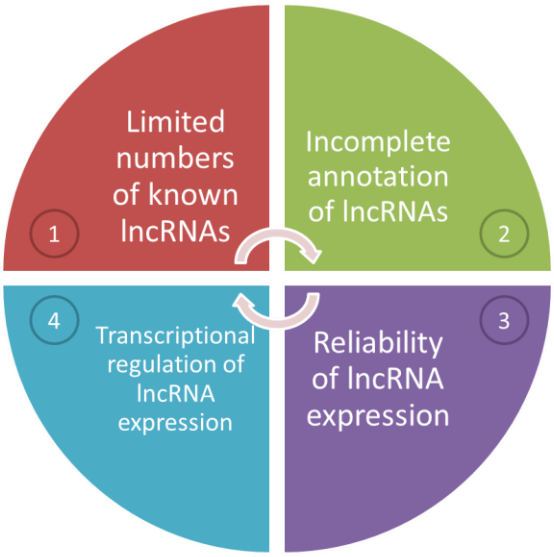

Limited numbers of known lncRNAs
1
Incomplete annotation of lncRNAs
2
4
Transcriptional regulation of lncRNA expression
3
Reliability of lncRNA expression

Figure 1. Four challenges for characterizing lncRNAs in *D. melanogaster*

doi:10.6342/NTU201601384

lncRNAs. Finally, it remains challenging to infer transcriptional regulation of lncRNA expression, because experimentally validated TF binding sites (TFBS; usually represented as sequence motifs) are currently scarce. In this regard, *in silico* predictions are needed for characterizing this issue in a large scale. In this thesis, we will discuss these four problems in detail and give an integrative approach to solve these problems by adopting multiple data sources from fly genomics and transcriptomics.

1.1 Challenges of lncRNA studies in *D. melanogaster*

Limited numbers of known lncRNAs in *D. melanogaster*

To assess the current state of lncRNA in the fruit fly, this thesis collected fruit fly lncRNAs from databases and literature and found that the number of known lncRNA genes in fruit fly (Table 1) was much smaller than those reported in human (~102,000) and mouse (~87,000) [4]. We suspect that the set of known lncRNAs in fruit fly is far from exhaustive. In this thesis, we first collect known lncRNA loci from databases and literature to establish an extensive list of annotated lncRNAs. Second, we produce two tissue-specific RNA-seq datasets from brain samples, respectively using the poly(A)-enriched and the ribo-zero method, and develop a computational pipeline to identify new lncRNAs from the two RNA-seq datasets.

doi:10.6342/NTU201601384

Table 1. Numbers of fly lncRNAs from different data sources

Type	Source	Number of fly lncRNA genes
Database	FlyBase (Release 6.06)	2,460
Database	UCSC genome browser	980
Literature	Young *et al*. (2012)	1,119
Literature	Brown *et al*. (2014)	1,875

Incomplete annotation of lncRNAs in *D. melanogaster*

The annotations of the collected lncRNAs are found to be incomplete. For example, Young *et al.* [5] reported 1,119 lincRNAs for *D. melanogaster* in 2012, but provided no detailed information because the RNA-sequencing reads were not generated with a strand-specific library construction [6]. In particular, transcriptional directions and exon regions are scarce for some of the previous published lncRNAs. Transcriptional direction is an important characteristic in lncRNAs. The transcripts of lncRNAs are able to disrupt the transcription of coding genes, a phenomenon known as convergent transcription in which the transcriptional direction of the lncRNA and the mRNA are head-to-head against each other [7, 8]. Conversely, for divergent transcription, the lncRNA/mRNA gene pair exhibit coordinated changes in transcription [9]. In this regard, the direction of lncRNA transcription is an important feature to be annotated. Another essential characteristic is the exon regions, which is important for most of subsequent biological experiments such as quantitative reverse transcription polymerase chain reaction (RT-qPCR). This thesis improves lncRNA annotation by integrating a

doi:10.6342/NTU201601384

large number of sequencing datasets (93 sets in total) from multiple sources (lncRNAs, RNA-seq and ChIP-seq). With these efforts, four general characteristics of lncRNAs are summarized in this thesis, including (1) genomic location distribution of lncRNAs, (2) length and structure of lncRNAs, (3) evolutionary conservation of lncRNAs, and (4) supporting evidences for lncRNA expression in the developmental stages.

Reliability of lncRNA expression detected from RNA-seq data

RNA-seq as a kind of high-throughput technology remains a possibility of certain bias and errors; for example, false lncRNAs detection might be caused by contaminated genomic DNA or unprocessed pre-mRNA during library construction. Recent studies have also revealed that the quantification results might be estimated differently by using different types of reads [10] or different bioinformatics/statistics methods [11, 12] . Therefore, it remains uncertain whether a lncRNA discovered from RNA-seq data is truly expressed. In this thesis, the reliability of lncRNA expression is assessed by adopting additional supporting evidences from genomics (ChIP-seq) or transcriptomics (RNA-seq) data, other data sources (such as coding potential predictors, and Conserved Domains database), and RT-qPCR validation.

Transcriptional regulation of lncRNA expression

While many studies have focused on annotating the function of lncRNAs, the

doi:10.6342/NTU201601384

knowledge about how the expression of lncRNAs is regulated is considerably limited. Only a few studies went upstream to ask how lncRNAs are regulated [13]. In fact, it is quite challenging to study this issue in a genome-wide level owing to the fact that transcription factor binding sites (TFBSs) with experimental validation are currently scarce (Table 2). In this regard, *in silico* predictions of TFBSs may be needed to investigate the regulation of lncRNA expression. This thesis incorporates *de novo* motif discovery to systemically investigate the presence of *cis*-elements shared by the promoters of coding and long non-coding (C-LNC) genes.

Table 2. Statistics of the public data used for studying transcriptional regulation in yeast, fruit fly and human

Data types	**Yeast**	**Fruit fly**	**Human**
Estimated number of TFs	312 TFs (~5% of all protein-coding genes; [14])	~750 TFs (~6%; [15])	~1850 TFs (~8%;[16])
Annotated PFMs[a]	307 matrices for 170 TFs [17-19]	815 matrices (~300 matrix clusters) [20-22]	~900 matrices [21]
Expression data	Cell cycle (Microarray; [23]) Environmental stresses (Microarray; [24])	Developmental (RNA-seq; [25]) (Microarray; [26, 27]) Early embryogenesis stage (Immuno-stained; [28])	Tissues / Disease stages
ChIP[b] experiments	350 ChIP-chips for 203 TFs [29]	93 ChIP-chip for 50 TFs [30, 31] 6 ChIP-seq for 2 TFs [30]	129 ChIP-chip [32] 16 ChIP-seq [33]

[a] PFM: position frequency matrix, which is utilized for representing the frequency of nucleotides (A, T, C and G) in a TF binding motif.

[b] ChIP: Chromatin Immunoprecipitation

doi:10.6342/NTU201601384

1.2 Integrative approach for characterizing lncRNAs by utilizing genomics and trnascriptomics data

To assess the current state of lncRNAs and their annotation in *D. melanogaster*, we collected known fly lncRNAs from databases and the literature, and then used strand-specific RNA-seq datasets (Table 3) to add to the characterization of the annotations. The collected lncRNAs contained approximately 3,300 genes. To investigate whether many more lncRNAs could be discovered, we obtained additional RNA-seq datasets from the brain (Table 3). We selected the brain, instead of the whole body, because many lncRNAs were tissue-specific according to lncRNA studies in mammals [34]. Also, the brain is important for studying neuron-related diseases. Since some lncRNAs may not contain poly(A) tails, both poly(A)-enriched and ribo-zero libraries were constructed in this thesis. For the purpose of discovering novel lncRNAs, we developed a reference-based assembly approach to identify potential lncRNA transcripts.

The next question addressed in this thesis is whether RNA-seq is a reliable platform for the discovery of novel lncRNAs. A previous study used chromatin immunoprecipitation sequencing (ChIP-seq) data of chromatin signatures to detect transcription of lncRNAs [35]. A lncRNA locus, similar to that of a protein coding gene, contains the promoter and gene body and associates with the active chromatin

doi:10.6342/NTU201601384

Table 3.   Summary statistics of datasets used in thesis.

Platforms	Types	Total number of datasets	Experimental condition	Number of datasets
Public RNA-seq (59 datasets in total)	Paired-end without strand-specific	30	Time course / whole body	30
	Paired-end with strand-specific	29	Tissue / head	9
			Tissue / ovary	2
			Tissue / accessory glands	1
			Tissue / testis	1
			Tissue / carcass	4
			Tissue / digestive system	4
			Tissue / CNS	2
			Tissue / fat body	3
			Tissue / imaginal discs	1
			Tissue / salivary glands	2
In-house RNA-seq (2 in total)	Paired-end with poly(A)-enriched	1	Tissue / brain	1
	Paired-end with ribo-zero	1	Tissue / brain	1
ChIP-seq (32 in total)	H3K36me3	3	Embryos	1
			Larvae	1
			Mixed Adult	1
	H3K4me3	14	Embryos	7
			Larvae	3
			Pupae	1
			Adult Female	1
			Adult Male	1
			Mixed Adult	1
	RNA polymerase II	15	Embryos	8
			Larvae	5
			Pupae	1
			Mixed Adult	1

Detailed information of these datasets can be seen in Additional File 3: Table S2 of the published work [36] and Appendix Table 1 in this thesis.

doi:10.6342/NTU201601384

signatures such as H3K4me3 and H3K36me3 [37-39]. It is also known that lncRNA expression also requires specific binding of transcription factors to promote RNA polymerase II (Pol II)-mediated transcription [40-42]. In combination with the information of expression profiles and these three chromatin signatures which are believed to be present in the actively transcribed regions, an lncRNA with these three chromatin signatures would be considered to be transcribed with higher confidence. As for a lncRNA discovered from a specific tissue sample, three more analyses could be conducted to investigate the reliability. For one of this kind of lncRNAs, it could be examined whether (1) it was observed to be expressed in the RNA-seq datasets from developmental stages; (2) it was predicted with a low coding probability by two or more predictors; and (3) it was not predicted to contain any conserved domains of proteins. Last but not least, RT-qPCR validation is the gold standard to assess a lncRNA transcribed or not.

While we integrated multiple sets of RNA-seq and ChIP-seq data (Table 3) to investigate transcription of lncRNAs during the development of *D. melanogaster*, we observed that a large proportion of genomic regions for lncRNAs expressed in RNA-seq were not occupied by chromatin signatures (H3K4me3, H3K36me3 and Pol II) that are usually associated with active transcription. However, no studies have discussed which feature (chromatin signatures or expression intensities) is better for inferring the

doi:10.6342/NTU201601384

existence of lncRNAs. To answer this question, we designed RT-qPCR experiments to evaluate the confidence level of lncRNAs discovered from RNA-seq.

Additionally, to investigate transcriptional regulation of lncRNA expression, this thesis incorporated *de novo* motif discovery to systemically investigate the presence of cis-elements shared by the promoters of coding and long non-coding (C-LNC) genes. For this purpose, the time-course RNA-seq data set of 30 developmental stages of D. melanogaster (Table 3) was adopted. Co-expressed C-LNC gene clusters were constructed by applying hierarchical clustering on the expression profiles of fly mRNAs and the compiled lncRNAs in this thesis. To identify potential regulatory elements, *de novo* motif discovery was conducted on the promoters of coding genes in a cluster. Then, the discovered motifs were examined to see whether they are also present in the promoters of LNC genes in the same cluster. The discovered motifs were also used to identify potential common regulators of these C-LNC genes.

In summary, this thesis aims to demonstrate that ambitious integration of sequencing data followed by computational procedures can largely facilitate novel lncRNA discovery as well as enhance lncRNA annotation and characterization.

1.3 Thesis structure

This thesis address the above challenges from the four corresponding aspects as showed in Figure 2: (1) collection and discovery; (2) annotation; (3) expression; and (4)

doi:10.6342/NTU201601384

regulation. CHAPTER 2 provides literature reviews for the related works about lncRNA studies and the current status in *Drosophila melanogaster*. CHAPTER 3 presents the collection and discovery of fruit fly lncRNAs. A computational approach is developed for identifying novel lncRNAs from the generated RNA-seq data with two types of library constructions. CHAPTER 4 is then focused on improving the annotations of the published and the newly discovered fly lncRNAs. Several general properties are characterized for the curated fly lncRNAs. Next, the reliability of the lncRNA expression was investigated and validated by RT-qPCR in CHAPTER 5. Then, CHAPTER 6 moves to the upstream of the lncRNA expression. A novel method incorporating motif discovery is proposed for systematically investigating the potential *cis*-elements and how it affects lncRNA expression. CHAPTER 7 discusses the limitations of this work. The conclusion and future work are given in CHAPTER 8.

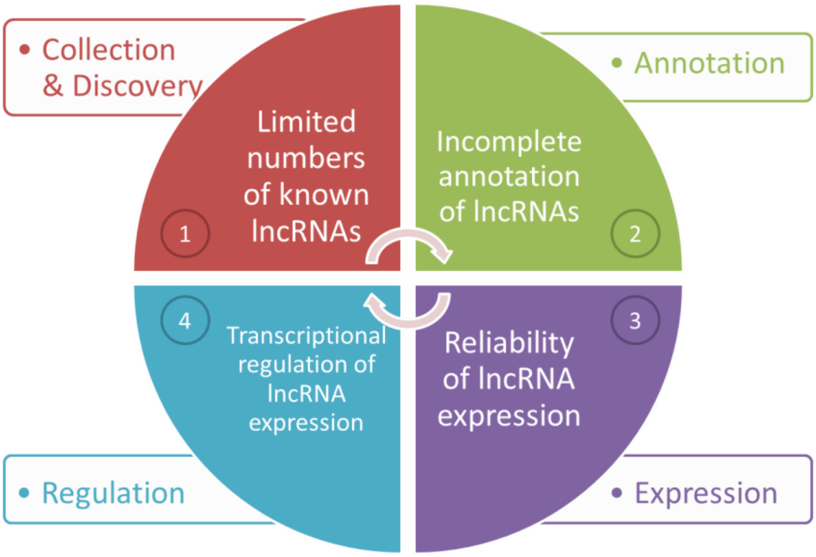

• Collection & Discovery
Limited numbers of known lncRNAs
1
• Annotation
Incomplete annotation of lncRNAs
2
4
Transcriptional regulation of lncRNA expression
3
Reliability of lncRNA expression
• Regulation
• Expression

Figure 2. Characterizing lncRNAs in *D. melanogaster*

doi:10.6342/NTU201601384

CHAPTER 2 Related Works

2.1 Brief history of long non-coding RNAs studies

Okazaki *et al*. (2002) investigated the mouse transcriptome by using 60,770 cDNAs, and found that around two third of mouse transcriptome was consisted by non-coding RNAs (ncRNAs) [43]. At the time, ncRNAs was comprehended as transcriptional noises. The fact of that ncRNAs is the major component of the transcriptome brought the attention of researchers to these geek transcripts. In 2004, Cawley *et al*. found that a great proportion of ncRNAs have transcription factor binding sites (TFBSs) in their promoters by an unbiased mapping of human TFBSs on chromosome 21 and 22 [44]. This study revealed the potential for ncRNAs to be transcriptionally regulated. This idea was relayed by Ravasi *et al*. In 2006, they provided experimental validation for the expression of several ncRNAs in mouse, and demonstrated that transcription of ncRNAs is the real event [45]. Thus, the view of transcriptional noises on ncRNAs was completely overthrown. In the next ten years, ncRNAs, including short ncRNAs (such as miRNAs) and long ncRNAs (lncRNAs), became the hot spots in RNA research. A RNA sequence is classified as a lncRNA if it lacks coding potential and has a length >200 base pairs (bp) [46]. The functional roles of lncRNAs have been investigated in several studies [47-50]. A review paper reported that lncRNAs serve as regulators of

doi:10.6342/NTU201601384

diverse cellular functions such as epigenetic silencing or transcriptional regulation [48]. Moreover, the advance of sequencing technology has facilitated the accumulation of a large amount of data. Thus, developing systematic approaches for integrating and interpreting these data is essential for the current academia research.

2.2 Integrative and systemic studies on lncRNAs

In the state-of-art of lncRNAs studies, several integrative and systemic studies have been conducted for the investigation of lncRNAs. These studies could be roughly categorized into four types: (1) LncRNA identification [51, 52]; (2) RNA-protein interactions [53, 54]; (3) LncRNA function identification [55]; and (4) Transcriptional regulation of lncRNA expression [13]. However, most lncRNAs studies were for mammalian species such as human and mouse. The accumulated information about *Drosophila melanogaster* lncRNAs is lacking when compared with mammalian organisms. Besides, over the past years, most studies have focused on investigating lncRNA functions [47-50], but few studies went upstream to ask how lncRNAs are regulated [13].

2.2.1 Related works for characterizing lncRNAs in in *Drosophila melanogaster*

Many studies have developed bioinformatics methods to systematically identify and characterize lncRNAs [51, 52] in human or mouse. However, in *D. melanogaster*, the

doi:10.6342/NTU201601384

related works are only a few. Young *et al*. (2012) [5] was the first work which systematically identified a large amounts of lncRNAs from RNA-seq data in fruit fly. But due to the RNA-seq datasets that were not constructed by strand-specific library, only limited annotations and characteristics of fly lncRNAs could be provided in their study. In 2014, Brown *et al*. [56] incorporated RNA-seq data from 10 types of tissues to study all types of transcripts in fly transcriptome, which also included lncRNAs. This study provided some interesting insights of lncRNAs, but failed to comprehensively discuss characteristics of fly lncRNAs. In this thesis, we integrated the information provided by the above two studies, compensated the scarce information of them, and thus presented the most up-to-date list of fly lncRNAs with comprehensive annotations. .

2.2.2 Related works for transcriptional regulation of lncRNA expression

Studies on mouse and human have reported that lncRNA genes are similar to protein coding genes in that they contain promoters and transcribed regions [44]. Upon transcription, these regions will have active chromatin signatures such as the tri-methylation of histone H3 lysine 4 (H3K4me3) and the tri-methylation of histone H3 lysine 36 (H3K36me3) [38, 39, 57]. It has also been revealed that lncRNA expression may require specific binding of transcription factors to drive RNA polymerase II (Pol II)-mediated transcription [40-42]. Wu *et al.* (2010) found that the expression of

doi:10.6342/NTU201601384

lncRNAs was regulated though EzH2-mediated H3K27 methylation on embryonic stem cells, which is known as a similar way to the regulation of protein coding genes [58]. In plant, it has been demonstrated that the expression of the lncRNA, COOLAIR, was inhibited by covered COOLAIR promoter with AtNDX aim to form R-loop in *Arabidopsis* [59]. Moreover, Yang *et al.* (2013) showed that histone acetylation-mediated modulation of the promoter region could suppress lncRNA, and cause low expression in tumor (lncRNA-LET) [60]. The above-mentioned studies have provided a firm support to that lncRNA expression is associated with the molecular modification of its promoted region.

To fully understand the function of lncRNA, the key driver of lncRNA expression may be also essential but less study systematically investigated this issue. For example, Yang *et al.* [13] constructed the ChIPBase database providing a user-friendly interface for users to browse transcription factor (TF) binding sites from ChIP-seq experiments in the regulatory region of a lncRNA. Though, the information provided by the ChIPBase included all of the peaks across different cell lines or tissues without telling from which experimental condition a TFBS is. Therefore, users cannot obtain specific TFBS information in a specific experimental condition. This inspired Jiang *et al.* [61] developed a web-based tool, TF2LncRNA, to enables users to obtain the specific information of TFs, TFBSs, and the experimental conditions. However, both of the two

doi:10.6342/NTU201601384

studies highly replied on ChIP experiments, where only limited number of ChIP datasets for TFs is available in *D. melanogaster*. To be more specific, only ~100 ChIP experiments for ~50 TFs are available currently, the number of which is far less than the estimated number of TFs (as showed in Table 2). An alternative approach is to adopt *de novo* motif discovery on the promoters of co-expressed genes for investigate transcriptional regulation. This approach may be easily frustrated by the fact that the number of co-expressed lncRNAs is usually limited. In this regards, this thesis proposed a procedure that performing *de no* motif discovery only on coding gene promoters in a co-expressed gene cluster, and then used the discovered motifs to identify regulatory elements in the co-expressed lncRNA promoters.

doi:10.6342/NTU201601384

CHAPTER 3 Collection and Discovery of lncRNAs in *Drosophila melanogaster*

In this thesis, we compiled an the most update list of fruit fly lncRNAs from databases and literature and found that the number of known lncRNA genes in fruit fly (~3,300) was much smaller than those reported in human (~102,000) and mouse (~87,000) [4]. We suspected that the set of known lncRNAs in fruit fly was far from exhaustive. Indeed, 462 novel lncRNA genes were discovered when two brain-specific RNA-seq datasets were produced in the present study. Thus, more lncRNA genes will likely be found when more RNA-seq studies of fruit fly are conducted in the future. The final set of curated fly lncRNAs, including known and novel lncRNAs, contains 3,816 lncRNA genes (4,599 lncRNA transcripts).

3.1 Known lncRNAs collected from databases and literatures

A non-redundant set of 1,999 lncRNA genes (2,347 transcripts) from FlyBase (r5.57) [62] and the UCSC genome browser [63] was first constructed. Next, the long intergenic non-coding RNAs (lincRNAs) reported in the study by Young *et al*. [5] and Brown *et al*. [56] were collected to expand the list. Among the 1,119 lincRNAs reported by Young *et al*. and the 3,088 lncRNAs by Brown *et al*., some potentially redundant

doi:10.6342/NTU201601384

lincRNAs or lncRNAs were excluded by a selection procedure (see the section of 3.4.1). In the end, 583 lincRNA genes (583 transcripts) from Young *et al*. and 772 lncRNA genes (1,207 transcripts) form Brown *et al*. were added to the non-redundant set reported in the present study.

3.2 Novel lncRNAs identified from brain samples

We developed an approach to discover lncRNAs from the brain-specific RNA-seq datasets of fruit fly produced in this thesis (SRP051132), which were obtained using two types of library construction, the poly(A)-enriched and ribo-zero protocols. This approach can be applied to future studies for the same purpose. The proposed pipeline consists of several steps, including reference-based assembly (using an earlier version of gene annotations downloaded from UCSC genome browser on March 13th, 2013), coding potential estimation, ribosomal RNA exclusion, and read remapping (see the section of 3.4.3). The results consisted of 754 intergenic transcripts that have not been previously annotated. After excluding transcripts with lengths less than 200 bp, 725 transcripts remained as putative lncRNAs. Then, we retained 591 putative lncRNA genes which showed a low potential to encode proteins. After excluding ribosomal RNA contamination, 587 putative lncRNA transcripts remained. We further excluded 57 transcripts that had no sufficient read support during the follow-up read remapping. Before finalizing the list, we compared the discovered lncRNAs with the most updated gene annotations from UCSC genome browser (Sep. 21st, 2015), and removed 68 transcripts that overlapped some newly reported coding genes in the sense direction. Finally, we obtained 462 novel lncRNA transcripts that have not been reported previously. RT-qPCR experiments were conducted for validation. The results showed that all of the selected novel lncRNAs were validated, which revealed the high reliability of the discovered novel lncRNA genes (details in CHAPTER 5).

3.3 Up-to-date list of long non-coding RNAs in *D. melanogaster*

In total, a set of 3,816 curated lncRNA genes (4,599 transcripts) in *D. melanogaster* was constructed in this thesis (Additional File 1 and Additional File 2 of the published work [36]). The final set of curated fly lncRNAs is larger than the 2,460 lncRNA genes in FlyBase (Release 6.06 [62]), and the 2,446 lncRNA transcripts recently reported by Matthews *et al*. [64]. Our final list is also larger than the latest version (version 4) of a well-known lncRNA database, NonCode (961 lncRNA genes) [65].

3.4 Methods for collection and discovery of fruit fly lncRNAs

3.4.1 Collection of published lncRNAs

The lncRNAs were collected from FlyBase [62], the UCSC genome browser [63], Young *et al.* [5], and Brown *et al*. [56]. A set of lncRNAs was obtained using the keyword term “non_protein_coding_genes” when querying FlyBase *D. melanogaster*

doi:10.6342/NTU201601384

(r5.57). LncRNA transcripts shorter than 200 bp were filtered out. First, the lncRNA transcripts from FlyBase were chosen as the primary set of lncRNA sequences. Second, BLASTn [66] was used to align the lncRNA transcripts collected from the UCSC genome browser against the primary set. Afterwards, by checking the alignments with E-value < 10^{-10} in the BLASTn results, redundant lncRNA transcripts were removed when either of the following two conditions was satisfied: (1) a lncRNA has the same loci with another lncRNA, or (2) a lncRNA overlaps another lncRNA with an overlapping region covering 50% of the transcript length. With the specified criteria, 972 redundant sequences were excluded. Third, 1,119 lincRNAs were collected from the study by Young *et al.* [5], where 415 sequences were excluded because they contained overlapping regions with the non-redundant set of lncRNA transcripts from FlyBase and the UCSC genome browser. Additionally, 3,088 lncRNA transcripts were collected from Supplementary Data 2 of the study of Brown *et al.* [56]. We removed 49 lncRNA transcripts with a length < 200 bp and 19 transcripts that were annotated as coding genes in the file provided by Brown *et al*. The remaining 3,020 lncRNA transcripts were next aligned to the above non-redundant set of lncRNA transcripts from FlyBase, UCSC, and Young *et al*. by using BLASTn. The alignments with E-value < 10^{-10} in the BLASTn results were further examined by the following selection procedure. We removed lncRNA transcripts that were annotated with an already included FlyBase

doi:10.6342/NTU201601384

lncRNA ID. LncRNA transcripts containing overlapping regions with the curated FlyBase/UCSC lncRNA transcripts (covering > 50% of the either transcript length) were removed unless the new lncRNA transcripts contain multiple exons and the number of exons differs from that of FlyBase/UCSC lncRNA transcripts. Afterwards, lncRNA transcripts aligned to lncRNA transcripts of Young *et al*. were removed only if they have the same loci or have an overlapping region covering 90% of transcript length. As a result, 1,635 redundant lncRNA transcripts were removed. All lncRNA transcripts were then aligned to 156 ribosomal RNAs collected from FlyBase r6.07 (2 sequences) and the NCBI database (154 sequences) using BLASTn. Sequences (10 sequences) with E-value < 10^{-10} and identity > 99% were removed to exclude ribosomal RNA contamination.

To ensure that the lncRNAs curated in this thesis did not contain newly reported coding genes present in the most updated FlyBase annotations, we retrieved ‘Feature Type’ and ‘Gene Model Status’ for the curated lncRNA transcripts from FlyBase by submitting transcript IDs to the batch download tool of FlyBase r6.07. Additionally, we utilized ‘Coordinates Converter’ provided by FlyBase to see whether a transcript location is no longer present in the release 6 genome (R6). Moreover, for the lncRNA transcripts from Young *et al*., FlyBase recently incorporated these lncRNA transcripts and provided update annotations based on a manual review (FBrf0220965). By taking

doi:10.6342/NTU201601384

the above-mentioned information from FlyBase into account, we removed 673 transcripts that were annotated as protein coding genes, pseudogenes, rRNA genes, snRNA, snoRNA, scaRNA, out-of-date IDs, or located within TE regions or the sequences dropped by the BDGP in the R6 genome. In the end, this thesis constructed a set of lncRNAs from FlyBase, the UCSC genome browser, and the studies by Young *et al.* [5] and Brown *et al*. [56], consisting of 3,354 lncRNA genes, corresponding to 4,137 lncRNA transcripts.

3.4.2 RNA-seq data of the fly brain

Brain samples were collected from four-day post-eclosion *Canton S* male adults. At a time, 20 to 30 adults were gassed with carbon dioxide and dissected. The collected brains were preserved in refrigerator until 100 brains were collected. Afterwards, total RNA was purified from the 100 brains, using the NucleoSpin® RNA II Purification Kit. RNA-seq was performed using the strand-specific library with poly(A)-enriched protocol or Ribo-Zero™ Gold Kit to generate paired-end 90-bp reads on the Illumina Hi-seq 2000 platform. In total, ~25 million and ~50 million pair-end reads of 90-bp in length were obtained from the poly(A)-enriched library and the total RNA (with Ribo-Zero™ Gold Kit) library, respectively. The raw reads have been submitted to NCBI Sequence Read Archive database (SRP051132).

doi:10.6342/NTU201601384

3.4.3 Novel lncRNA discovery

To discover novel lncRNAs from the two new datasets described above, we first mapped all short reads onto the unmasked *D. melanogaster* genome sequences (BDGP R5/dm3; from the UCSC genome browser), using TopHat [67]. Cufflinks [67] was then used to assemble the mapped reads and the assembled transcripts were compared to the reference annotation (Dmel refseq) from the UCSC genome browser (downloaded on March 13th, 2013) using Cuffcompare, a utility included in Cufflinks. The two sets of assembled transcripts, from poly(A)-enriched RNA and total RNA, respectively, were compared to the reference annotation at the same time to get a union set of intergenic transcripts. We set a length of 200 bp as the cutoff to exclude shorter non-coding RNAs. We then calculated the coding potential of all putative lncRNA loci using the Coding Potential Calculator (CPC) [68]. The putative lncRNA transcripts were then aligned against a set of ribosomal RNAs (the same set described in the "Collection of published lncRNAs" section) to exclude ribosomal RNA contamination. Afterwards, we remapped both poly(A)-enriched RNA and total RNA sequencing reads to the putative lncRNA transcripts, using Cufflinks. After remapping, we excluded transcripts with no read support as reported by Cufflinks. The developed computational pipeline is shown in Figure 3. Then, we compared the identified lncRNAs with the most updated R5 genome annotations downloaded from the UCSC genome browser (Sep. 21st, 2015), and

doi:10.6342/NTU201601384

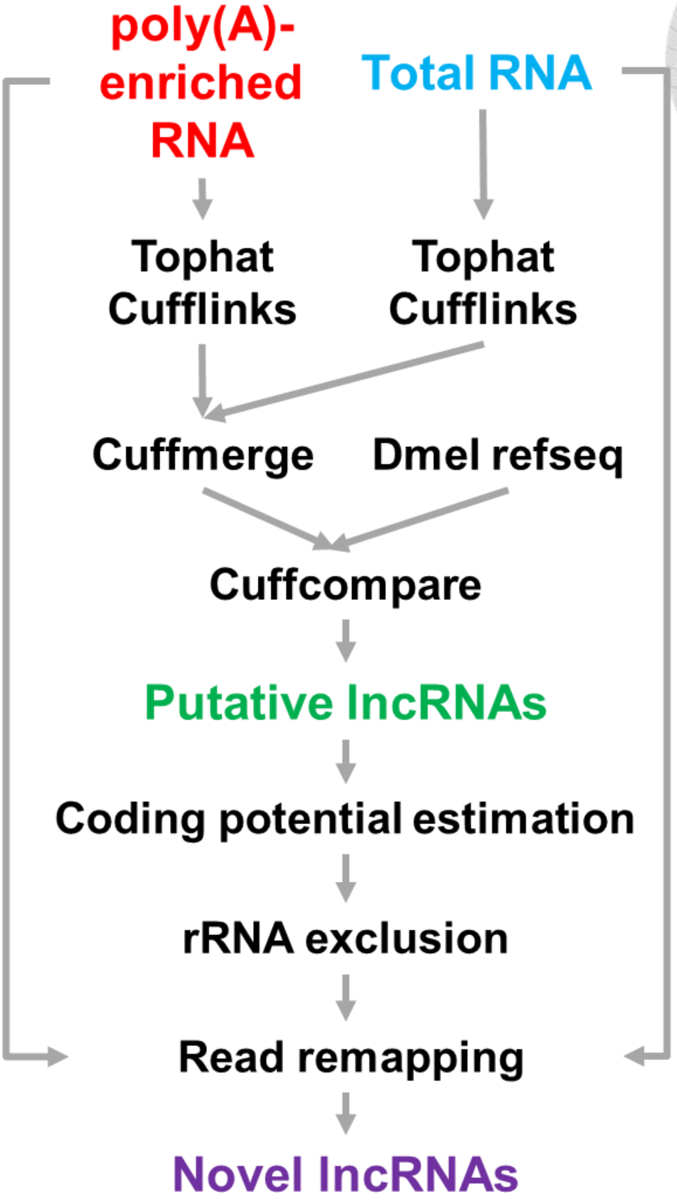


Figure 3. Procedures for discovering novel lncRNAs from RNA-seq data of the present study. The sequencing read datasets of mRNA and total RNA were respectively mapped to the reference genome sequence using TopHat and Cufflinks. Putative lncRNAs were then discovered by Cuffcompare. Sequencing reads were again mapped to the set of putative lncRNAs to construct the final set of novel lncRNAs.

removed lncRNA transcripts that overlapped with some newly reported coding genes in a sense direction. The resulting set of putative lncRNA transcripts were then compared

doi:10.6342/NTU201601384

to the set of non-redundant lncRNA transcripts collected from FlyBase, the UCSC genome browser, and the studies by Young *et al*. [5] and Brown *et al*. [56] to remove redundant sequences.

doi:10.6342/NTU201601384

CHAPTER 4 Annotation of the curated lncRNAs

This thesis showed that integrating multiple public datasets can largely facilitate the annotations and characterization of fly lncRNAs. A great amount of sequencing datasets, including 59 RNA-seq datasets and 32 ChIP-seq datasets collected from the modENCODE database, were used for improving the annotation. Next, according to the improved annotations, we observed four general characteristics of fruit fly lncRNAs and discussed these characteristics in this chapter.

4.1 Improving the annotation of the lncRNAs from Young *et al.*

Young *et al.* [5] reported 1,119 lincRNAs for *D. melanogaster* in 2012, but provided no detailed information because the RNA-sequencing reads were not generated with a strand-specific library construction [6]. In this thesis, we collected the original 30 RNA-seq datasets [6] used by Young *et al*. (Table 3and modENCODE IDs: 4433-4462 as shown in Additional File 3: Table S2 of the published work [36]) and adopted 29 additional stranded poly(A)-enriched RNA-seq datasets at different developmental stages (Table 3 and modENCODE IDs: 4291-4319 as shown in Additional File 3: Table S2 of the published work [36]) to determine the exon regions and transcriptional directions for the lincRNAs reported in Young *et al.*'s study. After excluding redundant lincRNAs against the annotated lncRNAs from the databases and removed transcripts

doi:10.6342/NTU201601384

which are no longer lincRNAs in the current FlyBase annotations (FBrf0220965), 583 lincRNA genes remained. To identify the exon regions of these 583 lincRNA genes, we remapped the 30 RNA-seq datasets to the lincRNA sequences using Cufflinks [67]. We found that most of lincRNA genes from Young *et al*. consisted of only one or very few exons (Table 4 and Additional File 4 of the published work [36]). As for transcriptional

Table 4. Statistics of exon numbers in lncRNA and mRNA genes from different sources.

Exon num.	FlyBase + UCSC	Young et al.	Brown et al.	Present study	mRNA
1	1167	444	465	422	2751
2	495	93	163	33	4739
3	196	32	60	6	4109
4	68	12	35	1	3659
5	36	1	15	0	2863
6	17	0	8	0	2268
7	8	0	7	0	2003
8	2	1	7	0	1586
9	3	0	2	0	1281
10	1	0	2	0	995
11	1	0	5	0	781
12	3	0	0	0	612
13	0	0	0	0	471
14	0	0	0	0	391
15	0	0	1	0	331
16	0	0	0	0	240
17	0	0	0	0	200
18	1	0	0	0	145
>=19	1	0	2	0	837
Total	1999	583	772	462	30262

Table 5. Statistics of transcriptional direction in the lncRNA genes from different sources.

Transcriptional direction	FlyBase + UCSC	Young et al.	Brown et al.	Present study	mRNA
Positive (+)	1011	200	392	268	14,941
Negative (-)	988	192	380	194	15,321
Unknown (*)	0	191	0	0	0
Total	1999	583	772	462	30262

direction, similar procedures were conducted. We annotated the direction of transcription in about 67% of the 583 lincRNA genes from the study by Young *et al.* (Table 5). To be more specific, 200 lincRNA genes were identified on the positive strand and 192 on the negative strand of the fruit fly genome (Table 5 and Additional File 2 of the published work [36]).

4.2 Utilizing additional RNA-seq datasets to improve the annotation of the 4,599 curated lncRNA transcripts

We utilized the RNA-seq datasets from multiple sources as well as those generated in this thesis to improve the annotation of the curated lncRNAs. Three properties were emphasized here: (1) the classification of a lncRNA in terms of its genome location and transcriptional direction; (2) whether the lncRNA is expressed in the brain or not; and (3) whether the lncRNA has a poly(A) tail or not.

The lncRNAs collected in the present study were classified into several groups according to their genome locations with respect to the closest adjacent coding gene.

doi:10.6342/NTU201601384

Table 6. Types of lncRNA transcripts.

Types	Number of lncRNAs	Averaged length (±sd)	Number of exons (counts of lncRNAs)	Transcriptional direction (counts of lncRNAs)
Intergenic	**2602**	**1002 (±1305.81)**	**Single (1805); multiple (797)**	**+(1375); −(1227)**
Exonic				
Anti-sense	832	1161 (±1059.20)	single (373); multiple (459)	+(448); −(384)
Sense	268	1380 (±1317.87)	single (154); multiple (114)	+(131); −(137)
Total	**1100**			
Intronic				
Anti-sense	495	770 (±581.83)	single (292); multiple (203)	+(239); −(256)
Sense	211	733 (±633.81)	single (149); multiple (62)	+(108); −(103)
Total	**706**			
Unknown	**191**	**813 (±782.66)**	**Single (164); multiple (27)**	**NA**
Total	**4599**			

+: positive strand.

−: negative strand.

NA: not available.

For lncRNAs located in regions that overlap with coding genes, the transcriptional direction was also considered to be an essential aspect for classification. In this regard, lncRNAs are classified into anti-sense exonic, sense exonic, anti-sense intronic and sense intronic lncRNAs, according to the transcriptional direction with respect to the overlapping coding gene. Among the curated 4,599 lncRNA transcripts, 2,602 were classified as intergenic lncRNA transcripts, 1,100 as exonic lncRNA transcripts (Table 6 and Additional File 2 of the published work [36]) and 706 as intronic lncRNA transcripts. There were 191 lncRNA transcripts for which the transcriptional direction could not be determined and were classified as ‘unknown’.

Additionally, this thesis provided two sets of sequencing reads of RNA samples from the brain (Table 3). With the two datasets, we could infer which lncRNAs were expressed in the brain. If the criterion ‘RPKM > 1’ was used, the data revealed that about one third of lncRNAs (1,464 transcripts, Additional File 2 of the published work [36]) were expressed in the brain. In Figure 13(b) we showed the RT-qPCR experiments of seven lncRNA genes with RPKM > 1 and three lncRNA genes with RPKM = 0. The RT-qPCR results showed that the –delta Ct values of the seven lncRNA genes with ‘RPKM > 1’ were distinguishable from the three lncRNA genes with ‘RPKM = 0’. In this regard, ‘RPKM > 1’ is considered as a safe criterion to infer the expression of lncRNAs in the brain. In summary, we found that 33% of the 3,816 lncRNA genes were expressed in the brain, when the criterion ‘RPKM > 1’ was used (Additional File 2 of the published work [36]). This number is considerably higher than that observed in other tissues reported by Brown *et al.* [56]. The study of Brown *et al.* incorporated RNA-seq data from 10 types of tissues and the testis tissue showed the highest number of expressed lncRNA genes (~30% of the 1,875 lncRNA genes).

We further examined whether a lncRNA contains the poly(A) tail. Both poly(A)-enriched and ribo-zero library constructions were used in the present study because some lncRNAs were previously found to contain no poly(A) tails in mammals [69-71]. Among the 1,464 lncRNA transcripts observed in the brain RNA-seq data,

doi:10.6342/NTU201601384

there were 190 lncRNA transcripts with a high probability of not containing poly(A) tails when expressed in the brain (Additional File 2 of the published work [36]).

4.3 General characteristics of the fruit fly lncRNAs

To understand the general characteristics of lncRNAs, we further processed the improved annotation in the previous sections, and characterized lncRNAs from four aspects, including (1) Location distribution of lncRNAs in Genome; (2) Length and structure of lncRNAs; (3) Evolutionary conservation of lncRNAs; and (4) Supporting evidences for lncRNA expression in the developmental stages.

4.3.1 Location distribution of lncRNAs in Genome

The numbers of lncRNAs from the three different sources are shown in Table 7 which indicated that lncRNAs are everywhere in the genome. In general, the euchromosome acquired more lncRNAs than the heterochromosome. Among the curated 4,599 lncRNA transcripts, 2,602 were classified as intergenic lncRNA transcripts, 1,100 as exonic lncRNA transcripts (Table 6 and Additional File 2 of the published work [36]) and 706 as intronic lncRNA transcripts. Table 6 shows that the number of lncRNAs for the four groups decreased as follows: anti-sense exonic lncRNAs > anti-sense intronic lncRNAs > sense exonic lncRNAs > sense intronic lncRNAs. The lncRNA numbers of the four groups in the different euchromatin regions were also provided (Figure 4). Here, we only considered lncRNAs located in euchromatin because most lncRNAs were

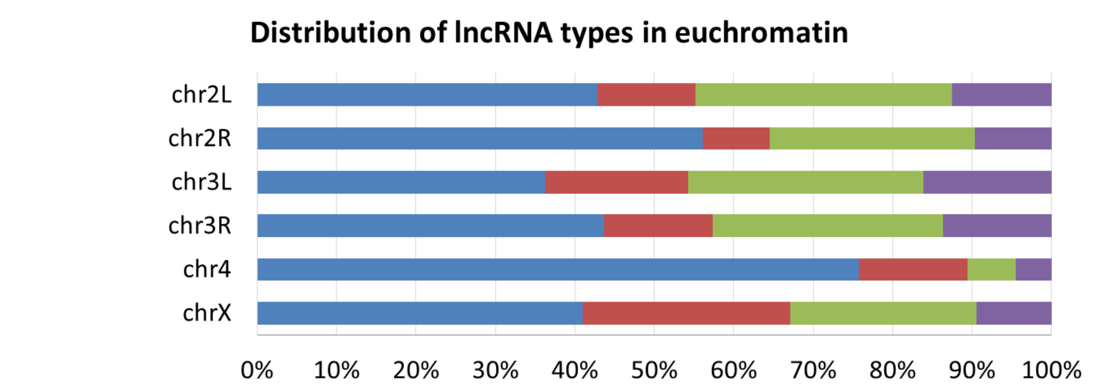
doi:10.6342/NTU201601384

Table 7. The number of lncRNAs from three different sources in each of the euchromosomes and heterochromosomes

Chromosome	**FlyBase + UCSC**	**Young *et al*.**	**Brown *et al*.**	**Present study**	**Summary**
chr2L	564	109	135	73	881
chr2LHet	1	0	2	0	3
chr2R	353	87	97	67	604
chr2RHet	4	0	9	14	27
chr3L	378	171	188	129	866
chr3LHet	4	0	4	24	32
chr3R	368	147	200	86	801
chr3RHet	2	0	7	5	14
chr4	23	4	10	14	51
chrU	30	0	27	17	74
chrX	271	65	92	33	461
chrXHet	1	0	1	0	2
total	1999	583	772	462	3816

expressed from the euchromatin in fruit fly.

However, in the curated list, we observed that there are some lncRNA transcripts from different sources partially sharing common genomic regions. These lncRNA transcripts might be in fact the same lncRNA, might be different splicing forms of a single lncRNA gene, or might be actually independent lncRNA genes. We realized that it remained difficult to learn the fact and determine the exact boundaries for these putative lncRNAs based on the limited information collected so far. Before a mature methodology can be developed, manual examination on RNA-seq data in a genome



	chrX	chr4	chr3R	chr3L	chr2R	chr2L
🟦 anti-sense exonic	121	50	163	115	185	150
🟥 sense exonic	77	9	51	57	28	43
🟩 anti-sense intronic	69	4	108	94	85	113
🟪 sense intronic	28	3	51	51	32	44

Figure 4. Distribution of lncRNA types in euchromatin.

browser is highly recommended. We highlighted the overlap information in Additional File 2 of the published work [36] to remind the readers that more investigations on such lncRNAs are needed. In addition, we also observed that the types of lncRNA transcripts (exonic, intronic, or intergenic lncRNAs) would potentially be changed once the annotation of protein-coding genes is updated. As the loci and boundaries of protein-coding genes continue to be refined, noncoding RNAs originally classified as intergenic may be found to be exonic, intronic or even become a new splicing form of a coding gene. In addition, Some of the Young *et al*. lincRNAs have been found by a follow-up FlyBase analysis (FBrf0220965) to overlap UTRs and are probably not lncRNAs. Therefore, the readers should be aware that the number of exonic sense lncRNAs in the curated list might be inflated by these lncRNAs.

4.3.2 Length and structure of lncRNAs

Transcriptional length of lncRNAs

The average length of the curated lncRNA transcripts is 1,008 bp with a diverse range and which is shorter than the average length of mRNAs (2,869 bp). More than 97% of the lncRNA transcripts have lengths from 200 bp to 4,000 bp (Table 8) which are consistent to the numbers reported by Novikova et al. [72].

Transcriptional direction of lncRNAs

When comparing lncRNAs with fruit fly mRNAs, we found that about half of the curated lncRNA genes were transcribed in the positive strands and half in the negative strands (Table 5). For each specific group of the lncRNA transcripts in Table 6 (the classification of a lncRNA in terms of its genome location and transcriptional direction), the lncRNA transcripts were equally derived from both strands. Moreover, 988 lncRNA genes (25.89% among the 3,816 lncRNA genes) were found to be transcribed in a

Table 8. Length of lncRNA transcripts

Range	200~500	500~1000	1000~2000	2000~4000	4000~up	Total
FlyBase + UCSC	707	997	463	131	49	2347
Young *et al*.	189	179	130	60	25	583
Brown *et al*.	390	443	240	104	30	1207
Present study	130	214	93	23	2	462
Total	1416	1833	926	318	106	4599

direction antisense to protein coding genes. This number is larger than that (15%) reported in human [73].

Exons of lncRNAs

As for the number of exons in lncRNAs, fruit fly lncRNAs tend to have fewer exons than mRNAs (Table 4), which is consistent with the observation in rat by Wang *et al.* [74]. Figure 5 showed that ~60% of mRNAs contain no more than five exons. The percentage of mRNAs with different exon numbers were roughly equally distributed (9% for one exon, 16% for two exons, 14% for three exons, 12% for four exons and 9% for five exons). In contrast, ~94% of lncRNAs contain one to three exons, and more than half of the lncRNAs contain only single exon. The exon numbers of lncRNAs were

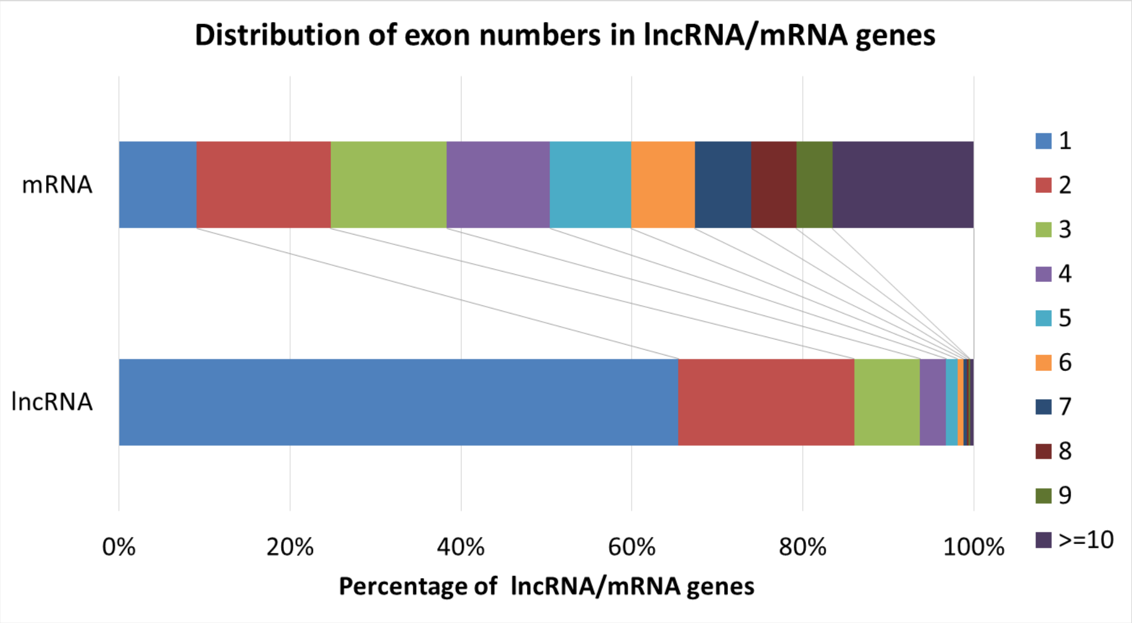


Figure 5. Distribution of exon numbers in lncRNA and mRNA genes.

doi:10.6342/NTU201601384

apparently smaller than that of mRNAs. It is not clear whether this was because the average length of the curated lncRNAs (1,008 bp) is shorter than that of mRNAs (2,869 bp). Additionally, in Table 6, we showed that intergenic lncRNAs were the major type of lncRNAs that contained only one exon.

Next, we utilized the peak detection results of 34 CAGE datasets from the study of Brown *et al.* to investigate the 5' end completeness of the curated lncRNA transcripts. The result showed that about ~55% of the curated lncRNA transcripts can find a CAGE peak within the ±50-bps region with respect to the 5' end of lncRNA transcripts (Additional File 2 of the published work [36]).

Possession of a poly(A) tail for lncRNA transcripts

In the thesis, the influence of RNA-seq data with two different types of library constructions, poly(A)-enriched and ribo-zero libraries, was also investigated. The data showed that 190 lncRNA transcripts were only detected in the reads from the ribo-zero library, but not in the reads from the poly(A)-enriched library. This indicates that some lncRNA transcripts do not contain poly(A) tails when they are expressed in the brain. Such lncRNA transcripts can be detected only by the ribo-zero library construction.

Five prime (5') end completeness of the curated lncRNA transcripts

Next, we utilized the peak detection results of 34 CAGE (Cap Analysis Gene

doi:10.6342/NTU201601384

Expression) datasets from the study of Brown *et al.* [56] to investigate the 5' end completeness of the curated lncRNA transcripts. The result showed that ~55% of the curated lncRNA transcripts can find a CAGE peak within the ±50-bps region with respect to the 5' end of lncRNA transcripts (Additional File 2 of the published work [36]).

4.3.3 Evolutionary conservation of lncRNAs

The coverage of scored bases in UCSC-15-way alignment is about 100% in euchromosomes and 83% in heterochromosomes (Table 9). These conservation scores are used to estimate the conservation level of lncRNAs when compared to the other regions of the genome. The conservation analysis showed that the conservation score in each euchromosome was higher than the heterochromatin, which confirmed that the euchromosome enriched in highly conserved functional element across 15 species (Table 10). In this thesis, we calculated the conservation score for each lncRNA according to the 15-way alignment. Many functional lncRNA are evolutionary conserved. In this regard, we wonder whether lncRNAs have higher conservation scores than the repeated sequences. By the 15-way alignments, lncRNA exons are significantly more conserved than the repeated sequences masked in the genome, but are less conserved than mRNAs (Table 11). We unexpectedly found that both the lncRNA sequences and the 500 bp upstream and 200 bp downstream regions from transcriptional

doi:10.6342/NTU201601384

Table 9. Coverage of scored bases in UCSC 15-way alignment

Chromosome	Length (bp)	Covered length (bp)	Coverage (%)
chr2L	23011544	22988826	100
chr2LHet	368872	295131	80
chr2R	21146708	21090805	100
chr2RHet	3288761	2671629	81
chr3L	24543557	24508307	100
chr3LHet	2555491	2387033	93
chr3R	27905053	27844634	100
chr3RHet	2517507	2230008	89
chr4	1351857	1292099	96
chrM	19517	19509	100
chrU	10049037	7342984	73
chrUextra	29004656	20560285	71
chrX	22422827	22250064	99
chrXHet	204112	202603	99
chrYHet	347038	228816	66

Table 10. Average conservation scores of each chromosome

Chromosome	Avg. Score	Avg. Score (including unaligned sites)
chr2LHet	0.089	0.071
chr2L	0.426	0.426
chr2RHet	0.096	0.078
chr2R	0.433	0.432
chr3LHet	0.091	0.085
chr3L	0.435	0.434
chr3RHet	0.096	0.085
chr3R	0.469	0.468
chr4	0.208	0.199
chrM	0.731	0.731
chrUextra	0.174	0.123
chrU	0.112	0.082
chrXHet	0.163	0.162
chrX	0.384	0.381
chrYHet	0.157	0.104

doi:10.6342/NTU201601384

Table 11. Conservation scores of different sequence groups

Seq. Type	Number of Scored Sequences	Avg. Score
mRNA promoter [-500, +200]	29,615	0.328
mRNA	22,620	0.480
lncRNA promoter [-500, +200]	4,220	0.381
lncRNA	4,286	0.419
3'end UTR exons	17,523	0.378
masked region (>= 200 bp)	14,977	0.069

starting site of the lncRNAs, the potential promoter regions, were more conserved than the mRNA promoters. This finding was inconsistent with human lncRNA. We suspect that it is due to the compact genome in *Drosophila* species [73].

4.3.4 Supporting evidences for lncRNA expression in the developmental stages

There is an increasing interest in the use of ChIP-seq data (H3K4me3, H3K36me3 and Pol II) to detect signatures of lncRNA transcription. Existing data of chromatin signatures and expression profiles of *D. melanogaster* were applied to examine the associated chromatin modifications and the expression levels of lncRNAs. For each lncRNA, the presence of transcription-related chromatin signatures chromatin signatures was provided in Additional File 2 of the published work [36]. In combination with the information of expression profiles and chromatin signatures, we found that a large proportion of expressed lncRNA transcripts (RPKM > 1) were not occupied by

doi:10.6342/NTU201601384

H3K4me3, H3K36me3 and Pol II chromatin signatures, which are believed to be present in the actively transcribed regions [38, 39, 57].

Expression profiles

To quantify the expression level of lncRNAs, the RPKM value of every lncRNA transcript at each developmental stage was calculated along with the averaged values of all lncRNA molecules and the averaged values of all mRNA molecules. Figure 6(a) shows that mRNA, on average, had ~8-fold higher expression than lncRNA at developmental stage. Moreover, Figure 6(b) shows that the numbers of transcripts expressed at the developmental stages are similar to those reported in the original study [6]. On average, lncRNA molecules occupied ~4.3% of all transcripts expressed at the

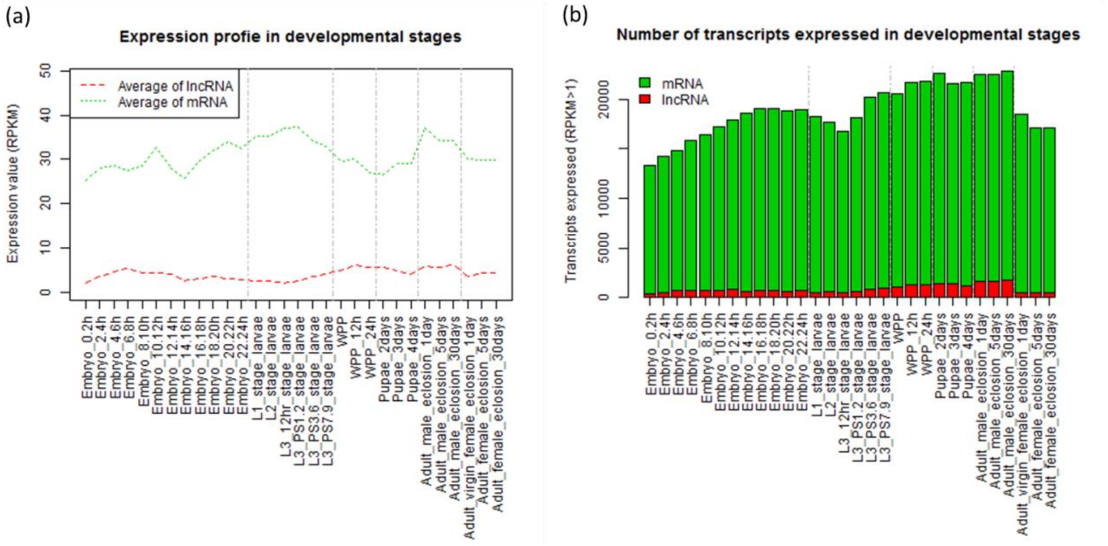


Figure 6. Expression profiles at different developmental stages of fruit fly. (a) Averaged RPKM values at different developmental stages for lncRNAs and mRNAs. (b) Numbers of expressed transcripts (RPKM > 1) at different developmental stages for lncRNAs and mRNAs, respectively.

doi:10.6342/NTU201601384

each developmental stages.

We further investigated the expression profile of every curated lncRNA by Heatmap. In this analysis, only the lncRNAs (2,926 lncRNAs) which varied between at least two developmental stages are considered. As showed in Figure 7, it was observed that lncRNAs were expressed across all of the developmental stages, while around half of the lncRNAs were highly expressed in the stages from white prepupae to adult male.

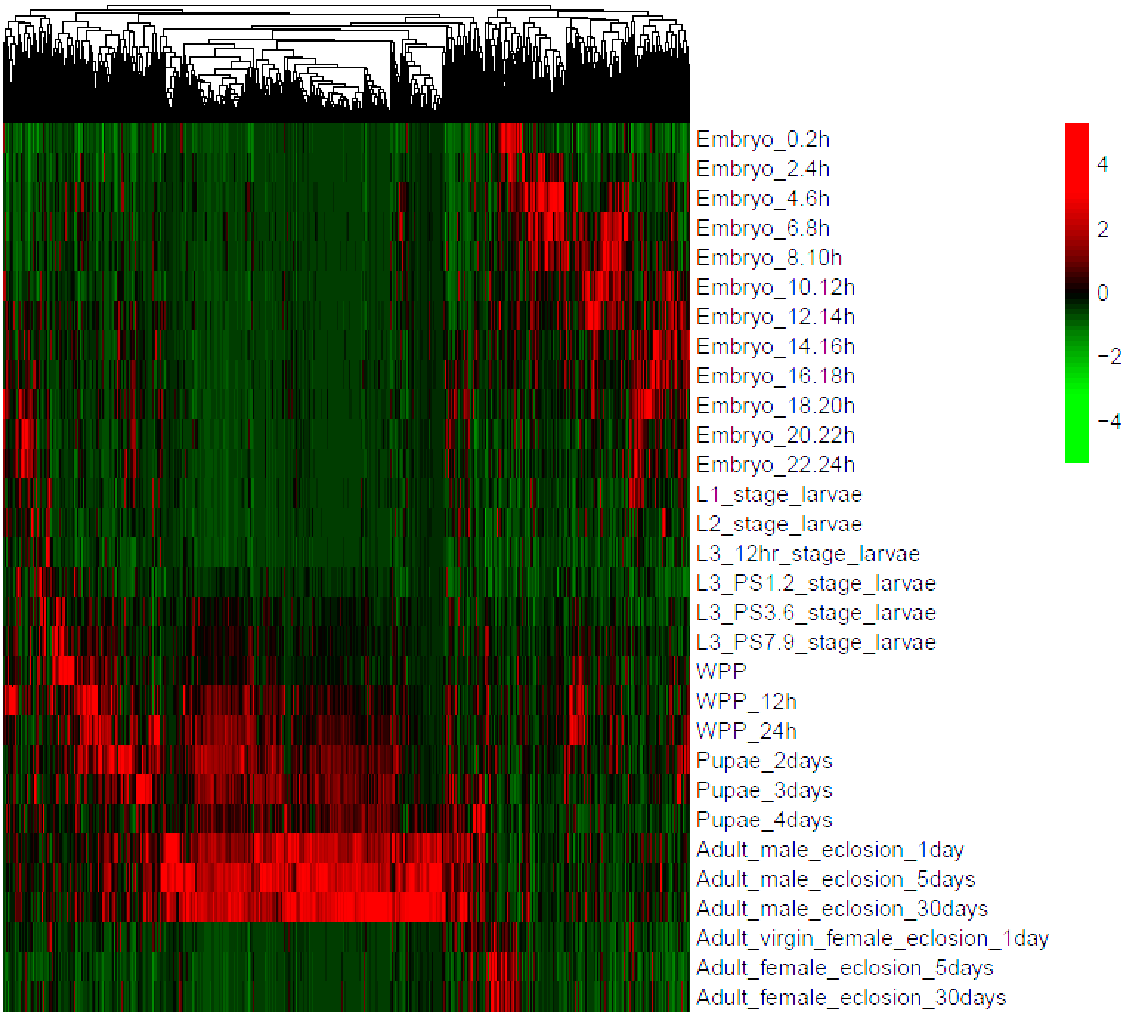


Figure 7. Expression profiles of the 2,926 lncRNAs which varied between at least two developmental stages

doi:10.6342/NTU201601384

This might be explained by the property, which have been frequently observed in other species, that lncRNAs are usually expressed in a cell type-, tissue-, developmental stage- or disease state-specific manner [1, 49, 75]. The dendrogram in Figure 7 also revealed that the lncRNAs expressed in the stages from L3 to adult male agglomerated into a cluster faster than that expressed in other stages. This suggested that these L3/while prepupae/pupae/adult male related lncRNAs represented highly correlated expression with each other. While using a stringent cutoff of correlation, 0.9, to identify co-expressed lncRNA clusters, about one third lncRNAs were found that are not co-expressed with any other lncRNA (singleton). In fact, it was observed that the distribution of members in a cluster followed the power-law distribution (Figure 8), which suggests that most of lncRNAs were not co-expressed with other lncRNAs. Nevertheless, we found that, in fruit fly, there are a few co-expressed lncRNA clusters contain a number of lncRNAs more than ten. Figure 9 showed that most of these clusters are associated with the stages of L3/while prepupae/pupae/adult male.

doi:10.6342/NTU201601384

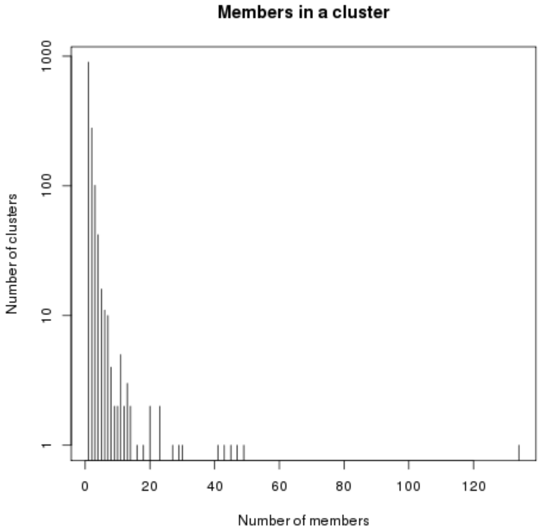


Figure 8. Histogram of members in a co-expressed lncRNA cluster among the 2,926 lncRNAs which varied between at least two developmental stages

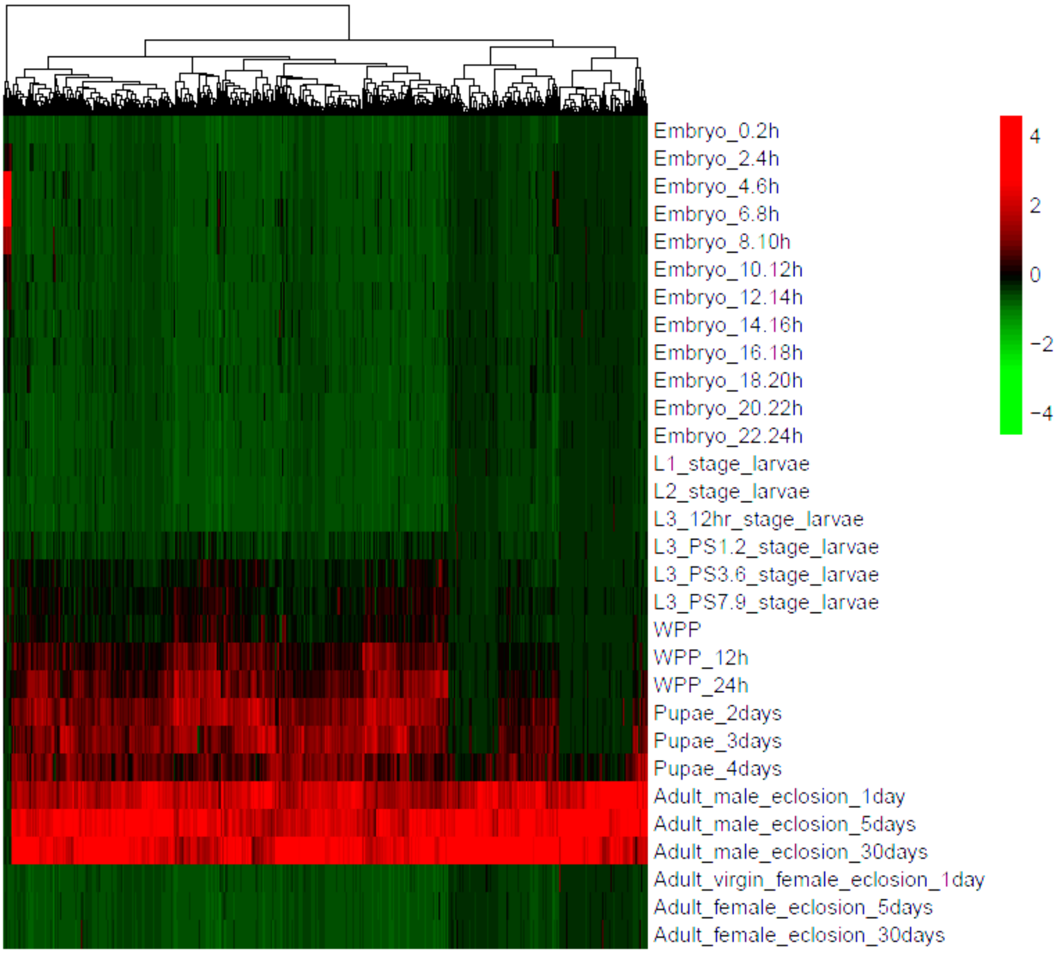


Figure 9. Expression profiles of the lncRNAs which are co-expressed with at least 9 other lncRNAs (Namely, a co-expressed lncRNA cluster is selected for this figure while it has at least 10 members).

doi:10.6342/NTU201601384

Chromatin signatures

In the set of curated lncRNAs, 1,119 of the 3,625 lncRNA genes with well-defined transcriptional direction had a detectable H3K4me3 signal at the proximal region of the genes (Figure 10). In addition, 650 lncRNA genes had detectable H3K36me3 signals, covering, on average, ~70% of the transcribed regions. We also examined the Pol II ChIP-seq data and found that 1,687 (44%) lncRNA genes had Pol II signals with an average coverage of ~60% over the transcribed regions. In summary, 433 lncRNA genes showed 'K4–K36' and Pol II signatures, strongly suggesting that these lncRNAs were epigenetically regulated like protein coding genes. We were aware of the possibility that the chromatin signatures assigned to the lncRNA genes were actually associated with

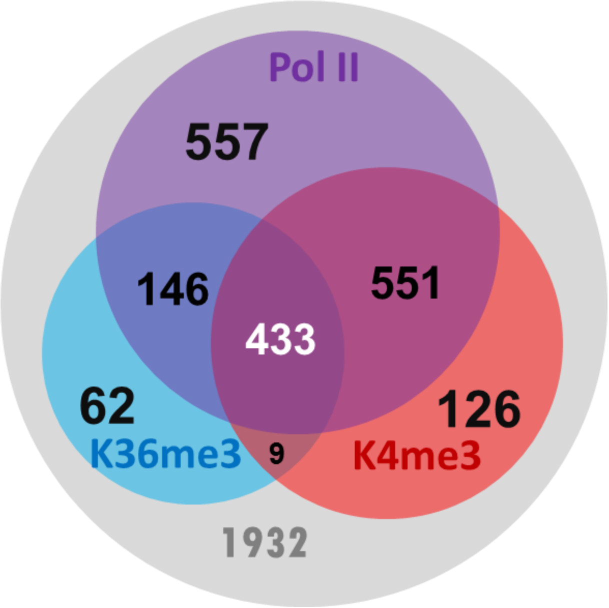


Figure 10. Analysis of chromatin signatures (Pol II, H3K36me3 and H3K4me3) in the curated lncRNA genes.

doi:10.6342/NTU201601384

the overlapped coding genes. There are 340 sense exonic/intronic lncRNA genes that may encounter such a situation.

Combination with the information of expression profiles and chromatin signatures

To study whether the lncRNAs reported by RNA-seq were associated with chromatin modifications, we collected ChIP-seq datasets of the three chromatin signatures, H3k36me3, H3k4me3 and Pol II, which are known to be strongly associated with transcription [38, 39, 57]. The collected datasets involved samples from embryos, larvae, pupae and adults of *D. melanogaster*, with the exception of H3k36me3 datasets in which pupae were not found. Furthermore, RNA-seq datasets of different fly developmental stages were included to quantify the expression of lncRNAs. We found that a large proportion of the expressed lncRNAs (RPKM >1) were not occupied by chromatin signatures, H3K4me3, H3K36me3 and Pol II. This observation raised the question of whether RNA-seq is a reliable platform for detecting transcription of lncRNAs. As both the inference of lncRNA expression and chromatin signatures of transcription were obtained using high-throughput technologies, we used RT-qPCR to detect the transcription of lncRNAs and address this inconsistency issue between expression and chromatin signature data in CHAPTER 5.

doi:10.6342/NTU201601384

4.4 Methods for annotation of the curated lncRNAs

4.4.1 Improving the annotation of curated lncRNAs

To understand the characteristics of the collected and the newly discovered lncRNAs, we integrated a great amount of sequencing datasets to curate information on transcriptional direction, exon regions, classification, expression in the brain, possession of a poly(A) tail, and evolutionary conservation as follows.

Transcriptional direction and exon regions

We determined the transcriptional direction and exon regions of each lncRNA based on the existing annotation from databases as well as the strand-specific RNA sequencing data, from both the present study and the modENCODE database [76]. For the lncRNAs discovered in the present study, both sequencing reads from poly(A)-enriched and total RNA libraries were generated by a strand-specific protocol, so that the transcriptional direction and the exon regions of the assembled transcripts could be unambiguously determined by Cufflinks. As for the lincRNAs from the study by Young *et al.* [5], 29 stranded poly(A)-enriched RNA-seq datasets sampled from different developmental stages and multiple tissues (modENCODE IDs: 4291-4319 as shown in Table 3 and Additional File 3: Table S2 of the published work [36]) were additionally collected and used to determine the transcriptional direction and exon regions, as the RNA library construction of the datasets originally used by Young *et al.* [5] was not strand-specific.

Classification of lncRNAs

Based on the relative location and direction to the closest adjacent coding gene, we divided the lncRNA transcripts into three major classes by in-house perl scripts: (a) lncRNAs imbedded in the introns of protein-coding genes are classified as intronic lncRNAs; (b) lncRNAs that do not overlap with any coding genes are classified as intergenic lncRNAs; and (c) lncRNAs that overlap with an exon in protein-coding genes are classified as exonic overlapping lncRNAs (Figure 11). All exonic and intronic overlapping lncRNAs were then subdivided into sense and antisense depending on the direction of the protein-coding gene. Unclassified lncRNAs were denoted as an unknown group. Here, as in Young *et al.* [5], we used the annotated gene reference from the UCSC genome browser (Sep. 21st, 2015).

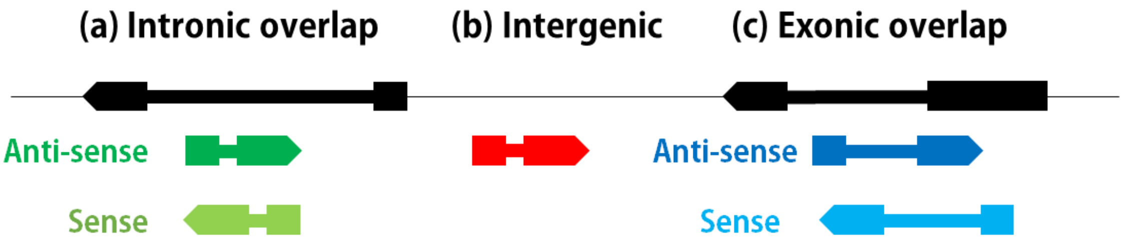


Figure 11. Rules for classifying lncRNAs. Black arrows (transcripts) represent coding genes and colored transcripts are lncRNAs. (a) lncRNAs with intronic overlaps. This group includes lncRNAs (dark green and light green transcripts) located in intronic regions of coding genes (black transcripts). (b) Intergenic lncRNAs. This group includes lncRNAs (red transcripts) located in regions between two coding genes (black transcripts). (c) lncRNAs with exonic overlaps. This group includes lncRNAs (dark blue and light blue transcripts) overlapping exonic regions of coding genes (the black transcript).

doi:10.6342/NTU201601384

Expression in the Brain

As the sequencing reads of the present study were sampled from the brains of fruit flies, we could thus tell whether a lncRNA was expressed in the brain or not. For each of the sequencing read datasets produced in the present study, the two paired-end sequencing reads (read 1 and read 2) were first concatenated into one read set. Next, we remapped the reads onto the transcript set of the collected and the newly discovered lncRNA transcripts using Bowtie [77] followed by eXpress [78] to normalize the read counts of transcripts as Reads Per Kilobase of transcript per Million mapped reads (RPKM). The lncRNA transcripts with a RPKM greater than 1 were defined as “expressed”.

Possession of a poly(A) tail

To answer the question regarding whether a poly(A) tail is required for an expressed lncRNA, the sequencing reads of the present study were generated by using two types of library construction: one was enriched by poly(A) tails (poly(A)-enriched protocol), while the other (ribo-zero protocol) was not. These two types of sequencing reads were quantified with the same procedure as described in ‘Expression in the brain.’ Then, we adopted a stringent criterion to define the group of expressed lncRNA transcripts containing no poly(A) tail if they were expressed in the ribo-zero RNAs (RPKM > 1) but not in the poly(A)-enriched RNAs (RPKM = 0). A stringent criterion is adopted because total RNA sequencing reads with ribo-zero library construction may include mature mRNAs (the major group of RNAs containing poly(A) tails), immature RNAs, partially transcribed RNAs, small RNAs, lncRNAs, etc.

Conservation scoring

Conservation scores were calculated based on the multiple sequence alignment (UCSC-15-way alignment) provided by the UCSC genome browser. The employed multiple sequence alignment included twelve *Drosophila* species, two *Anopheles* species and *Tribolium castaneum*. The position-wise conservation scores for each chromosome base can be downloaded directly from the UCSC web page. For a genome region of interest (e.g., a mRNA sequence, a lncRNA sequence, a promoter region, a 3'UTR exon, or a repeat-masked region), the conservation scores for all the bases within the region were averaged. In case there are conservation scores missing on any of the positions in the given region, the record was excluded before calculating the average score for a particular group (e.g., mRNA, lncRNA, mRNA promoter, or lncRNA promoter).

4.4.2 Genomic and transcriptomic data for supporting lncRNA expression in the developmental stages

4.4.2.1. Collection of *D. melanogaster* sequences

Sequence information of *D. melanogaster* were mainly collected from UCSC genome browser [79] and FlyBase [80]. Sequences from UCSC genome browser (version:

doi:10.6342/NTU201601384

Apr.2006; BDGP R5/dm3) included (i) genome sequences (15 chromosomes); (ii) 3' untranslated region of each coding gene (17,769 sequences); (iii) introns of each coding gene (18,617 sequences). Moreover, mRNA transcripts (30,306 sequences) were extracted from the genome annotation file (in gff format) provided by FlyBase (version: r5.57) for the subsequent expression profiling analysis.

4.4.2.2. LncRNA expression during development of *D. Melanogaster*

The gene expression profile of each lncRNA was measured by Illumina sequencing reads of 30 developmental stages (modENCODE IDs: 4433-4462 as shown in Additional File 3: Table S2 of the published work [36]), from 0-2 hr embryos through 30-day male and female adults, provided by Graveley *et al.* [6]. The sequencing reads were pre-processed by trimming 10 bp from the 5' end to eliminate random primer effects [81]. Bases from the 3' end were also trimmed until a quality score higher than 20 was reached. In addition, only reads that were at least 36 bp in length were retained for subsequent analysis. The qualified reads were then mapped onto all transcripts including both mRNA (collected from FlyBase r5.57; see section 4.4.2.1) and lncRNA sequences using Bowtie [77] and the read counts of transcripts were normalized as RPKM using eXpress [78].

4.4.2.3. Chromatin signatures during development of *D. Melanogaster*

Like protein coding mRNAs, many expressed lncRNAs in mammalian cells contain a

doi:10.6342/NTU201601384

‘K4–K36’ signature [82]. That is, H3K4me3 is present in the promoter region, followed by a longer stretch of H3K36me3 extending throughout the entire transcribed region. In this thesis, we integrated the ChIP-seq data containing information of ‘K4–K36’ histone modifications to further characterize the collected lncRNAs. To assign H3K4me3 signals to an lncRNA, we defined regions 500 bp upstream and 100 bp downstream, with respect to the transcription start, as the promoter region and used pre-defined protein binding sites from H3K4me3 ChIP-seq datasets collected from modENCODE [76]. Next, we examined H3K36me3 modifications and calculated the coverage as a percentage of the transcribed region in a lncRNA that was covered by the H3K36me3 signal. In addition, as Pol II occupancy can also reveal expression of transcripts, we also considered Pol II occupancy across the promoter region and the transcribed region for a lncRNA as an essential chromatin signature. The modENCODE IDs of all ChIP-seq datasets used in this thesis are listed in Appendix Table 1. The specific definition of occupied regions for each chromatin signatures is shown in Figure 12.

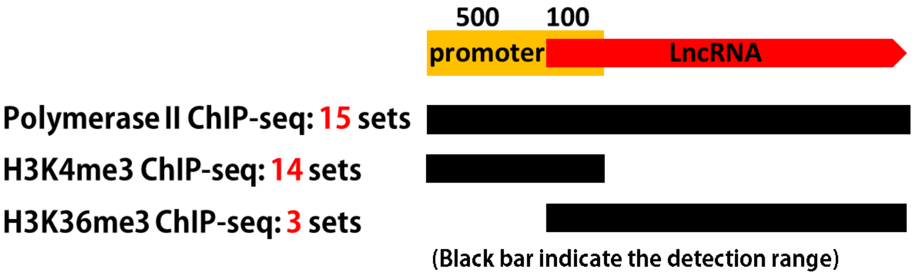

500
100
promoter
LncRNA
Polymerase II ChIP-seq: 15 sets
H3K4me3 ChIP-seq: 14 sets
H3K36me3 ChIP-seq: 3 sets
(Black bar indicate the detection range)

Figure 12. Occupied regions for each chromatin signature

doi:10.6342/NTU201601384

CHAPTER 5 Reliability of lncRNA expression

In section 3.2, “ Novel lncRNAs identified from brain samples”, we discovered a set of novel lncRNAs from the generated RNA-seq datasets. Since these lncRNAs were newly discovered, it is essential to have supporting evidences for the discovery. In fact, it is uncertain whether the lncRNAs predicted from other studies are truly expressed as well. Therefore, we selected a set of novel lincRNA genes and a set of the curated lncRNAs for RT-qPCR validation (section 5.1 and 5.2). As for novel lincRNA genes, we conducted three more analysis to investigate the quality of the found lncRNAs (section 5.1). Moreover, in section 4.3.4, “Supporting evidences for lncRNA expression in the developmental stages”, the inconsistency between RNA-seq and Chromatin signatures have been observed and raised an issue whether RNA-seq is a reliable platform for identifying lncRNAs. In this chapter, we addressed this issue by conducting a series of RT-qPCR experiments (section 5.2).

5.1 Reliability of the lncRNAs newly discovered identified from brain samples

To investigate the quality of the lincRNAs discovered in the present study, we conducted three analyses and selected a set of lincRNA genes for RT-qPCR validation to investigate the reliability of these newly discovered lncRNAs. The following results revealed the high reliability of the discovered novel lncRNA genes.

For a lncRNA, it was examined whether (1) it was observed to be expressed in the collected RNA-seq datasets from developmental stages; (2) it was predicted with a low coding probability by the second predictor; and (3) it was not predicted to contain any conserved domains of proteins. As shown in Additional File 5 of the published work [36], 86.15% of the 462 novel lncRNA genes discovered from fly brain were also observed expressed in at least three developmental stages. In the proposed workflow of discovering lncRNAs, we applied a SVM-based prediction tool, Coding Potential Calculator (CPC) [68], to filter out potential coding sequences. Here, we applied another tool for estimating coding potential, Coding-Potential Assessment Tool (CPAT) [83], on the discovered lncRNAs. The result (Additional File 5 of the published work [36]) showed that only seven transcripts were with a coding probability ≥ 0.39. This cutoff threshold 0.39 was an optimum cutoff for fruit fly suggested by Wang *et al*. [83], where 96% of fly coding genes were shown to have a coding probability ≥ 0.39 (data shown on the tool download page). Moreover, the results of invoking RPS-BLAST showed that only nine newly discovered lncRNA transcripts might contain conserved domains from the Conserved Domains database (CDD, version 3.4), as shown in Additional File 5 of the published work [36] as well.

Additionally, 22 novel lncRNA genes were randomly selected for RT-qPCR

doi:10.6342/NTU201601384

experiments applied on fly brains. In Figure 13(a), the results showed that 17 novel lncRNA genes have adequate expression (–delta Ct ≥ 1). For the five lncRNAs of which the expression was not clear (–delta Ct < 1), we doubled the amount of template brain cDNA and performed RT-qPCR again on these five low-expressed lncRNA genes. In the second RT-qPCR validation experiment, seven FlyBase lncRNA genes that were believed to be expressed in brains and three FlyBase lncRNA genes that were believed to be unexpressed in brains were also included for comparison. The ten FlyBase

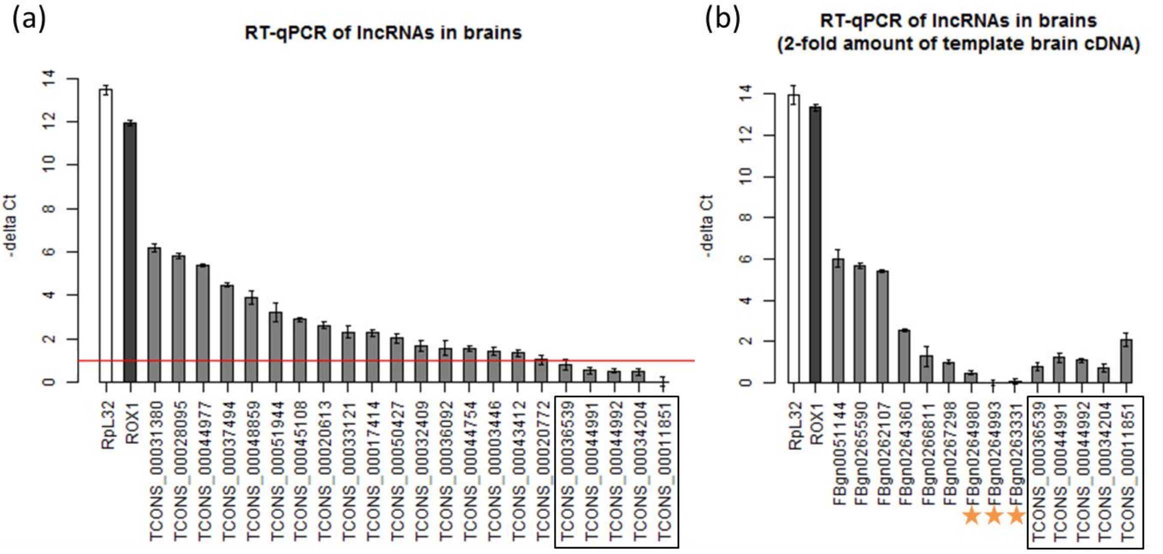


Figure 13. RT-qPCR experiments for a selected set of lncRNAs in brains.

(a) 22 novel lncRNAs discovered in the present study were selected for validation. RpL 32 (a coding gene) and ROX1 (a non-coding gene) were included as positive controls. The horizontal line indicated –delta Ct ≥ 1. The rectangle indicated the five lncRNAs with considerably low expression, and was tested again by the second RT-qPCR experiment shown in (b). (b) The five lncRNAs from the rectangle of (a) were tested again by RT-qPCR with twofold amount of template cDNA. Ten FlyBase lncRNAs were included for comparison. The three FlyBase lncRNAs highlighted by the orange stars were selected because their RPKM values in our brain RNA-seq data was 0.

doi:10.6342/NTU201601384

lncRNAs were selected according to the RPKM values from our poly(A)-enriched RNA-seq data of brain (RPKM > 1 suggested expressed; RPKM = 0 suggested unexpressed). The results in Figure 13(b) revealed that the expressed and unexpressed FlyBase lncRNA genes showed distinct values in RT-qPCR experiments. When compared with the three unexpressed FlyBase lncRNA genes, the five novel lncRNA genes were also considered expressed in brains. Here, we demonstrated that novel lncRNAs can be found in a tissue-specific manner, as suggested by a previous study in mammals [34].

5.2 Experimental validation of a selected set from the curated lncRNAs by RT-qPCR

To investigate whether the collected lncRNA genes were indeed actively transcribed, we used RT-qPCR to detect the expression of a selected set of lncRNAs in adult male flies. A set of lncRNAs expressed in adult male flies (RPKM >1) were selected and divided into four groups according to two properties: (a) lncRNAs with all of the three chromatin signatures (H3K4me3, H3K36me3 and Pol II) or without any of the three chromatin signatures, and (b) lncRNAs with high expression (RPKM > 3rd quartile, i.e., 12.92) or with low expression (RPKM < 1st quartile, i.e., 2.78). In each group, at least 10 lncRNAs were randomly selected to be validated with RT-qPCR. The four groups were defined as (G1) high expression with chromatin signatures (11 lncRNA genes),

doi:10.6342/NTU201601384

(G2) low expression with chromatin signatures (11 lncRNA genes), (G3) high expression without chromatin signatures (10 lncRNA genes) and (G4) low expression without chromatin signatures (10 lncRNA genes). In total, we selected 42 lncRNA genes which were reported by RNA-seq.

The results revealed that most lncRNA genes (95.24%) were indeed present at the chosen stage (male adults) of the fruit flies (Figure 14 and Appendix Table 2). Among

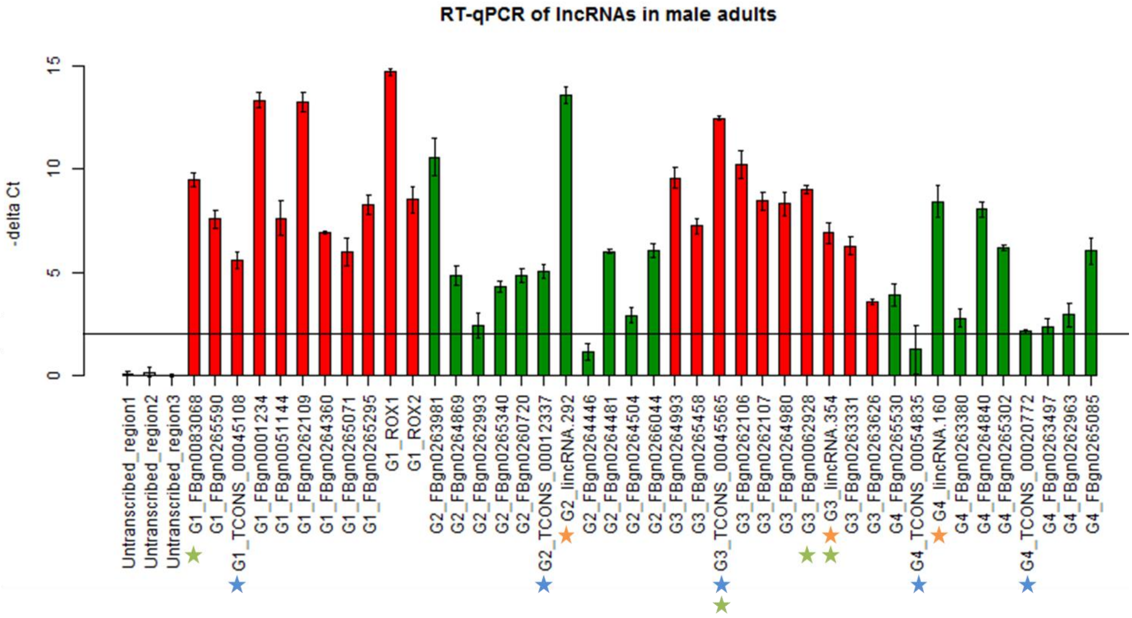


Figure 14. RT-qPCR experiments of a selected set of lncRNAs in male adults. G1: high expression with chromatin signatures (11 lncRNAs); G2: low expression with chromatin signatures (11 lncRNAs); G3: high expression without chromatin signatures (10 lncRNAs); and G4: low expression without chromatin signatures (10 lncRNAs). Three negative controls (un-transcribed region 1, 2, and 3) were all around zero. Stars were used to highlight the lncRNAs that were not from the databases (Orange stars: the selected lncRNAs from Young *et al*. [5]. Blue stars: the lncRNAs from the present study). The horizontal line indicated the cutoff (–delta Ct ≥ 2) used to define a validated lncRNA. Green stars: the transcripts that are now annotated as other types of transcripts by FlyBase, and thus were removed from the list of the curated lncRNAs in the present study.

doi:10.6342/NTU201601384

the validated lncRNA genes, three lincRNA genes (lincRNA.354 is now annotated as a protein-coding gene in FlyBase) were discovered by Young *et al.* [5] and five lncRNA genes (TCONS_00045565 is now annotated as an rRNA gene in FlyBase) were reported by the present study. Two known lncRNA genes expressed in male adults, roX1 and roX2 [3], were also validated by RT-qPCR. These observations confirmed that most of the lncRNA genes identified by RNA-seq are not transcriptional noise, and provided strong support that RNA-seq is a reliable tool to identify lncRNA genes. In addition, by dividing the 42 selected lncRNA genes into four groups with all possible combinatorial conditions of chromatin signatures (present or absent) and expression (high or low), the data showed that in all four groups, all lncRNA genes except two with low expression (one lncRNA gene in G2 and one lncRNA gene in G4) could be successfully detected by RT-qPCR. This observation held even for the expressed lncRNA genes that had none of the three chromatin signatures (G3 and G4). Our results suggested that the lack of associated H3K4me3, H3K36me3 and Pol II signatures might not directly imply no transcription of lncRNAs, since most of the expressed lncRNA genes without these three chromatin signatures (G3 and G4) were successfully detected by RT-qPCR. However, it should be noted that the collected ChIP-seq datasets were not sampled from the stages as precisely as the RNA-seq datasets, which were collected from 30 time points (12 for embryos, 6 for larva, 3 for white pupae, 3 for pupae, 3 for male adults and

3 for female adult stages) during the development of *D. melanogaster*. The inconsistency between RNA-seq and ChIP-seq data may be because the collected ChIP-seq data were not extensive. In particular, ChIP data of H3K36me3 sampled from pupae was not found during data collection.

5.3 Details of the RT-qPCR experiments

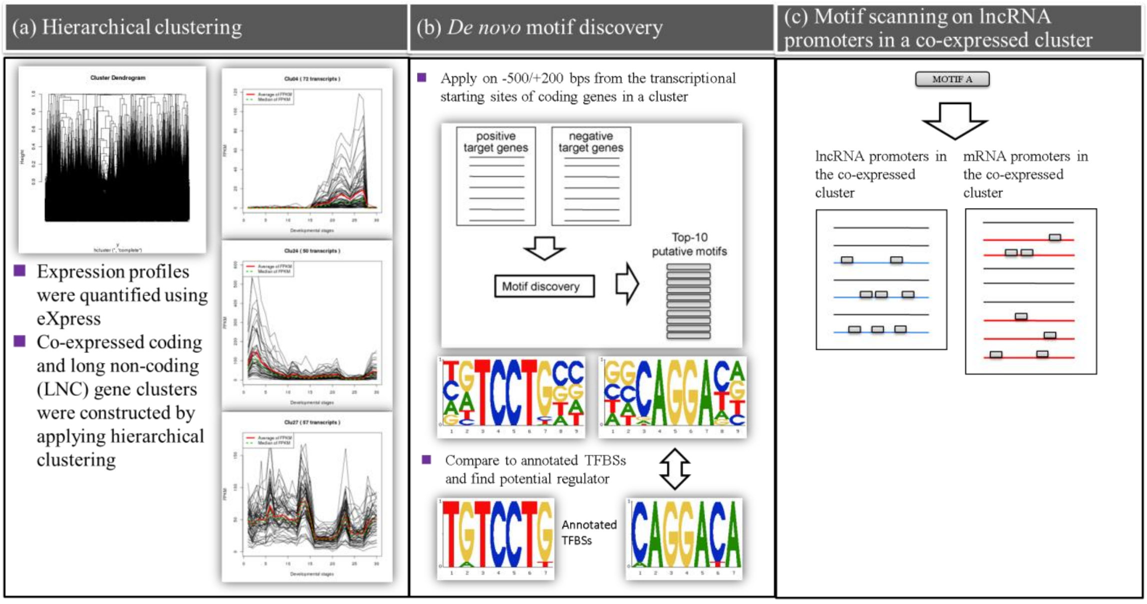
In this thesis, real-time quantitative PCR (RT-qPCR) experiments were adopted for validating the expression of two selected lncRNA sets in two types of samples, brains and whole bodies of young male adults (*Canton S*). Total RNA samples were purified from 100 brains and 20 whole bodies, respectively, by using TRIzol® (Invitrogen) and were subsequently treated with DNase to eliminate genomic DNA contamination. Next, 1μg of total RNA were converted to cDNA by random hexamer primers and SuperScript^{TM} reverse transcriptase (Invitrogen) according to manufacturer's protocol along with a negative control without reverse transcriptase. A primer pair for each of the selected lncRNAs was designed, using the Primer-BLAST tool provided by NCBI [84]. The functionality of the designed primer pairs was pre-tested by polymerase chain reactions applied on the genomic DNA purified from 5 *Canton S* larvae. The tests revealed that 35 primer pairs (used in Figure 13) and 42 primer pairs (used in Figure 14) worked well which were then used in subsequent analysis (the primer list is shown in Additional File 3: Table S6 of the published work [36]). Finally, the RT-qPCR

doi:10.6342/NTU201601384

experiments (four technical replicates) were performed for each of the selected lncRNA using OmicsGreen qPCR 5X Master Mix (Omics Bio) on a CFX96^{TM} connect Real-Time PCR System (Bio-Rad). 1/100 of total converted cDNA was used as template cDNA for all RT-qPCR experiments, except for those shown Figure 13(b) in which 1/50 of total converted cDNA was used. In addition, for the experiments of whole bodies (Figure 14), RT-qPCR experiments were also performed on three negative controls randomly picked up from un-transcribed regions (intergenic regions that are not expected to see any transcripts) for comparison.

CHAPTER 6 Regulation of lncRNA Expression

To tackle the issue about transcriptional regulation of lncRNA expression, this thesis proposed a workflow (Figure 15) to identify shared *cis*-elements of co-expressed coding and long non-coding genes (C-LNC). We incorporates *de novo* motif discovery to systemically investigate the presence of *cis*-elements shared by the promoters of C-LNC gene clusters. Co-expressed C-LNC gene clusters were constructed by applying hierarchical clustering (Figure 15(a)) on the expression profiles of 30,306 mRNA from FlyBase (r5.57) and 4,599 lncRNA curated in this thesis. To identify potential regulatory elements, *de novo* motif discovery was conducted on the promoters of coding genes in a cluster. The discovered motifs were also used to identify potential common regulators of



- Expression profiles were quantified using eXpress
- Co-expressed coding and long non-coding (LNC) gene clusters were constructed by applying hierarchical clustering
- Apply on -500/+200 bps from the transcriptional starting sites of coding genes in a cluster
- Compare to annotated TFBSs and find potential regulator

Figure 15. Workflow for identifying shared *cis*-elements of co-expressed coding and long non-coding genes

doi:10.6342/NTU201601384

these C-LNC genes (Figure 15(b)). Then, the discovered motifs were examined to see whether they are also present in the promoters of LNC genes in the same cluster (Figure 15(c)). In the following sections, we provided results and discussions for each step in Figure 15(a)-(c).

6.1 Hierarchical clustering of co-expressed coding and long non-coding genes

To identify co-expressed clusters of coding and lncRNA genes, we applied hierarchical clustering on the expression profiles of the developmental stages for a filtered set of coding and lncRNA genes (see filtering procedure in section 6.4.1). Table 12 showed the numbers of identified clusters decreased along with the descending cutoffs of Pearson’s correlation coefficient. As shown in Figure 16, when investigating the

Table 12. Statistics of clusters with different cutoff of correlation

Cutoff of correlation	total # of clusters	# of clusters with members more than 50
0.9	8082	27
0.8	2823	78
0.7	1255	125
0.6	636	155
0.5	342	146
0.4	185	117
0.3	109	84
0.2	59	53
0.1	38	37

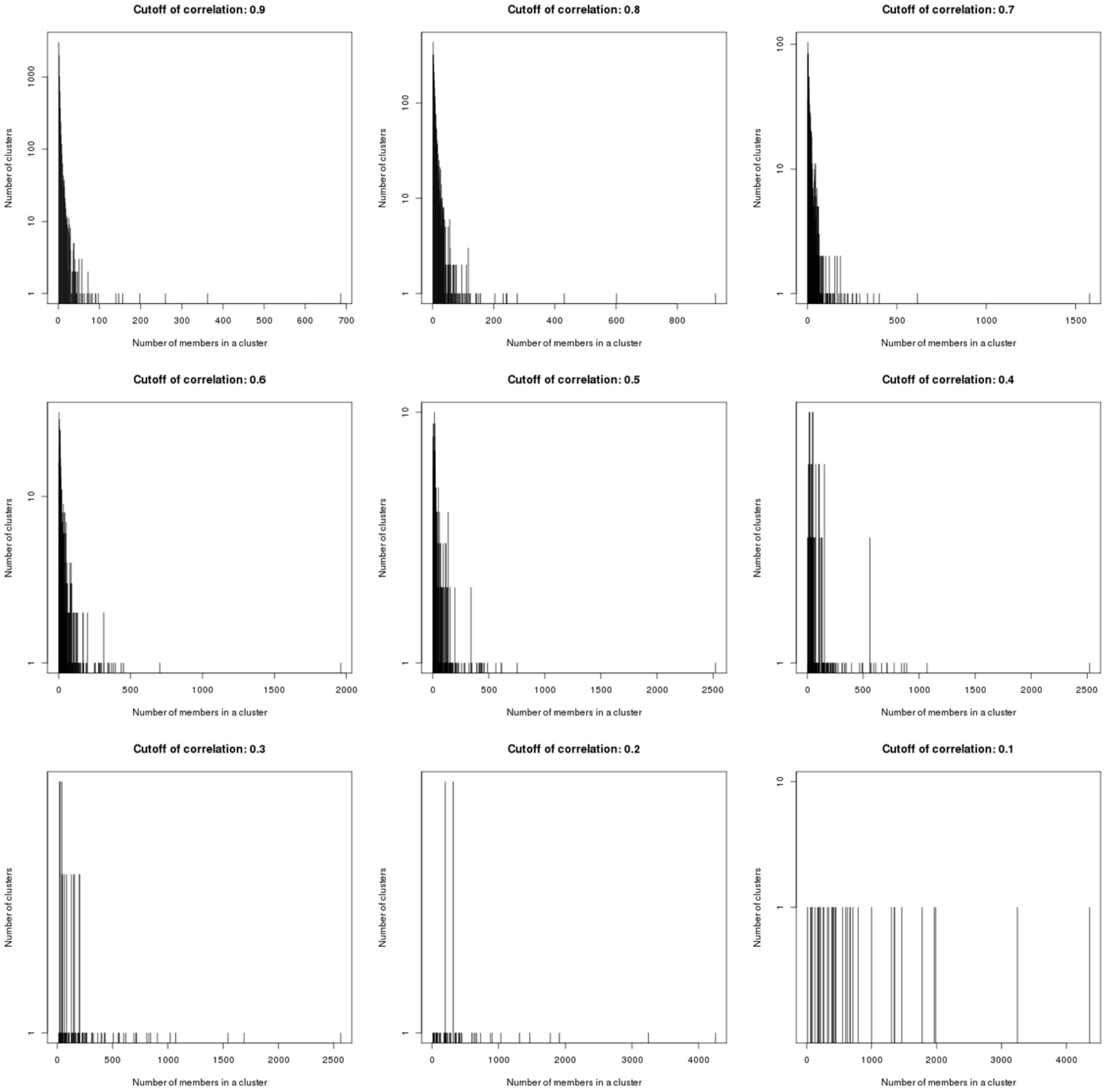


Figure 16. Frequency distribution of clusters along with different member numbers in a cluster

frequency distribution of the clusters with different member numbers in the range of adequate correlation cutoffs (ranging from 0.5 to 0.9), it was observed that a power-law distribution is formed for the frequency of clusters along with the number of members in a cluster. This indicated that a lot of clusters contain only a few members in them. Additionally, to identify highly correlated C-LNC gene clusters, a stringent cutoff of

doi:10.6342/NTU201601384

correlation, 0.9, was adopted. About ~10% of all transcripts (3,021 singletons among the total 29,508 transcripts) were observed as singletons when using 0.9 as the cutoff.

For conducting *de novo* motif discovery in the subsequent analysis, we required the used clusters to have considerable co-expressed transcripts. Therefore, the 27 clusters with members more than 50 transcripts were selected (Table 12), where 20 clusters of the selected clusters included at least one co-expressed lncRNA (Table 13). The Heatmap in Figure 17 showed that the expression profiles of the 27 clusters. It was observed that some of the clusters shared similar patterns with each other. Therefore, we invoked the second hierarchical clustering procedure on the averaged expression profile of each cluster to further categorize the 27 clusters into 7 groups (Table 13). We next combined the group information and the expression profile of the clusters as shown in Figure 18. Interestingly, the results revealed that the co-expressed C-LNC clusters containing at least one lncRNAs were only associated with the stages of L3 to adult male (Group 1), pupae and adult male (Group 2), adult male only (Group 3), and pupae only (Group 4). The other clusters associated with the stages of embryo and adult female (Group 5-7) are found that do not have co-expressed lncRNAs. We suspected that requiring at least 50 transcripts is too stringent for some C-LNC clusters associate with the stages of embryo and adult female, and thus conducted additional Heatmap analysis on the co-expressed clusters with at least 30 transcripts to see whether

doi:10.6342/NTU201601384

Table 13. Summary of the selected clusters which contain at least 50 transcripts

Group ID	Cluster ID	Number of transcripts	Number of lncRNAs (523 in total)	Number of mRNAs (2781 in total)	Number of lncRNA promoters without 'Ns' (517 in total)	Number of mRNA promoters without 'Ns' (2776 in total)
Group 1	Clu01	72	9 (12.5%)*	63	8	63
Group 1	Clu02	147	23 (15.6%)	124	23	124
Group 1	Clu03	90	14 (15.6%)	76	14	76
Group 1	Clu04	72	19 (26.4%)	53	19	53
Group 1	Clu05	363	51 (14%)	312	51	311
Group 1	Clu06	57	19 (33.3%)	38	19	38
Group 1	Clu07	57	10 (17.5%)	47	10	47
Group 2	Clu08	140	30 (21.4%)	110	30	110
Group 2	Clu09	260	38 (14.6%)	222	38	219
Group 2	Clu10	50	19 (38%)	31	19	31
Group 2	Clu11	686	70 (10.2%)	616	68	616
Group 2	Clu12	198	36 (18.2%)	162	35	161
Group 2	Clu13	77	23 (29.9%)	54	23	54
Group 2	Clu14	91	25 (27.5%)	66	24	66
Group 2	Clu15	59	25 (42.4%)	34	25	34
Group 3	Clu16	55	15 (27.3%)	40	15	40
Group 3	Clu17	70	24 (34.3%)	46	24	46
Group 3	Clu18	97	26 (26.8%)	71	26	71
Group 3	Clu19	156	45 (28.8%)	111	44	111
Group 4	Clu20	51	2 (3.9%)	49	2	49
Group 5	Clu21	73	0	73	0	73
Group 5	Clu22	63	0	63	0	63
Group 5	Clu23	83	0	83	0	83
Group 5	Clu24	50	0	50	0	50
Group 5	Clu25	80	0	80	0	80
Group 6	Clu26	50	0	50	0	50
Group 7	Clu27	57	0	57	0	57

* Percentage of lncRNA numbers among all transcripts in a cluster.

doi:10.6342/NTU201601384

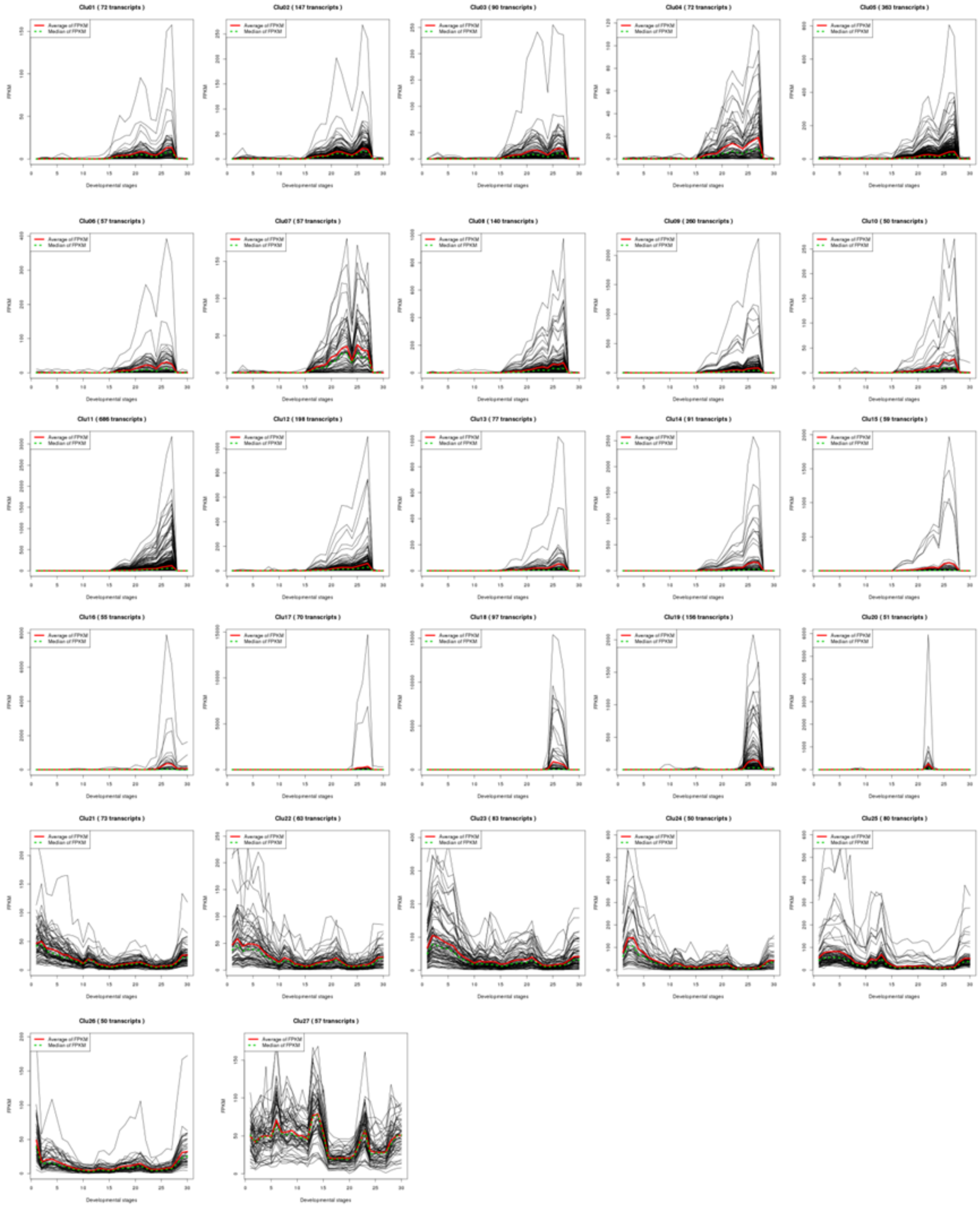


Figure 17. Expression profiles for the lncRNAs of the 27 co-expressed clusters

doi:10.6342/NTU201601384

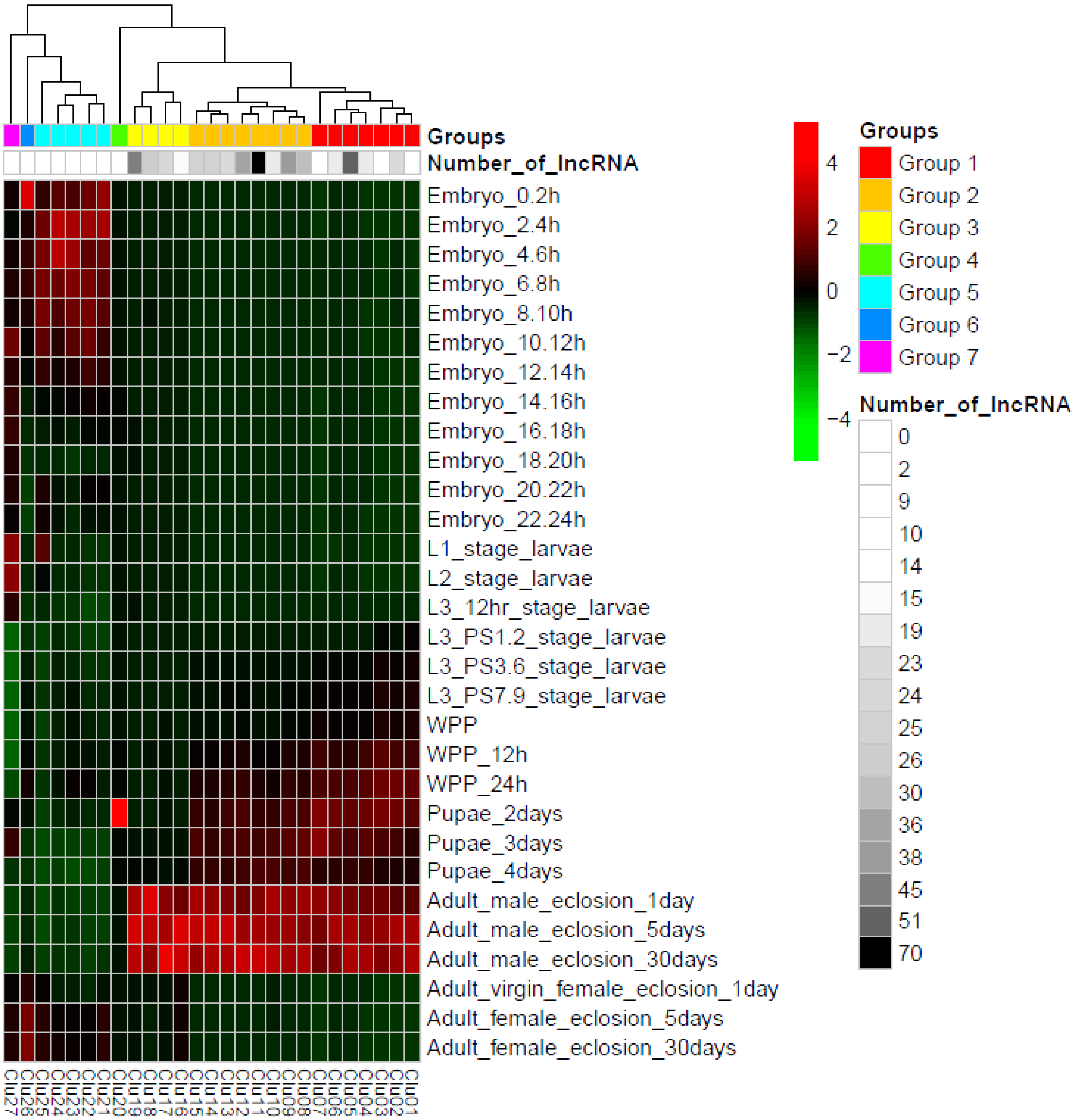


Figure 18. Second-phase hierarchical clustering for the 27 co-expressed clusters with at least 50 transcripts. The rainbow color bar indicates the groups categorized by the second-phase hierarchical clustering. The gray color bar represents the number of lncRNAs within a cluster. Details of group and cluster information can be found in Table 13.

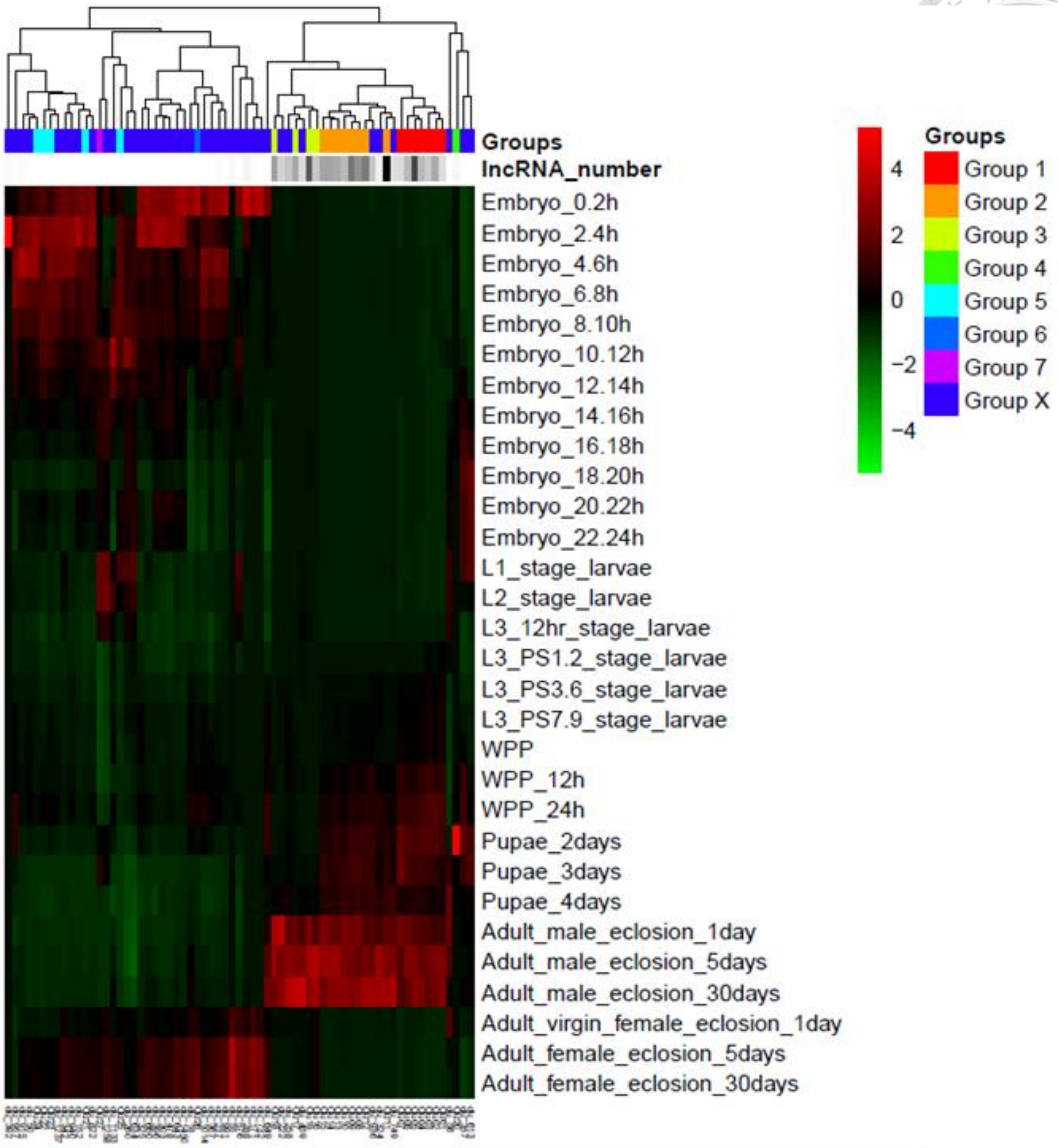


Figure 19. Second-phase hierarchical clustering for the co-expressed clusters with at least 30 transcripts. The rainbow color bar indicates the groups categorized by the second-phase hierarchical clustering. The gray color bar represents the number of lncRNAs within a cluster. Details of Group 1-7 can be found in Table 13, while Group X represents the clusters that contain members of 30~49 transcripts.

the clusters with less members would have co-expressed lncRNAs within them. Figure 19 showed that only limited number of C-LNC clusters (the ones with gray stripes in Group X) could be saved back from the lost information caused by the stringent member number requirement. Especially, for the clusters associated with the stages of embryo and adult female, we rarely observed co-expressed lncRNAs. Thus, we still required 50 transcripts in a cluster for the subsequent analysis of *de novo* motif discovery. In fact, these results were consistent with the observation of lncRNA expression profiles in the developmental stages, which suggested that the L3/while prepupae/pupae/adult male related lncRNAs were more likely co-expressed with other genes (see section 4.3.4).

6.2 *De novo* motif discovery on the promoters of co-expressed coding genes

De novo motif discovery was conducted on the promoters of “coding genes” in a cluster to identify potential regulatory elements. In this section, we first discussed about the issue for how to define the promoter region of a gene. Second, parameters used when conducting *de novo* discovery were tuned and analyzed. Third, the quality of the discovered motifs was evaluated.

6.2.1 Promoter regions of genes in *D. melanogaster*

The gene promoter regions may have lengths varying from hundreds to thousands long, and locate upstream or downstream from transcription start site (TSS), in different species [85, 86]. In *D. melanogaster*, some studies have used (–1,000 to +200 bp) as the [87, 88], and some others used (–100 to +200 bp) [89]. To clarify which

doi:10.6342/NTU201601384

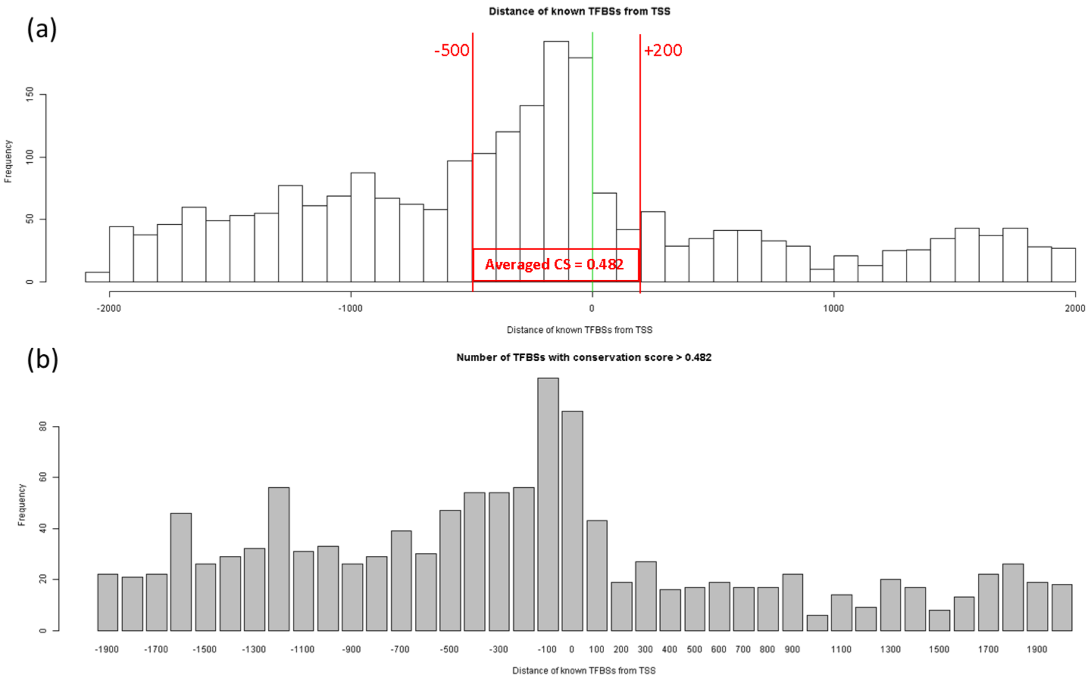


Figure 20. Distribution and Conservation scores (CS) analysis of the 2,059 annotated binding sites collected from REDfly database [90] (170 TFs and 2,048 target genes included). (a) Position distribution. The averages CS of TFBSs located within (−500 to +200 bp) is 0.482; (b) Frequency of TFBSs that have a CS value ≥ 0.482.

region should be considered as the promoters of the identified co-expressed gene cluster for the *de novo* motif discovery, we matched the annotated TFBSs (collected from REDfly database [90]) back onto the gene promoter regions for investigating the patterns of promoter structure. As shown in Figure 20(a), most of annotated TFBSs located at the regions adjacent to the TSS. For the annotated TFBSs located in the region of (−500 to +200 bp), we calculated the average conservation scores (CS). The calculated CS value (0.482) is much higher than the average CS value of mRNA and lncRNA promoters (0.328 and 0.381, respectively; Table 11). In addition, the annotated

doi:10.6342/NTU201601384

TFBSs with the supports of evolutionary conservation (CS ≥ 0.482) were usually located at the regions adjacent to the TSS as shown in Figure 20(b). In this thesis, the region of (–500 to +200 bp) was used for the subsequent *de novo* motif discovery.

6.2.2 Parameter tuning for the weights of nucleosome occupancy and evolutionary conservation while conducting *de novo* motif discovery

To optimize the performance of *de novo* motif discovery using eTFBS [91], we adopted an analyzed procedure to find the best parameter set for the weights of nucleosome occupancy and evolutionary conservation. Here, we selected a fixed pattern support during pattern mining step, 0.15, for the subsequence analysis. The patter support was defined as the proportions of sequences in the coding gene promoters of a co-expressed cluster that contains an observed pattern. With the selected pattern support, it has a possibility to achieve highest precision for the prediction of TFBSs as validated by the annotated TFBSs collected from REDfly database [90] (Appendix Figure 1.   ).

As described in the motif discovery procedure (section 6.4.3), a pattern ranking scheme (*Eq 1*) is used for selecting reliable patterns. In the equation, there are three parameters (*a*, *b*, and *c*) that can be tuned, where *a*, *b*, and *c* are the relative weights given to the position score, nucleosome occupancy score and conservation score. Nevertheless, the position score (with weight *a*) was designed for positive sequences

doi:10.6342/NTU201601384

with scores relevant to reliability, such as *P* value estimated from ChIP-seq experiments. In this thesis, the weight *a* should thus be set as ‘0’, since the positive promoters used in this thesis were collected form each co-expression cluster and have no measured scores relevant to reliability. Therefore, in this section, only the weights (*b*, *c*) of nucleosome occupancy and evolutionary conservation were analyzed. The weights, (0, 1, 2, 3) were used for *b*, while (1, 2, 3) for *c*. In total, there are 12 parameter sets were tested.

To evaluate the performance of the predictions considering different parameter sets, we collected the annotated TFBSs from REDfly database [90] for validation. For each

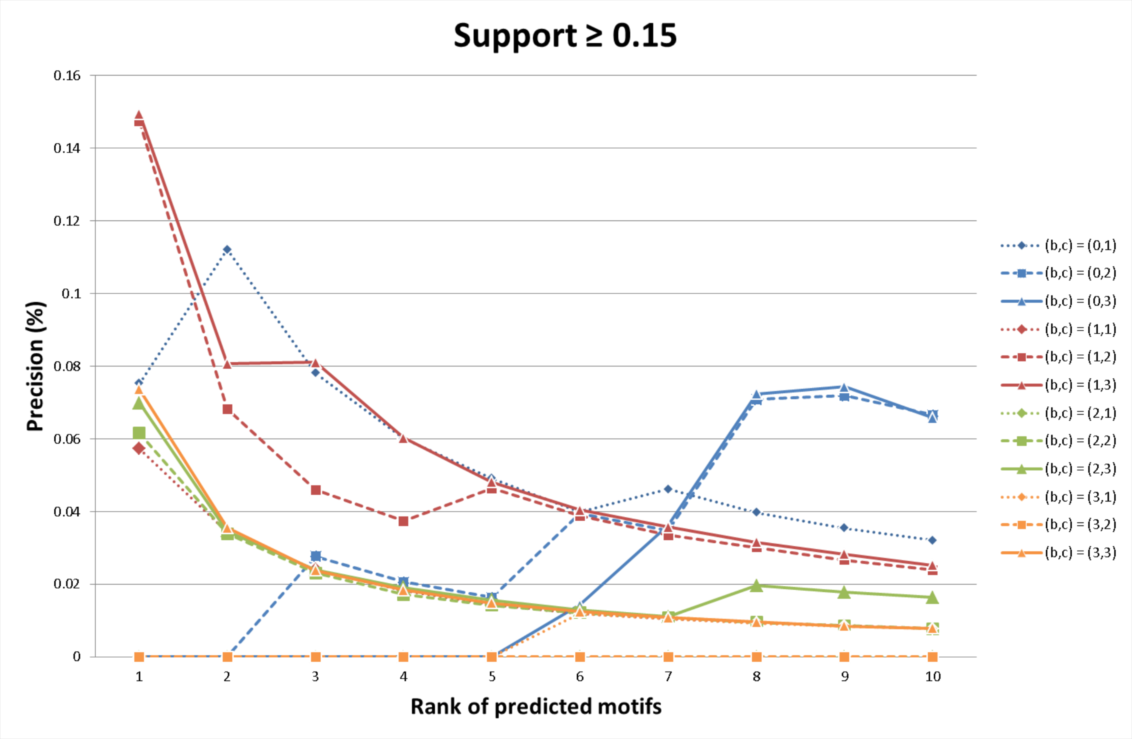


Figure 21. Parameter tuning for the weights (*b*, *c*) which are given to nucleosome occupancy and evolutionary conservation. Different colors denote different weights for nucleosome occupancy. The colors, (blue, red, green, orange), indicate $b \in (0, 1, 2, 3)$. Different types of lines represent different weights for evolutionary conservation. The line types, (solid line, thick broken line, broken line), indicate $c \in (1, 2, 3)$.

doi:10.6342/NTU201601384

run of prediction, a list of top-10 putative motifs, along with their corresponding positions in the positive promoters (instances), was reported. We validated these predicted instances by comparing to the collected annotated TFBSs. Precision scores were calculated by the ratio of (True positives/Predicted instances), where ‘True positives’ were counted when a predicted instance was overlapped with an annotated TFBS. Figure 21 suggested that the information of evolutionary conservation was useful for finding true TFBSs, since it was observed that the highest *c* (solid line) obtained the best precision for each fixed *b* (each line color). Moreover, along with the ranks of the predicted motifs, the result showed that the information of nucleosome occupancy helped to make real TFBSs better ranked when comparing lines in read to lines in blue. Taken together, the parameter set of (*b*, *c*) are empirically set to (1, 3), where the best performance on the prediction of TFBSs was obtained.

6.2.3 Evaluation of the discovered motifs

A list of top-10 putative TFBSs for each cluster was reported, and resulted in 270 putative TFBSs in total for the 27 clusters. About 80% of the predicted motifs (212 motifs among the total 270 motifs) were similar to annotated TFBSs (Table 14). To confirm the results were not random events caused by genome-wide motif mapping, we further mapped the discovered motifs onto 3’ untranslated regions (3’ UTRs) and introns of coding genes. The frequency of motif hits in LNC gene promoters was

doi:10.6342/NTU201601384

Table 14. Summary of *de novo* motif discovery results

Promoter region	**-500/+200**
Num. of clusters with annotated TFBS	27
Num. of predicted motifs	270
Num. of predicted motifs supported by annotated PFMs	212 (78.52%)
Num. of involved annotated PFMs	73

Table 15. Investigation of similarity between lncRNA and mRNA promoters

Discovered motifs matched onto different sequence sets	**p-value of paired t test**
mRNA promoter vs. lncRNA promoter	0.157
mRNA promoter vs. 3' UTR	0.012
lncRNA promoter vs. 3' UTR	0.031
mRNA promoter vs. intron	0.035
lncRNA promoter vs. intron	0.027

significantly hits in LNC gene promoters was significantly higher than 3’ UTRs and introns, while it was not different from the coding gene promoters. Table 15 showed that frequency distribution of motif hits for all the predicted TFBSs has no significant difference between the mRNA and lncRNA promoters (*P*-value of paired t-test: 0.157). Nevertheless, in comparison to 3’ UTR regions or introns of mRNA, lncRNA promoters showed significantly difference (*P*-value lower than 0.05) of motif-hit frequency distribution from those two types of sequences (*P*-value: 0.031 and 0.027, respectively) which behaved like the distribution calculated from mRNA promoters. In summary, these results provided evidences to the identified co-expressed clusters by showing that the promoters of coding genes in co-expressed clusters share motifs that were similar to the annotated TF PFMs.

6.3 Co-occurrence of TF binding motifs in the promoter regions of co-expressed coding and non-coding genes

The discovered motifs were examined to see whether they are also present in the promoters of LNC genes in the same cluster. By adopting the procedure described in section 6.4.4, we identified 12 co-expressed C-LNC clusters that shared at least one *cis*-element in both of coding and lncRNA promoters (60% of the 20 clusters with at least one co-expressed lncRNA). Though these *cis*-elements were dicovered from coding gene promoters, they were statistically enriched in their co-expressed lncRNA promoters. This suggested the possibility that some of the co-expressed C-LNC gene clusters might be co-regulated. This phenomina was majorly observed in the stages from L3 to male adlut. In section 4.3.4, it was found that most of the lncRNA co-expressed with other lncRNAs are associated with the stages from L3 to male adlut. The situation hold still when the co-expressed partners of lncRNAs changed to coding genes. However, futher investigations are needed for these unique co-expressed C-LNC genes that share one or more common *cis*-elements.

doi:10.6342/NTU201601384

Table 16. Summary of *cis*-elements shared by co-expressed coding and long non-coding genes

Group ID	Cluster ID	Number of transcripts	Number of lncRNA promoters without ‘Ns’ (517 in total)	Number of mRNA promoters without ‘Ns’ (2776 in total)	# of scanned motifs enriched in lncRNA promoters
Group 1	**Clu01**	**72**	**8**	**63**	**1**
Group 1	**Clu02**	**147**	**23**	**124**	**1**
Group 1	**Clu03**	**90**	**14**	**76**	**3**
Group 1	Clu04	72	19	53	0
Group 1	**Clu05**	**363**	**51**	**311**	**5**
Group 1	**Clu06**	**57**	**19**	**38**	**1**
Group 1	Clu07	57	10	47	0
Group 2	**Clu08**	**140**	**30**	**110**	**2**
Group 2	**Clu09**	**260**	**38**	**219**	**1**
Group 2	**Clu10**	**50**	**19**	**31**	**2**
Group 2	**Clu11**	**686**	**68**	**616**	**3**
Group 2	Clu12	198	35	161	0
Group 2	**Clu13**	**77**	**23**	**54**	**1**
Group 2	Clu14	91	24	66	0
Group 2	Clu15	59	25	34	0
Group 3	Clu16	55	15	40	0
Group 3	**Clu17**	**70**	**24**	**46**	**1**
Group 3	Clu18	97	26	71	0
Group 3	**Clu19**	**156**	**44**	**111**	**3**
Group 4	Clu20	51	2	49	0
Group 5	Clu21	73	0	73	–
Group 5	Clu22	63	0	63	–
Group 5	Clu23	83	0	83	–
Group 5	Clu24	50	0	50	–
Group 5	Clu25	80	0	80	–
Group 6	Clu26	50	0	50	–
Group 7	Clu27	57	0	57	–

6.4 Materials and methods for the proposed workflow

6.4.1 Collection of annotated transcription factor binding sites

To evaluate the quality of the motifs discovered and to find potential transcription factors for each co-expressed cluster, we collected known motifs (PFMs) from the JASPAR [20], the TRANSFAC [21] and Fly Factor Survey (FFS) database [22] for comparison. In total, 815 motifs were included in the final list of known motifs. Additionally, 2,059 annotated binding sites were collected from REDfly database [90] (170 TFs and 2,048 target genes included) for defining gene promoters, conservation analysis, and serving as the validation set to perform parameter tuning at *de novo* motif discovery procedure.

6.4.2 Hierarchical clustering

To form co-expressed C-LNC gene clusters, hierarchical clustering was applied on the expression profiles of all transcripts in the developmetnal stages. By the method described in section 4.4.2.1, the expression profiles of 30,306 mRNA from FlyBase and 4,599 lncRNA were constructed using the time-course RNA-seq datasets of 30 developmental stages [6] (which were also used in the previous chapters). First of all, we filtered out the transcripts showing constent expression along with the development stages, since only the transcripts differentially expressed between at least two developmental stages were of interest. The filtered transcript set contained 2,926

doi:10.6342/NTU201601384

lncRNAs and 26,582 mRNAs in total. Co-expressed coding and long non-coding (LNC) gene clusters were then constructed by applying Hierarchical clustering on the filtered expression profiles. Specifically, the filtered expression (RPKM) profiles of 30 time points were loaded into R 3.1.0 and clustered by hierarchical clustering (R package: amap) utilizing complete linkage and Pearson's correlation coefficient. Additionally, only a cluster containing at least 50 transcripts was selected for the subsequent analysis.

6.4.3 *De novo* motif discovery of cis-elements from co-expressed coding gene promoters

We utilized eTFBS [91] to discover gapped or ungapped TFBSs in the promoters (upstream 500 bp and downstream 200 bp from transcription start site) of co-expressed "coding" genes in each selected cluster (positive promoter set). In the discovery procedure, a negative promoter set was constructed for each cluster by randomly selecting promoters of genes that are not included in the positive promoter set. A list of top-10 putative TFBSs for each cluster was reported. The reported motifs were ranked by an adjusted measurement (showed as *Eq 1*) according to the ranking scheme provided by eTFBS [91].

$$S_{pattern} = S_d \times \left(S_p\right)^a \times (S_n)^b \times (S_c)^c \qquad (Eq\ 1)$$

, where $S_{pattern}$ denotes the final pattern score for a discovered pattern. There are four components incorporated to form the pattern score. The first two components, S_d and S_p

represent two-sample proportion test score (by comparing pattern support between positive and negative promoters) and pattern position score, respectively [91]. The third, S_n is the newly added component by this thesis, which is the averaged score for each pattern calculated from the nucleosome occupancy scores provided by Kaplan *et al*. [92]. Last, S_c is the conservation score in single base resolution collected from UCSC genome browser (UCSC-15-way alignment). As for *a*, *b* and *c*, they are the relative weights given to the position score, nucleosome occupancy, and conservation score, respectively. In this thesis, the parameter set (*a*, *b*, *c*) = (0, 1, 3) was used for *de novo* motif discovery.

6.4.4 Identification of shared *cis*-element in the co-expressed lncRNA promoters

For each cluster, the discovered TFBSs were mapped onto the LNC gene promoters to see whether the discovered motifs from the coding gene promoters could be also found in the co-expressed LNC gene promoters. Next, to investigate whether the discovered TFBSs are enriched in the LNC gene promoters, the propotions of the hit LNC gene promoters in all LNC promoters, and the hit gene promoters in all gene promoters were calculated, respectively. Then, the one-tailed two-sample proportion test was adopted to find the enriched TFBSs. Howeve, for some of the TFBSs, the number of the hit LNC gene promoters was less than 5. In this case, the *P* value might be mis-caculated, since the propotion test is based on Chi-square distribution. For this kind of TFBSs, we used

doi:10.6342/NTU201601384

Fisher's excat test instead for the calculation. A cutoff of P value < 0.05 was adopted for identifying the TFBSs that are enriched in the LNC gene promoters. If a cluster contained at least one enriched motif in the co-expressed LNC gene promoters, we denoted this cluster as a potential co-regulated C-LNC gene cluster.

doi:10.6342/NTU201601384

CHAPTER 7 Limitations of this work

In the curated list, we observed that there are some lncRNA transcripts from different sources partially sharing common genomic regions. These lncRNA transcripts might be in fact the same lncRNA, might be different splicing forms of a single lncRNA gene, or might be actually independent lncRNA genes. We realized that it remained difficult to learn the fact and determine the exact boundaries for these putative lncRNAs based on the limited information collected so far. Before a mature methodology can be developed, manual examination on RNA-seq data in a genome browser is highly recommended. We highlighted the overlap information in Additional File 2 of the published work [36] to remind the readers that more investigations on such lncRNAs are needed. In addition, we also observed that the types of lncRNA transcripts (exonic, intronic, or intergenic lncRNAs) would potentially be changed once the annotation of protein-coding genes is updated. As the loci and boundaries of protein-coding genes continue to be refined, noncoding RNAs originally classified as intergenic may be found to be exonic, intronic or even become a new splicing form of a coding gene. Some of the Young *et al*. lincRNAs have been found by a follow-up FlyBase analysis (FBrf0220965) to overlap UTRs and are probably not lncRNAs. Therefore, the readers should be aware that the number of exonic sense lncRNAs in the curated list might be inflated by these

doi:10.6342/NTU201601384

lncRNAs.

Again, by the follow-up FlyBase analysis (FBrf0220965), some of the Young *et al*. lincRNAs have been found to actually consist of two or more independent lncRNA genes which map to opposite strands. We observed that the characterization process performed in the present study failed to clarify these cases based on the stranded RNA-seq data collected so far. In this regard, the readers should be aware that such complicated cases were not easily to be discovered automatically by the proposed computational approach, and might be still present in the remaining 583 Young *et al*. lincRNA genes curated in the list.

doi:10.6342/NTU201601384

CHAPTER 8 Conclusions and Future Directions

In this thesis, I have developed a procedure to discover novel lncRNAs using RNA-seq technology, and used a large number of RNA-seq datasets as well as lncRNA databases and ChIP-seq datasets to improve the annotation of lncRNAs in fruit fly. From these efforts, I have provided an enlarged set of *D. melanogaster* lncRNAs, including known lncRNAs and novel lncRNAs from the two tissue-specific RNA-seq datasets generated in this thesis. The novel lncRNAs I identified suggests that many fruit fly lncRNAs remain to be identified. In order to discover lncRNAs that do not contain poly(A) tails, I have developed a computational approach to identify novel lncRNAs by integrating sequencing read datasets from two different library construction protocols, the poly(A)-enriched and ribo-zero protocols. This approach can be applied to future studies for the same purpose. Moreover, I have also improved the annotation of the curated lncRNAs regarding transcriptional direction, exon regions, classification, expression in the brain, possession of a poly(A) tail, and presence of conventional chromatin signatures by utilizing the strand-specific RNA-seq and the ChIP-seq datasets from the modENCODE database and data from the present study. Through RT-qPCR experiments, we demonstrate that RNA-seq is a reliable platform to discover lncRNAs. In summary, the present study provided a solid foundation for studying the functions of lncRNAs in *Drosophila*.

With the improved annotation of transcriptional direction, researchers can investigate the co-expression relationships between lncRNAs and coding genes in order to further understand the functional roles of the set of curated lncRNAs. In conclusion, the present study has integrated many RNA-seq and ChIP-seq datasets to increase the compilation breadth and annotation detail of lncRNAs. The set of curated lncRNAs along with improved annotation serves as an important resource in lncRNA studies.

doi:10.6342/NTU201601384

REFERENCE:

1. Batista, P.J. and H.Y. Chang, *Long noncoding RNAs: cellular address codes in development and disease.* Cell, 2013. **152**(6): p. 1298-307.
2. Wapinski, O. and H.Y. Chang, *Long noncoding RNAs and human disease.* Trends Cell Biol, 2011. **21**(6): p. 354-61.
3. Deng, X. and V.H. Meller, *roX RNAs are required for increased expression of X-linked genes in Drosophila melanogaster males.* Genetics, 2006. **174**(4): p. 1859-66.
4. Zhao, Y., et al., *NONCODE 2016: an informative and valuable data source of long non-coding RNAs.* Nucleic Acids Res, 2016. **44**(D1): p. D203-8.
5. Young, R.S., et al., *Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome.* Genome Biol Evol, 2012. **4**(4): p. 427-42.
6. Graveley, B.R., et al., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-9.
7. Gullerova, M. and N.J. Proudfoot, *Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells.* Nat Struct Mol Biol, 2012. **19**(11): p. 1193-201.
8. Hobson, D.J., et al., *RNA polymerase II collision interrupts convergent transcription.* Mol Cell, 2012. **48**(3): p. 365-74.
9. Sigova, A.A., et al., *Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2876-81.
10. Gonzalez, E. and S. Joly, *Impact of RNA-seq attributes on false positive rates in differential expression analysis of de novo assembled transcriptomes.* BMC Res Notes, 2013. **6**: p. 503.
11. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.
12. Gierlinski, M., et al., *Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment.* Bioinformatics, 2015. **31**(22): p. 3625-30.
13. Yang, J.H., et al., *ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data.* Nucleic Acids Res, 2013. **41**(Database issue): p. D177-87.

doi:10.6342/NTU201601384

14. Schlitt, T. and A. Brazma, *Current approaches to gene regulatory network modelling.* BMC Bioinformatics, 2007. **8**.

15. Adryan, B. and S.A. Teichmann, *The developmental expression dynamics of Drosophila melanogaster transcription factors.* Genome Biol, 2010. **11**(4): p. R40.

16. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-151.

17. Tsai, H.K., et al., *MYBS: a comprehensive web server for mining transcription factor binding sites in yeast.* Nucleic Acids Research, 2007. **35**: p. W221-W226.

18. Badis, G., et al., *A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters.* Molecular Cell, 2008. **32**(6): p. 878-887.

19. Zhu, C., et al., *High-resolution DNA-binding specificity analysis of yeast transcription factors.* Genome Research, 2009. **19**(4): p. 556-566.

20. Mathelier, A., et al., *JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2016. **44**(D1): p. D110-5.

21. Wingender, E., *The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.* Brief Bioinform, 2008. **9**(4): p. 326-32.

22. Enuameh, M.S., et al., *Global analysis of Drosophila Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants.* Genome Res, 2013. **23**(6): p. 928-40.

23. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

24. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.

25. Graveley, B.R., et al., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-479.

26. Chintapalli, V.R., J. Wang, and J.A.T. Dow, *Using FlyAtlas to identify better Drosophila melanogaster models of human disease.* Nature Genetics, 2007. **39**(6): p. 715-720.

27. Hooper, S.D., et al., *Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis.* Mol Syst Biol, 2007. **3**.

28. Pisarev, A., et al., *FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution.* Nucleic Acids Research, 2009. **37**: p. D560-D566.

29. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.

doi:10.6342/NTU201601384

30. Roy, S., et al., *Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-1797.

31. MacArthur, S., et al., *Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions.* Genome Biology, 2009. **10**(7).

32. Massie, C.E. and I.G. Mills, *ChIPping away at gene regulation.* Embo Reports, 2008. **9**(4): p. 337-343.

33. Hoffman, B.G. and S.J.M. Jones, *Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing.* Journal of Endocrinology, 2009. **201**(1): p. 1-13.

34. Moran, I., et al., *Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes.* Cell Metab, 2012. **16**(4): p. 435-48.

35. Ilott, N.E. and C.P. Ponting, *Predicting long non-coding RNAs using RNA sequencing.* Methods, 2013. **63**(1): p. 50-9.

36. Chen, M.J., et al., *Integrating RNA-seq and ChIP-seq data to characterize long non-coding RNAs in Drosophila melanogaster.* BMC Genomics, 2016. **17**(1): p. 220.

37. Schuettengruber, B., et al., *Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos.* PLoS Biol, 2009. **7**(1): p. e13.

38. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

39. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells.* Cell, 2007. **130**(1): p. 77-88.

40. Navarro, P., et al., *Molecular coupling of Xist regulation and pluripotency.* Science, 2008. **321**(5896): p. 1693-5.

41. Donohoe, M.E., et al., *The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting.* Nature, 2009. **460**(7251): p. 128-32.

42. Nesterova, T.B., et al., *Pluripotency factor binding and Tsix expression act synergistically to repress Xist in undifferentiated embryonic stem cells.* Epigenetics Chromatin, 2011. **4**(1): p. 17.

43. Okazaki, Y., et al., *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.* Nature, 2002. **420**(6915): p. 563-573.

44. Cawley, S., et al., *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell, 2004. **116**(4): p. 499-509.

45. Ravasi, T., et al., *Experimental validation of the regulated expression of large*

numbers of non-coding RNAs from the mouse genome. Genome Research, 2006. **16**(1): p. 11-19.

46. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nat Rev Genet, 2009. **10**(3): p. 155-9.
47. Ponting, C.P., P.L. Oliver, and W. Reik, *Evolution and Functions of Long Noncoding RNAs.* Cell, 2009. **136**(4): p. 629-641.
48. Wang, K.C. and H.Y. Chang, *Molecular Mechanisms of Long Noncoding RNAs.* Molecular Cell, 2011. **43**(6): p. 904-914.
49. Quinn, J.J. and H.Y. Chang, *Unique features of long non-coding RNA biogenesis and function.* Nature Reviews Genetics, 2016. **17**(1): p. 47-62.
50. Fatica, A. and I. Bozzoni, *Long non-coding RNAs: new players in cell differentiation and development.* Nature Reviews Genetics, 2014. **15**(1): p. 7-21.
51. Lee, C. and N. Kikyo, *Strategies to identify long noncoding RNAs involved in gene regulation.* Cell and Bioscience, 2012. **2**.
52. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & Development, 2011. **25**(18): p. 1915-1927.
53. Wang, Y., et al., *De novo prediction of RNA-protein interactions from sequence information.* Molecular Biosystems, 2013. **9**(1): p. 133-142.
54. Nacher, J.C. and N. Araki, *Structural characterization and modeling of ncRNA-protein interactions.* Biosystems, 2010. **101**(1): p. 10-19.
55. Guo, X.L., et al., *Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks.* Nucleic Acids Research, 2013. **41**(2).
56. Brown, J.B., et al., *Diversity and dynamics of the Drosophila transcriptome.* Nature, 2014. **512**(7515): p. 393-9.
57. Schuettengruber, B., et al., *Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos.* PLoS Biol, 2009. **7**(1): p. e1000013.
58. Wu, S.C., E.M. Kallin, and Y. Zhang, *Role of H3K27 methylation in the regulation of lncRNA expression.* Cell Res, 2010. **20**(10): p. 1109-16.
59. Sun, Q.W., et al., *R-Loop Stabilization Represses Antisense Transcription at the Arabidopsis FLC Locus.* Science, 2013. **340**(6132): p. 619-621.
60. Yang, F., et al., *Repression of the Long Noncoding RNA-LET by Histone Deacetylase 3 Contributes to Hypoxia-Mediated Metastasis.* Molecular Cell, 2013. **49**(6): p. 1083-1096.
61. Jiang, Q.H., et al., *TF2LncRNA: Identifying Common Transcription Factors for a List of lncRNA Genes from ChIP-Seq Data.* Biomed Research International,

doi:10.6342/NTU201601384

2014.

62. dos Santos, G., et al., *FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations.* Nucleic Acids Res, 2015. **43**(Database issue): p. D690-7.

63. Karolchik, D., et al., *The UCSC Genome Browser database: 2014 update.* Nucleic Acids Res, 2014. **42**(Database issue): p. D764-70.

64. Matthews, B.B., et al., *Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data.* G3 (Bethesda), 2015. **5**(8): p. 1721-36.

65. Xie, C., et al., *NONCODEv4: exploring the world of long non-coding RNA genes.* Nucleic Acids Res, 2014. **42**(Database issue): p. D98-103.

66. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinf, 2009. **10**: p. 421.

67. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.

68. Kong, L., et al., *CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W345-9.

69. Yang, L., et al., *Genomewide characterization of non-polyadenylated RNAs.* Genome Biol, 2011. **12**(2): p. R16.

70. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

71. Livyatan, I., et al., *Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation.* Nucleic Acids Res, 2013. **41**(12): p. 6300-15.

72. Novikova, I.V., S.P. Hennelly, and K.Y. Sanbonmatsu, *Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure?* Bioarchitecture, 2012. **2**(6): p. 189-99.

73. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**(9): p. 1775-89.

74. Wang, F., et al., *Characteristics of long non-coding RNAs in the Brown Norway rat and alterations in the Dahl salt-sensitive rat.* Sci Rep, 2014. **4**: p. 7146.

75. Flynn, R.A. and H.Y. Chang, *Long noncoding RNAs in cell-fate programming and reprogramming.* Cell Stem Cell, 2014. **14**(6): p. 752-61.

76. Washington, N.L., et al., *The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details.* Database (Oxford), 2011. **2011**: p. bar023.

77. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

78. Roberts, A. and L. Pachter, *Streaming fragment assignment for real-time analysis of sequencing experiments.* Nat Methods, 2013. **10**(1): p. 71-3.

79. Karolchik, D., et al., *The UCSC Genome Browser database: 2014 update.* Nucleic Acids Research, 2014. **42**(D1): p. D764-D770.

80. St Pierre, S.E., et al., *FlyBase 102-advanced approaches to interrogating FlyBase.* Nucleic Acids Research, 2014. **42**(D1): p. D780-D788.

81. Hansen, K.D., S.E. Brenner, and S. Dudoit, *Biases in Illumina transcriptome sequencing caused by random hexamer priming.* Nucleic Acids Res, 2010. **38**(12): p. e131.

82. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.

83. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.* Nucleic Acids Res, 2013. **41**(6): p. e74.

84. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC Bioinf, 2012. **13**: p. 134.

85. Butler, J.E. and J.T. Kadonaga, *The RNA polymerase II core promoter: a key component in the regulation of gene expression.* Genes Dev, 2002. **16**(20): p. 2583-92.

86. Pedersen, A.G., et al., *The biology of eukaryotic promoter prediction--a review.* Comput Chem, 1999. **23**(3-4): p. 191-207.

87. Lee, D.H., et al., *Functional characterization of core promoter elements: the downstream core element is recognized by TAF1.* Molecular and Cellular Biology, 2005. **25**(21): p. 9674-9686.

88. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.

89. Chen, F., X. Gao, and A. Shilatifard, *Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide.* Genes & Development, 2015. **29**(1): p. 39-47.

90. Gallo, S.M., et al., *REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila.* Nucleic Acids Res, 2011. **39**(Database issue): p. D118-23.

91. Chen, C.Y., et al., *Discovering gapped binding sites of yeast transcription factors.* Proc Natl Acad Sci U S A, 2008. **105**(7): p. 2527-32.

92. Kaplan, N., et al., *The DNA-encoded nucleosome organization of a eukaryotic genome.* Nature, 2009. **458**(7236): p. 362-U129.

doi:10.6342/NTU201601384

APPENDIX

List of Publications

Journal paper

1. **Chen MJM**[#], Chen LK[#], Lai YS, Lin YY, Wu DC, Tung YA, Liu KY, Shih HT, Chen YJ, Lin YL, Ma LT, Huang JL, Wu PC, Hong MY, Chu FH, Wu JT*, Li WH*, Chen CY*. (2016) Integrating RNA-seq and ChIP-seq Data to Characterize Long Non-coding RNAs in *Drosophila melanogaster*, *BMC Genomics* 17(1):220. ([#]authors with equal contribution)
2. Hsu JC*, Lin YY, Chang CC, Hua KH, **Chen MJM**, Huang LH, Chen CY*. (2016) Discovery of Organophosphate Resistance-Related Genes in Well-known Resistance Mechanisms of the Diamondback Moth (Plutella xylostella) by RNA-Seq, *Journal of Economic Entomology* pii: tow070.
3. Lin KI*, Hung KH, Su ST, Chen CY, Hsu PH, Wu PC, Chen HY, Lin FR, Tsai MD, Huang SY, Wu WJ, **Chen MJM**. (2016) Aiolos collaborates with Blimp-1 to regulate the survival of multiple myeloma cells, *Cell Death and Differentiation*, doi:10.1038/cdd.2015.167.
4. Kuo TCY, Hu CC, Chien TY, **Chen MJM**, Feng HT, Chen LFO, Chen CY, Hsu JC. (2015) Discovery of genes related to formothion resistance in oriental fruit fly (*Bactrocera dorsalis*) by a constrained functional genomics analysis, *Insect molecular biology* 24(3): 338-347.
5. Chen WY, Shih HT, Liu KY, Shih ZS, Chen LK, Tsai TH, **Chen MJ**, Liu H, Tan BCM, Chen CY, Lee HH, Loppin B, Aït−Ahmed O, Wu JT*. (2015) Intellectual disability−associated dBRWD3 regulates gene expression through inhibition of HIRA/YEM−mediated chromatin deposition of histone H3. 3, *EMBO reports* e201439092.
6. Rajendran SK, Lin IW, **Chen MJM**, Chen CY, Yeh KW*. (2014) Differential activation of sporamin expression in response to abiotic mechanical wounding and biotic herbivore attack in the sweet potato, *BMC plant biology* 14:112.
7. Meyer P*, Cokelaer T, Chandran D, Kim KH, Loh PR, Tucker G, Lipson M, Berger B, Kreutz C, Raue A, Steiert B, Timmer J, Bilal E, **DREAM 6&7 Parameter Estimation consortium**, Sauro HM, Stolovitzky G and Saez-Rodriguez J*. (2014) Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach, *BMC Systems Biology* 8:13 (**Chen MJM** is one of the authors listed in DREAM 6&7 Parameter

Estimation consortium).

8. Meyer P*, Siwo G, Zeevi D, Sharon E, Norel R, **DREAM6 Promoter Prediction Consortium**, Segal E and Stolovitzky G. (2013) Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach, *Genome research* 23: 1928-1937 (**Chen MJM** is listed in DREAM6 Promoter Prediction Consortium).
9. Hsu JC, Chien TY$^{\#}$, Hu CC$^{\#}$, **Chen MJM**$^{\#}$, Wu WJ, Feng HT, Haymer DS and Chen CY*. (2012) Discovery of Genes Related to Insecticide Resistance in *Bactrocera dorsalis* by Functional Genomic Analysis of a *De Novo* Assembled Transcriptome, *PLoS ONE* 7(8): e40950. ($^{\#}$authors with equal contribution)
10. **Chen MJM**, Chou LC, Hsieh TT, Lee DD, Liu KW, Yu CY, Oyang YJ, Tsai HK* and Chen CY*. (2012) *De novo* motif discovery facilitates identification of interactions between transcription factors in *Saccharomyces cerevisiae*. *Bioinformatics* 1;28 (5): 701-708.
11. Liu HC, Shih LY, **Chen MJM**, Wnag CC, Yeh TC, Lin TH, Chen CY, Lin CJ and Liang DC*. (2011) Expression of HOXB genes is significantly different in acute myeloid leukemia with a partial tandem duplication of MLL vs. a MLL translocation: a cross-laboratory study. *Cancer Genetics* 204 (5): 252-259.
12. Liu LY, Chen CY, **Chen MJM**$^{\#}$, Tsai MS$^{\#}$, Lee CHS$^{\#}$, Phang TL, Chang LY, Kuo WH, Hwa HL, Lien HC, Jung SM, Lin YS, Chang KJ and Hsieh FJ*. (2009) Statistical identification of gene association by CID in application of constructing ER regulatory network. *BMC Bioinformatics* 10: 85**.**
13. Chen CY, Tsai HK, Hsu CM, **Chen MJM**, Hung HG, Huang GTW, Li WH*. (2008) Discovering gapped binding sites of yeast transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* 105: 2527-2532**.**

Oral presentation in conference

1. **Chen MJM**, Chou LC, Hsieh TT, Lee DD, Liu KW, Yu CY, Oyang YJ, Tsai HK* and Chen CY*. *De novo* motif discovery facilitates identification of interactions between transcription factors in *Saccharomyces cerevisiae*. October 19-21, 2012; *International Symposium on Evolutionary Genomics and Bioinformatics (ISEGB)*, Kaohsiung, Taiwan

Conference poster

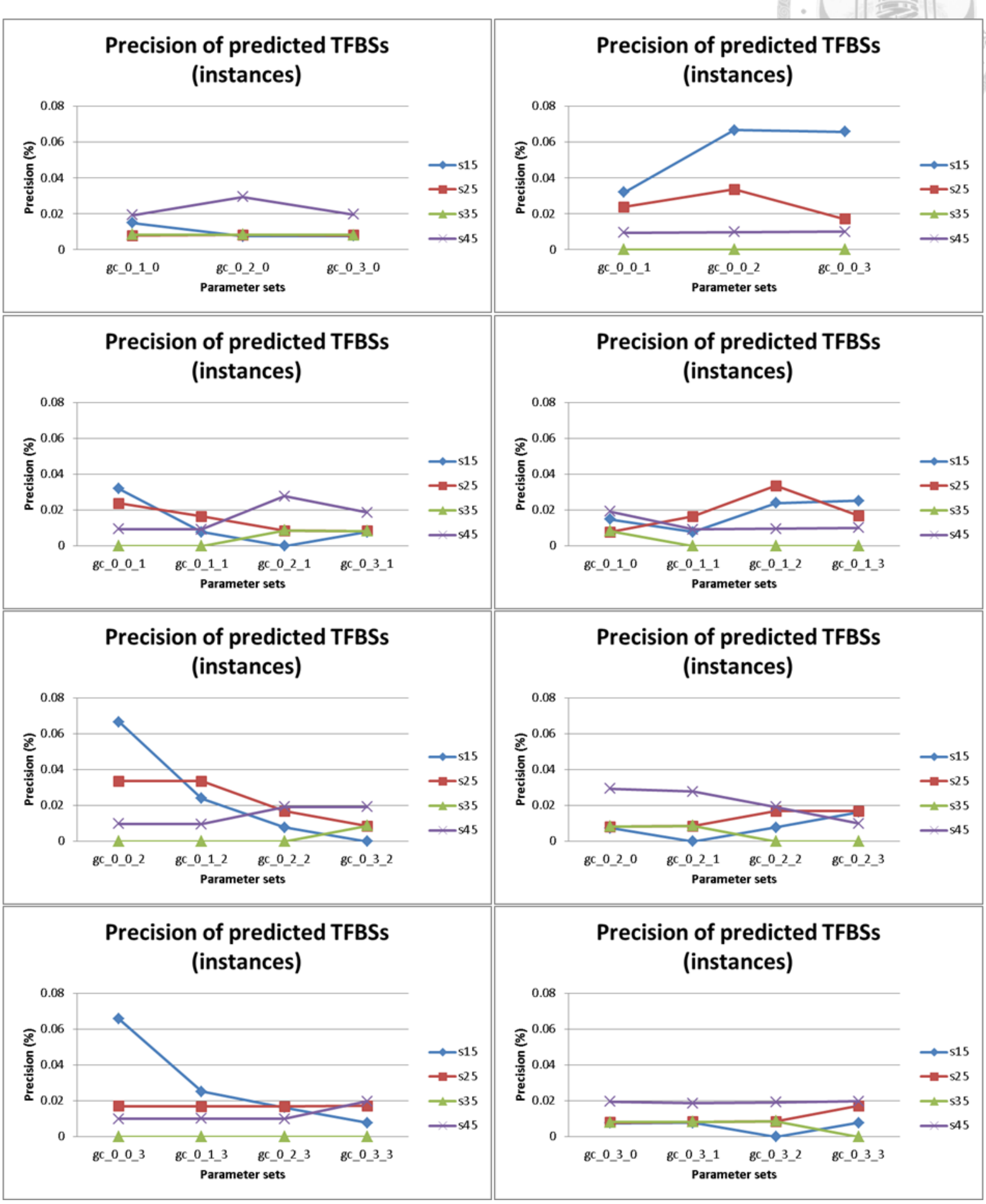
1. **Chen MJM**, Lin YY, Li WH, and Chen CY. Common cis-elements suggests co-regulation of coding and long non-coding genes in *Drosophila melanogaster*. Poster; September 16-20, 2016; *17th edition of International Conference on Systems Biology (ICSB)*, Barcelona, Spain.

doi:10.6342/NTU201601384

2. **Chen MJM**, Su YR, Chang P, Hong TR, Cherng BW, Tung YA and Chen CY. Potential of lncRNA to regulate gene expression through promoter binding in *Drosophila Melanogaster*. Poster; September 3-7, 2016; *15th European Conference on Computational Biology (ECCB),* The Hague, Netherlands
3. Poelchau M, Childers C, Moore G, Tsavatapalli V, Pieper U, **Chen MJM** and Lin YY. The i5k Workspace@NAL - Updates and new developments of an arthropod genome portal. Poster; January 9-13, 2016; *Plant and Animal Genome Conference XXIV (PAG),* San Diego, USA
4. **Chen MJM**, Lin YY, Li WH and Chen CY. Transcriptional regulation of long non-coding gene expression in *Drosophila melanogaster*: a genome-wide study using RNA-seq. Poster B39 in the Category of Gene Expression; September 7-10, 2014; *13th European Conference on Computational Biology (ECCB),* Strasbourg, France
5. Chang C, **Chen MJM**, Kuo T, Huang JL, Haymer DS, Hsu JC and Chen CY. Improving completeness of *de novo* transcriptome assembly and gene annotation by multi-species transcriptome sequencing in fruit fly genus *Bactrocera*. Poster A74 in the Category of Sequencing and Sequence Analysis for Genomics; September 7-10, 2014; *13th European Conference on Computational Biology*, Strasbourg, France
6. Lin YY, **Chen MJM** and Chen CY. A study of inter- and intra-protein corelated mutations on highly similar protein sequences. Poster E72 in the Category of Structural Bioinformatics; September 7-10, 2014; *13th European Conference on Computational Biology*, Strasbourg, France
7. **Chen MJ**, Hu CC, Hsu JC, Chen CY. Different transcript segments reveal consistent information about expression ratios in RNA-seq. Poster 91 in the Systems Biology Track; October 14-19, 2011; *4th RECOMB Conference on Regulatory Genomics, Systems Biology, and DREAM Challenges*, Barcelona, Spain
8. Chien TY, **Chen MJ**, Chen CY. Multi-level determination of confidence scores for alternatively spliced mRNA transcripts discovered by de novo assembler; October 14-19, 2011; *4th RECOMB Conference on Regulatory Genomics, Systems Biology, and DREAM Challenges*, Barcelona, Spain
9. Tung YA, Chen YS, **Chen MJM**, Chen CY. Predicting promoter activities by non-linear combination of sequence motifs; October 14-19, 2011; *4th RECOMB Conference on Regulatory Genomics, Systems Biology, and DREAM Challenges*, Barcelona, Spain
10. **Chen MJ**, Chen CY. Improving network completeness of yeast transcriptional interaction network by predicted TF-TF interactions; July 17 to 19, 2011; *19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 10th European Conference on Computational Biology (ECCB)*, Vienna, Austria

11. Wang YT, **Chen MJ**, Wu HY, Tsai CF, Hong TM, Yen CJ, Chen YJ. Personalized Tissue Phosphoproteomics Screening of Human Hepatocellular Carcinoma for Drug Target Discovery in Cancer Therapy. Arpil 27 to 28, 2011. 2011 *Translational Medicine Conference and Taiwan Proteomics Society Annual Symposium (TPS)*,Taiwan
12. Wu PC, **Chen MJM**, Chen CY. Exploiting Cross-Species Conservation to Improve Prediction Accuracy of Discovering Transcription Factor Binding Sites. Poster 44; Aug. 16 to 18, 2010; *9th Annual International Conference on Computational Systems Bioinformatics (CSB)*, Stanford, California
13. Hsieh TT, **Chen MJM**, Chen CY. Investigating Consistency between Curated Binding Profiles and PWMs Derived from Protein-DNA Structure Models. Poster 45; Aug. 16 to 18, 2010; *9th Annual International Conference on Computational Systems Bioinformatics (CSB)*, Stanford, California
14. **Chen MJM**, Chen CY. Identification of Transcription Factor Interacting Pairs by Mining ChIP-chip Data. O) Regulation; 2008 July 19 to 23; *16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Toronto

Appendix Figures



Appendix Figure 1. Parameter tuning for different pattern supports (ratio of pattern-hit promoters/all promoters in the positive set) and different weights used for pattern ranking. Precision is calculated by the ratio of (True Positives/Predicted instances), and presented as percentage.

Appendix Tables

Appendix Table 1. Primer list of the selected lncRNAs for RT-qPCR experiments

ID	5' primer	3' primer	Experiment results
TCONS_00031380	AGTCCTTCGAAACAAACTGTCT	TTGGTAAACAATGCGGCAATAC	Figure 13 (a)
TCONS_00028095	ATACATTGTGCCAAAATAGCCG	AATTCACAGCCCTTCTTAGCAT	Figure 13 (a)
TCONS_00044977	TCGATGATTCTACGGTCAAGTT	TTTTTGTTTGCCGAACATCTCG	Figure 13 (a)
TCONS_00037494	AGCCTATGGACAAGGACATCTA	TATGATGTGTAATTGGTCGGCA	Figure 13 (a)
TCONS_00048859	CCACTTAAAGGAGGCGATCTTC	AAGATGCTGAGGATATGGATGC	Figure 13 (a)
TCONS_00051944	ATCCGGATATTCGACCTTGTTG	ATTTTAGTTGCGCTTGCTGTTC	Figure 13 (a)
TCONS_00020613	GAAAAGGCAGCAAGTGTTACAA	ACCAAACTGCTGGTATCGTTAT	Figure 13 (a)
TCONS_00033121	GCTTCGATCATTTCGCGTATC	CCACTAGCGATGATGGTGAAAG	Figure 13 (a)
TCONS_00017414	TCGCTGACGACAAAATCCTTAT	TACGTTTACTTTTCGTGAGGCT	Figure 13 (a)
TCONS_00050427	ATCCAGATGCCAGAATTCACC	ATGTGGATGTGACCTGAATCAC	Figure 13 (a)
TCONS_00032409	GTGTCGTGCTACATGTGTTTAC	GAGAAGAAAACAAGGTGCTGTG	Figure 13 (a)
TCONS_00036092	ATTTCCATTGTTGTTGCCATGC	CGGCGGTCCAATACAAACAATA	Figure 13 (a)
TCONS_00044754	GGAACTAGGGGCATTTAGTTGT	CAACATATGCGGAGGGATTTTG	Figure 13 (a)
TCONS_00003446	TCTTGGGCTGAGAATAATGCAA	ATATTCCAACAGCCCACTAACG	Figure 13 (a)
TCONS_00043412	CATGGCTACTCACTCAGGTAGA	CTAATGGCTTCTTGATGCGTTC	Figure 13 (a)
TCONS_00036539	ACCAACTCGGCAACAACTATAA	CTTACAGTTGCACGACAACAAC	Figure 13
TCONS_00044991	AATCGTTACACTAAACACCCGA	ACTCGCTACACATCCCTAAGTA	Figure 13
TCONS_00044992	TGACGACACATAGCTGAAAAGT	CAGAAGCTCAAGCAAATTCCTC	Figure 13
TCONS_00034204	CAGCTTGAATTGGGTCAAGTTT	CACACCAGCTGACAGTTATTTC	Figure 13
TCONS_00011851	GAACGGAACCGCAAAACTAAG	CTGCCCTTTGATGCTAAATGTC	Figure 13
FBgn0266811	TCATAATGGAACTATGCAGGCG	ATTTCAATACGTTTAGGCACGC	Figure 13 (b)
FBgn0267298	AAACACTTGAAATGGACTTGGC	TGTTCGGGTATCCTCGCTAAAT	Figure 13 (b)
Untranscribed_region1	ACTCTCGTAGAAACAATCTCGT	GCAAAAGTTAAAAGGACACAGC	Figure 14
Untranscribed_region2	CGCATTTATTATGCCATCCTCA	GTATTGATGCCGGTGTACTTTT	Figure 14
Untranscribed_region3	ATCACACGATAACAACAAAGGG	CTCCTCCGATGATTTTAGTCCT	Figure 14
G1_FBgn0083068	ATCGGACGGAAATGCAGAAG	CACTGGGAGGGCTAATGAAC	Figure 14
G1_FBgn0265590	CAAGAAGTGGAAGGGAGATGG	GACAGGCGCAACAACTAAAC	Figure 13 (b) and Figure 14
G1_TCONS_00045108	CTAACCAGACGCTCTCAGTC	CCCCTCCCTTCAAACAAGATAC	Figure 13 (a) and Figure 14
G1_FBgn0001234	CACTGGTGTATCGACTTCTCTG	GTATGTCTGCCCTTTACGGAAC	Figure 14
G1_FBgn0051144	CTAAGAGGCCGATCAGAAGG	CTTCCTACTCCATTTGTCGC	Figure 13 (b) and Figure 14
G1_FBgn0262109	TCGTAAAGGGAATCCAACGC	GATGCAATCGTCAGCGAAGTC	Figure 14
G1_FBgn0264360	ATATGCTGCTCTGCGTCTTC	TCTGTTTACGTGTTGGCGTC	Figure 13 (b) and Figure 14

doi:10.6342/NTU201601384

G1_FBgn0265071	CTTCTTCTTGCTACCCGCTTTG	TCTGCTCATAATTGCGCTCG	Figure 14
G1_FBgn0265295	GTAGTAGACGTGAGCCAAGTTC	GTTGGAGGTGCCCACAATTATC	Figure 14
G1_ROX1	ACATCAGGCCATAGCCAAGAAG	AACACGATCTACTTCTGGTCGG	Figure 13 and Figure 14
G1_ROX2	GGTCACACTAAGCTAGGGCTAC	CGGAAATCGTTACTCTTGCTTG	Figure 14
G2_FBgn0263981	CAGCTCCAGCATTTCCTTAACC	CGTACAGCTTATCCATATCGGC	Figure 14
G2_FBgn0264869	CTCGACTCAACACAATTCCGAC	CAACACGAGGTATGTTTCTCCC	Figure 14
G2_FBgn0262993	GGACAACCATAGAATGAGGGAG	CGAATGCGAGAAAGAGAGGTAG	Figure 14
G2_FBgn0265340	CCCAACCATTGATGAAGCTGTG	GTATAGTCTAACGGCGGAGATG	Figure 14
G2_FBgn0260720	CCATCACCATCTTCAATAGCCC	TGCTACATAAGCCAGTCAGTG	Figure 14
G2_TCONS_00012337	ATTTCAAGTTGCCCCCAGTC	CTCGATTTCAGGCCAAGAGAG	Figure 14
G2_lincRNA.292	CCTTCTGATAACCCTTGTGGC	GCTGATAGATACGGAAGTGGTC	Figure 14
G2_FBgn0264446	TACCTTCGCATCACTGCTTC	GGATTTGGGTTTTGGGCTTG	Figure 14
G2_FBgn0264481	CGTCATTCTCTTCCTCCGATG	GTCGTGTCTGTGTGTGCTTA	Figure 14
G2_FBgn0264504	CAAAGACTGTTCCTGCTCCTG	CCATGTTCCCAGCTTACGATTG	Figure 14
G2_FBgn0266044	GGAGTGAGTTAAGGGACAACAG	CGCTGCTGAGATTGGAGTTAG	Figure 14
G3_FBgn0264993	CTTCGATGAGCACCAGGATAC	CATGGGATTCAAGTACGACAGC	Figure 13 (b) and Figure 14
G3_FBgn0265458	CCCCAATGTCTTCGACTTACTC	CAGGAGGATCTGTTTCTGGAC	Figure 14
G3_TCONS_00045565	AGTCTAACCTGCCCACTGAA	CCAACCATTCATTCCAGCCTTC	Figure 14
G3_FBgn0262106	GTCATTCATACTGGGTCTTGCC	TCCATTTCGGGTTTGGTGAC	Figure 14
G3_FBgn0262107	ATGACCAAGAGGATGAGTCGC	GCTACTGCTGTCTATAAGGTGG	Figure 13 (b) and Figure 14
G3_FBgn0264980	CTAATTTCACTCTACCCGCCG	CTCAACTCAACCGACCCTTAC	Figure 13 (b) and Figure 14
G3_FBgn0062928	GAACCGAAAGCACCAGATCC	GGAGGAGAGTAAGCCACGTTAG	Figure 14
G3_lincRNA.354	GTGGCTATAATGATCCCGGTAG	GTGATGATCTCCCATTCTCTGC	Figure 14
G3_FBgn0263331	CGCTTGTGGGTGAAGCATTG	TGCCGCCAGAATGAGATTCC	Figure 13 (b) and Figure 14
G3_FBgn0263626	CTCTACCCCATCCATTTTCAGG	CTGTGTGCTCTGTTATGTGTCC	Figure 14
G4_FBgn0265530	CGAATCAACCAGACCCATAAGC	TGGCGATATTTGACAGACGG	Figure 14
G4_TCONS_00054835	CCCATTATCCTCTGCAAGTGTG	GAGAGTCGGAAATCGAGAATCG	Figure 14
G4_lincRNA.160	GTATGAAAAAGTGGAGCGACGG	CCCACCATCCCCTAAACAAAG	Figure 14
G4_FBgn0263380	CAATCATGGAGATGGAGGACC	CGGAGTCTTCAGTTCGAGTTC	Figure 14
G4_FBgn0264840	AAGACAGGTTAAGGCTAGTCGG	CTCATGCCGAAACACATTCG	Figure 14
G4_FBgn0265302	GCCTTCTCCAGTTTGGTATGAC	ACAATTAGCCCGACCATCTC	Figure 14
G4_TCONS_00020772	GAGTGGATAGCGGAGATTGC	GCCTTCTTGACTTCCTTCTCC	Figure 13 (a) and Figure 14
G4_FBgn0263497	ATCGAATCGGTGGTAAGTGAGG	GGAAAGTGAGCGGGTTAAAGTG	Figure 14
G4_FBgn0262963	GTTCTGGGGTCAGTTGGACT	AACCAAAGAGGGAAATGCGG	Figure 14
G4_FBgn0265085	CATCTGAACCCCAACCACTTC	GAGCACAAGCACCAACAATG	Figure 14

Appendix Table 2. Raw Ct values of RT-qPCR experiments for un-transcribed regions and the selected lncRNAs.

	RT+				RT-			
Replicates	**P1**	**P2**	**P3**	**P4**	**N1**	**N2**	**N3**	**N4**
Figure 13(a) RT-qPCR experiments for a selected set of lncRNAs in brains								
RpL32	21.89	21.92	22.04	21.97	35.73	35.23	35.29	35.55
ROX1	20.33	20.15	20.33	20.35	32.11	32.31	32.2	32.3
TCONS_00031380	26.67	26.47	26.46	26.57	32.55	32.87	32.77	NA
TCONS_00028095	26.37	26.32	26.24	26.11	NA	31.99	32.16	32.13
TCONS_00044977	28.07	27.98	27.97	28.12	NA	33.42	33.35	33.49
TCONS_00037494	NA	28.86	28.7	28.78	33.32	33.27	33.18	33.18
TCONS_00048859	28.29	28.37	28.14	27.91	32.32	32.24	32.02	31.72
TCONS_00051944	NA	33.7	33.46	33.77	37.43	37.08	36.58	36.45
TCONS_00045108	27.42	27.29	27.41	27.42	30.31	30.22	30.25	NA
TCONS_00020613	29.42	29.22	29.27	29.2	32.08	31.91	31.81	31.78
TCONS_00033121	31.7	31.57	31.47	31.39	34.13	34.05	33.47	33.73
TCONS_00017414	29.55	29.21	29.29	29.2	NA	31.6	31.62	31.55
TCONS_00050427	30.88	30.91	30.91	31.06	33.09	33.05	32.65	33.03
TCONS_00032409	30.43	30.35	30.18	30.16	NA	32.21	31.77	31.86
TCONS_00036092	30.25	30.12	29.75	29.61	31.62	31.73	31.52	31.14
TCONS_00044754	30.37	30.51	30.55	30.55	31.94	32.04	32.09	32.06
TCONS_00003446	31.42	31.65	31.32	31.36	32.68	32.88	33.03	32.88
TCONS_00043412	31.96	31.97	31.94	32.12	33.39	33.3	33.17	33.47
TCONS_00020772	26.1	25.95	26.08	26.16	27.42	26.93	26.87	27.15
TCONS_00036539	32.12	32.24	32.04	32.22	32.66	33.25	33.02	32.85
TCONS_00044991	31.11	31	31.01	31.06	31.58	31.42	31.76	NA
TCONS_00044992	NA	31.07	31.25	31.14	31.67	31.66	31.6	31.73
TCONS_00034204	31.28	31.24	31.15	31.17	31.93	31.68	31.6	31.5
TCONS_00011851	30.98	30.8	30.72	30.56	30.98	30.8	30.72	30.56
Figure 13(b) RT-qPCR experiments for a selected set of lncRNAs in brains: 2-fold amount of template brain cDNA								
RpL32	21.52	21.45	21.57	21.68	35.49	35.02	36.25	35.31
ROX1	20.03	19.79	19.86	19.65	33.18	33.08	33.18	33.27
FBgn0051144	27.26	27.05	27.07	26.92	33.16	33.65	32.63	32.94
FBgn0265590	NA	25.78	25.89	25.7	NA	31.33	31.55	31.48
FBgn0262107	25.58	25.62	25.62	25.66	31	30.98	31.13	31.04
FBgn0264360	30.27	30.19	30.11	30.16	NA	32.76	32.75	32.63
FBgn0266811	29.72	29.52	NA	29.72	31.42	31.26	30.33	30.66

FBgn0267298	32.68	32.53	32.62	NA	33.63	33.67	33.51	NA
FBgn0264980	31.32	31.39	31.39	NA	31.79	31.79	31.73	31.97
FBgn0264993	32.28	32.45	32.16	32.18	32.27	32.18	32.2	32.38
FBgn0263331	31.35	31.38	31.29	31.46	31.39	31.58	31.25	31.42
TCONS_00036539	31.91	32.23	32.02	32.26	32.98	32.66	32.92	33
TCONS_00044991	31.22	31.14	31.04	30.81	32.43	32.28	32.31	32.03
TCONS_00044992	31.05	31	30.84	30.83	NA	32.02	32.04	31.96
TCONS_00034204	31.37	31.26	31.43	31.23	32.17	32.2	31.79	32.02
TCONS_00011851	30.77	30.66	30.48	30.76	32.99	33.02	32.57	32.41

Figure 14. RT-qPCR experiments of a selected set of lncRNAs in male adults

Untranscribed_region1	33.19	33.17	33.43	33.38	33.40	33.40	33.33	33.36
Untranscribed_region2	33.40	33.89	33.55	33.73	33.79	33.58	34.02	33.80
Untranscribed_region3	33.20	33.19	33.19	33.21	33.15	33.13	33.26	33.16
G1_FBgn0083068	26.67	26.57	26.49	26.55	35.70	36.12	36.52	35.80
G1_FBgn0265590	27.39	27.39	27.36	27.41	35.17	35.13	34.18	35.32
G1_TCONS_00045108	29.33	29.36	29.08	29.11	34.20	35.07	34.81	35.13
G1_FBgn0001234	22.68	22.64	22.61	22.70	36.02	36.21	35.40	36.33
G1_FBgn0051144	27.12	27.14	27.01	27.04	33.54	NA	35.29	35.27
G1_FBgn0262109	26.91	27.03	26.80	27.04	NA	40.62	NA	39.78
G1_FBgn0264360	24.28	24.27	24.27	24.20	31.24	31.08	31.19	31.26
G1_FBgn0265071	26.47	26.46	26.42	26.35	32.25	33.23	NA	31.69
G1_FBgn0265295	26.19	26.24	26.12	26.18	34.28	34.07	35.24	34.25
G1_ROX1	23.62	23.57	23.47	23.56	38.14	NA	38.20	38.48
G1_ROX2	28.44	28.49	28.31	NA	37.70	NA	36.31	36.76
G2_FBgn0263981	26.79	26.69	26.57	26.53	37.53	NA	36.05	38.07
G2_FBgn0264869	32.12	31.71	31.70	31.69	37.19	36.72	36.58	36.05
G2_FBgn0262993	32.59	32.39	32.46	32.64	34.35	35.70	34.44	35.21
G2_FBgn0265340	29.58	29.62	29.35	29.53	33.52	33.80	33.72	34.24
G2_FBgn0260720	28.58	28.40	28.18	28.29	32.72	33.23	33.36	33.51
G2_TCONS_00012337	28.56	28.28	28.27	28.25	32.93	33.54	33.21	33.71
G2_lincRNA.292	21.14	21.14	21.21	21.19	35.42	34.48	34.40	34.66
G2_FBgn0264446	31.07	NA	31.09	31.04	32.59	31.56	32.22	32.41
G2_FBgn0264481	NA	30.00	29.83	30.00	NA	35.93	35.91	36.01
G2_FBgn0264504	29.99	NA	29.74	29.89	33.03	NA	32.32	32.99
G2_FBgn0266044	27.35	NA	27.25	27.29	32.93	33.33	NA	33.76
G3_FBgn0264993	27.11	27.17	27.06	27.04	36.93	36.59	35.89	37.22
G3_FBgn0265458	28.25	28.20	28.13	28.09	35.75	34.81	35.47	35.54
G3_TCONS_00045565	13.39	13.46	13.39	13.34	25.78	25.77	25.96	25.90

doi:10.6342/NTU201601384

G3_FBgn0262106	26.11	26.08	25.98	26.10	35.99	35.42	36.61	37.08
G3_FBgn0262107	27.02	26.89	26.67	26.76	35.96	35.01	35.10	35.01
G3_FBgn0264980	27.40	27.47	27.48	27.45	36.72	35.26	35.48	35.57
G3_FBgn0062928	25.30	25.24	25.18	25.11	34.47	34.03	34.03	34.25
G3_lincRNA.354	26.06	26.06	25.89	25.89	32.63	32.63	33.71	32.48
G3_FBgn0263331	26.48	26.45	26.46	26.42	33.08	33.20	32.47	32.17
G3_FBgn0263626	24.52	24.46	24.37	NA	27.95	27.83	27.97	28.16
G4_FBgn0265530	31.40	31.54	31.22	31.31	35.44	35.79	35.38	34.41
G4_TCONS_00054835	33.21	33.11	33.29	31.13	33.53	34.73	34.39	33.04
G4_lincRNA.160	27.14	27.10	27.02	28.53	36.44	35.51	35.98	35.52
G4_FBgn0263380	33.05	32.41	32.27	32.47	34.90	35.78	35.34	35.25
G4_FBgn0264840	28.17	28.33	28.24	28.29	35.99	36.14	NA	36.79
G4_FBgn0265302	34.08	34.29	34.19	NA	40.50	40.42	40.24	40.37
G4_TCONS_00020772	26.24	26.21	26.00	26.11	28.26	28.29	28.28	28.18
G4_FBgn0263497	31.50	31.62	31.35	NA	33.71	34.30	33.46	33.99
G4_FBgn0262963	NA	31.82	31.68	31.62	35.14	34.11	34.12	35.19
G4_FBgn0265085	29.31	29.38	29.26	29.38	35.02	35.57	34.57	36.25