2015

Investigation of the length distributions of coding and noncoding sequences in relation to gene architecture, function, and expression

Rachel Amber Caldwell
University of Wollongong, gobyfish01@gmail.com

Recommended Citation

Caldwell, Rachel Amber, Investigation of the length distributions of coding and noncoding sequences in relation to gene architecture, function, and expression, Doctor of Philosophy thesis, School of Biological Sciences and School of Mathematics and Applied Statistics, University of Wollongong, 2015. https://ro.uow.edu.au/theses/4645

Investigation of the Length Distributions of Coding and Noncoding Sequences In Relation to Gene Architecture, Function, and Expression

A thesis submitted in fulfilment of the requirements for
the award of the degree of **Doctor of Philosophy**

From
University of Wollongong

By
Rachel Amber Caldwell, B. MarSc

School of Biological Sciences
And
School of Mathematics and Applied Statistics

2015

For my mother, Glenda Rogers

Table of Contents

Abstract .... **I**

Declaration of Authorship .... **III**

Papers Already Published Out of This Thesis .... **IV**

Statement of Contribution .... **V**

Acknowledgements .... **VI**

List of Abbreviations .... **VII**

List of Figures .... **VIII**

List of Tables .... **X**

1 Introduction **2**

1.1 Background 2

1.1.1 Gene Structure 2

1.1.2 Gene Expression 3

1.1.3 Transcription 3

1.1.4 The Function of Exons and Introns 4

1.1.5 Promoters 5

1.1.6 Translation 5

1.1.7 Comparison between Prokaryotes and Eukaryotes 8

1.1.8 Protein Function 9

1.1.9 Model Organisms 11

1.1.10 Sequence Databases and Tools 13

1.2 From Sequence to Discovery – review of length distribution studies 15

1.3 Motivation and thesis outline 17

2 Data Collection of Organisms Studied **22**

2.1 Arabidopsis thaliana 24

2.1.1 *Arabidopsis thaliana* Genome 24

2.1.2 Gene Number 26

2.1.3 Coding and Untranslated regions 26

2.1.4 Gene Expression Data 26

2.2 Drosophila melanogaster 28

2.2.1 *Drosophila melanogaster* Genome 29

2.2.2 Gene Number 29

2.2.3 Coding and noncoding regions 32

2.2.4 Genome Sequence data 33

3 **Promoter Prediction in Relation to Coding and Noncoding Sequences 35**

3.1 Introduction 35

3.1.1 Currently Used Algorithms 35

3.1.2 Further Improvements in Promoter Prediction 36

3.2 Gene Expression 38

3.3 Statistical Characteristics on Quantitative Measurements 38

3.4 The Algorithms for TLS-NNPP and TSC-TSS-NNPP 43

3.4.1 Scenario 1 – TLS-NNPP Algorithm 44

3.4.2 Scenario 2 – TSC-TSS-NNPP Algorithm 46

3.5 Applications of the Algorithms TLS-NNPP and TSC-TSS-NNPP and the comparisons to NNPP2.2 48

3.5.1 *E. Coli* Sequence Study Using the TLS-NNPP Algorithm 48

3.5.2 Human Sequence Study Using the TSS-TSC-NNPP Algorithm 49

4 Coding Sequence Length Comparisons 53

4.1 Introduction 53

4.2 Data File Construction & Statistical Data Analysis 54

4.3 Empirical & Comparative Study 57

4.3.1 Protein number and density 58

4.3.2 Protein coding sequences and chromosomes 62

4.4 Discussion 71

5 Coding and Noncoding Sequence Length Comparison with *Arabidopsis* and *Drosophila* 77

5.1 Introduction 77

5.2 Statistical Analysis 78

5.3 Length Distributions among all Three Regions 79

5.4 Nonlinear Function relationship between D_1(d_1) and D_2^* (d_2^*) Values 82

5.5 Protein Function 87

5.6 Functional Protein Classification 88

5.7 Protein Function in relation to Coding and Noncoding Sequence Lengths 90

***6* Coding and Noncoding Sequences In Relation To Gene Expression – *Arabidopsis thaliana Case Study* . 96**

6.1 Introduction 96

6.2 Conventional Statistical Analysis 97

6.3 Gene expression and length distributions of coding and noncoding sequences in *Arabidopsis thaliana* 98

6.3.1 *Arabidopsis thaliana* Gene Expression and Coding and Noncoding Sequences 98

6.4 Gene expression and protein function in relation to coding and noncoding sequences 105

7 Canonical Correlation Analysis (CCA) - *Drosophila melanogaster Case Study* 113

7.1 Introduction 113

7.2 Canonical correlation analysis (CCA) 114

7.3 Pearson's Correlation 116

7.4 Discussion 119

8 Genome Comparisons using Quantile Regression Analysis between Gene Expression and Length 123

8.1 Introduction 123

8.2 Quantile regression Analysis on gene expression and length distributions in *Arabidopsis thaliana* . 123

8.3 Genome Comparisons using Quantile regression Analysis on gene expression and length distributions 128

8.4 Methods 130

8.4.1 Datasets 130

8.4.2 Statistical Analysis 131

8.4.3 Quantile Regression Analysis 132

8.5 Results – Length Subset Analysis 132

8.6 Quantile Regression Analysis 137

8.7 Discussion 150

9 Conclusion and Future Research 155

9.1 Conclusion 155

9.2 Outlook 157

References......159

Appendix A – Visual Basic Scripts......169

Appendix B – MS Excel Formulas & Macros......172

Appendix C – R and SPSS Codes......174

Abstract

The last 20 years has seen the birth of bioinformatics, and is defined as the combination of mathematics, biology, and computational approaches. This discipline has led to the era of ontology, extensive databases including sequences, structures, expression profiles, and genomes and database cross-referencing, (Ouzounis, 2012). Before this discipline, scientists referenced atlas books, such as Margret Dayhoff's protein sequence collection (Strasser, 2010) which required long hours of letter counting. Through the development of sequencing technology over the past forty years, a tremendous amount of genomic sequencing data has already been collected. With a surge of such data increasing, so does the challenges of data organisation, accessibility and interpretation, with interpretation being the most challenging (Ouzounis, 2012).

The primary structure of DNA and proteins has been predominantly the focus of sequence analysis. However, other attributes, such as sequence length are also important. The journey of gene length research commences with Zhang (2000) who conducted an investigation on protein length for three domains of life. Protein length was found to be 40-60% greater in eukaryotes than in prokaryotes (Zhang, 2000). This finding was substantiated by Xu and colleagues (2006) who found that the mean length of genetic coding sequences is highly conserved in prokaryotes and eukaryotes but diverges between the two kingdoms (Wang, 2005; Xu, et al., 2006). They reported that the coding sequence length is on average 445 bp longer in eukaryotes than in prokaryotes (Xu, et al., 2006). These findings still hold true today. Zhang's research also suggests that the differences in the length is not random, but has some biological significance (Zhang, 2000). These findings started a revolution with research now focusing on eukaryote protein size, conservation, complexity, and compactness.

As genome sequence data becomes readily available for different living organisms, and the explosion of data from biological experiments, there is a greater need for automated tools to classify and analysis this data, as well as increasing the scale and sophistication of the information technology, in order to draw conclusions from the data and to formulate new directions for research. The regulation of gene expression and its products is one of the important facets of an organism, and this regulation has been associated with different regions of the gene, including 5' and 3' un-translated regions. Variations in coding and noncoding sequence length, intron number and size differ significantly among living organisms. The main aim of this thesis is to explore and understand, using statistics and mathematical modelling, the length distribution relationship between the coding and noncoding regions of protein coding genes.

The project involved data acquirement from the internet, data formatting and creation of a database for the research, pattern search for target DNA elements, followed by the examination of the interrelationships between these regulatory elements. The research outlined in this thesis introduced a nonlinear model and incorporated gene expression data into the analysis. Other statistical methods such as Canonical Correlation Analysis (CCA) and quantile regression was used to determine the relationship of length and gene expression.

The research started with collaborations with several authors to assess neural network promoter prediction and the results found that for the *H. sapiens* data set, the TSC-TSS-NNPP method achieved better results than both NNPP2.2 and TSS-NNPP. A generalised understanding of the behaviour of the coding sequence and protein length (with and without introns) for 15 organisms was found, notably there were differences between the more complex organisms compared to the lower species. The nonlinear model has revealed a significant relationship with the coding sequence and the 5' UTR region and has complemented research that has already been investigated with these regions. Protein function was also investigated, and the results found significant differences between the available protein function classifications in relation to the coding and noncoding gene region lengths. Canonical Correlation Analysis (CCA)

was used in a *Drosophila melanogaster* study to determine a relationship between the length of the coding and noncoding regions and the gene expression levels subjected to various environmental conditions. The breakdown of the analysis showed two canonical correlation functions as being significant, and that for each dependent variable there was a weak relationship with the coding sequence. The results show the maximized correlation for each data set for each variable, was between longevity (extended life span under non-stressful conditions) and the 5' UTR length. Both of these values were negative, indicating that the higher the expression levels of longevity, the longer the length of the 5' un-translated region. However, interpretation of this method was difficult and is not widely used due to this constraint. All the work in the previous chapters has led to the most important discovery, which was the positive correlation between the 3' UTR length and gene expression. This is a unique result and was identified in both an animal and plant species.

Bioinformatics is an important discipline in the post-genomic era as it is used to convert genomics data into knowledge. Ultimately this project's goal was to discover new biological insights in the length distributions of coding and noncoding sequences and create a universal perspective on the importance of length and the relationship between the coding and noncoding sequences. The knowledge gained in this thesis can now complement and enhance other research in the areas of cancer studies (Dorairaj, et al., 2014; Mayr and Bartel, 2009; Skeeles, et al., 2013); neurodevelopmental and neurodegenerative disorders (Zylka, et al., 2015); and stress adaptation (Xue-Franzén, 2014). This research has validated the data that is publicly available on the web. It has scrutinized the data available, including gene expression data and has shown patterns not discovered previously by other research studies. In conclusion, in delving into the patterns of statistical properties of different gene regions and their correlation we intended to elucidate the spatial organization rules between various gene functional elements and the difference in such organizations among different living organisms and gene families. We believe that these rules and differences are the results of organism complexity and reflect the complexity differences in the regulation of gene expression. The information described in this thesis provides the basis for further exploration into gene regulation and architecture, with regard to sequence length of the coding and noncoding regions. With more organism genome-wide data becoming available to study and new methods and technologies to explore, we can look forward to a surge in genome-wide comparative research.

Declaration of Authorship

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes text and figures from six (6) original papers in peer reviewed journals and / or conference proceedings and two (2) unpublished works. The inclusion of co-authors in the published papers reflects the fact that this work came from active collaborations. I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

The core theme of the thesis is gene architecture and length. The ideas, developed and authorship of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the University of Wollongong under the supervision of Dr Ren Zhang and A/Professor Yan-Xia Lin.

RCaldwell

Ms Rachel Amber Caldwell
2015

Papers Already Published Out of This Thesis

Caldwell, R., Dai, Y., Srivastava, S., Lin, Y., and Zhang, R. (2008) Improving neural network promoter prediction by exploiting the lengths of coding and noncoding sequences, Advances of Computational Intelligence in Industrial Systems (Studies in Computational Intelligence) edited by Ying Liu, Aixin Sun, Han Tong Loh, Wen Feng and Ee-Peng Lim, Springer, 213-230. doi:10.1007/978-3-540-78297-1_10.

Caldwell R., Lin, Y., and Zhang, R. (2008) Correlations of Length Distributions between noncoding and coding sequences of the *Arabidopsis thaliana*, Chapter: 2008 IEEE International Conference On Bioinformatics and BioMedicine BIBM 2008 (Philadelphia, Pennsylvania, USA) edited by Xue-wen Chen, Xiaohua Hu, and Sun Kim, IEEE Computer Society, 72-77.

Caldwell, R., Kongcharoen, J., Lin, Y., and Zhang, R. The Length Distributions of Noncoding and Coding Sequences in Relation to Gene Expression: A Study on *Arabidopsis thaliana*, Proceedings of IEEE International Conference on Bioinformatics and Computational Biology, 2010, Las Vegas, USA.

Caldwell, R., Lin, Y., and Zhang, R. (2010) Assessment of length distributions between noncoding and coding sequences amongst two model organisms, *International Journal of Data Mining and Bioinformatics*, 4 (5), 535-552. doi:10.1504/IJDMB.2010.035899.

Kongcharoen, J., Lin, Y., Caldwell, R., Yang, Y., and Zhang, R. (2011) Case Study on the Pattern Change in *Arabidopsis thaliana* Intron Sequence, *GSTF International Journal on Bioinformatics & Biotechnology*, 1(1), 18-23. doi:10.5176/ 2251-3159_1.1.3.

Caldwell, R., Lin, Y., and Zhang, R. (2015) Comparisons between Arabidopsis thaliana and Drosophila melanogaster in relation to Coding and Noncoding Sequence Length and Gene Expression, *International Journal of Genomics*, vol. 2015, Article ID 269127, 13 pages, 2015. doi:10.1155/2015/269127.

Statement of Contribution

Article 1 - *Improving neural network promoter prediction by exploiting the lengths of coding and noncoding sequences (2008):* Rachel Caldwell composed the manuscript for publication. The data analysis was conducted by the other authors and discussed with Rachel, who composed the manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang gave feedback on drafts of the manuscript prior to publication.

Article 2 - *Correlations of Length Distributions between noncoding and coding sequences of the Arabidopsis thaliana (2008):* Rachel Caldwell collected and compiled all the data and ran the analysis, and composed the research as a manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang gave feedback on drafts of the manuscript prior to publication.

Article 3 - *The Length Distributions of Noncoding and Coding Sequences in Relation to Gene Expression: A Study on Arabidopsis thaliana (2010):* Rachel Caldwell collected and compiled all the data and designed the research, ran the analysis, and composed the research as a manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang gave feedback on drafts of the manuscript prior to publication. Jinda Kongcharoen designed and conducted the Quantile regression analysis and gave feedback on drafts of the manuscript.

Article 4 - *Assessment of length distributions between noncoding and coding sequences amongst two model organisms (2010):* Rachel Caldwell collected and compiled all the data and ran the analysis, designed the research and composed the research as a manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang gave feedback on drafts of the manuscript prior to publication.

Article 5 - *Case Study on the Pattern Change in Arabidopsis thaliana Intron Sequence (2011):* Rachel Caldwell collected and compiled the data, Jinda Kongcharoen composed the research as a manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang and Rachel Caldwell gave feedback on drafts of the manuscript prior to publication.

Article 6 - *Comparisons between Arabidopsis thaliana and Drosophila melanogaster in relation to Coding and Noncoding Sequence Length and Gene Expression (2015):* Rachel Caldwell collected and compiled all the data and ran the analysis, designed the research and composed the research as a manuscript for publication. Yan-Xia Lin helped with the statistical model design and provided feedback on drafts of the manuscript prior to publication. Ren Zhang gave feedback on drafts of the manuscript prior to publication.

The co-authors below indicate their agreement to this statement:

Yan-Xia Lin
L. Yanxia

Ren Zhang

Acknowledgements

Firstly I would like to acknowledge my husband, who has been extremely patient and supportive, throughout the whole process.

I have gained invaluable experience and knowledge in both faculties, Biological Sciences and Informatics and wish to extend my thanks to my supervisors, Dr Ren Zhang for his support and discussions, and A/Prof Yan-Xia Lin, for her patience with me in advanced statistics, which was very much appreciated. I would like to thank the University of Wollongong for giving me the opportunity to advance my study. I will cherish the knowledge, memories and experiences that I have gained over the course of my post graduate studies.

Finally I would like to thank my family and work colleagues for all their support and encouragement to pursue my dreams.

List of Abbreviations

bp	Base pairs
CBP	Cap-binding Protein
cDNA	Complementary Deoxyribonucleic Acid
CCA	Canonical Correlation Analysis
CDS	Coding Sequence
COG	Clusters of Orthologous Groups
d1	Coding region length
d2	Distance between transcription start site and translation start site
d3	Distance between Stop Codon and Terminator Site
D1	Coding region length with introns
D2	Distance between transcription start site and translation start site with introns
D3	Distance between Stop Codon and Terminator Site with Introns
DNA	Deoxyribonucleic Acid
Fmet	Formylmethionine
GTP	Guanosine-5’ Triphosphate
IF	Initiation Factor
Mb	Megabytes
mRNA	messenger RNA
RefSeq	Reference Sequence Data collection
rRNA	Ribosomal RNA
RNA	Ribonucleic Acid
tRNA	Transfer Ribonucleic Acid
snRNP	Small nuclear ribonucleio proteins
TLS	Translation Start Site
TSS	Transcription Start Site
TTS	Terminator Site
TF	Transcription Factor
UTR	Untranslated region

List of Figures

Figure 1.1 Diagrammatic illustration of a structural gene, and areas of interest in this study. 3
Figure 1.2 General sequence of steps in the formation of eukaryotic mRNA 4
Figure 1.3 Translation initiation in prokaryotes 10
Figure 1.4 DNA Sequences on the 16S rRNA 11
Figure 1.5 Translation initiation in eukaryotes 12
Figure 1.6 Diagrammatic illustration of a structural gene 19
Figure 2.1 Phylogeny of Eukaryotes 23
Figure 2.2 Image of Arabidopsis thaliana 24
Figure 2.3 The genome structure of the *Arabidopsis thaliana* separated into chromosomes 25
Figure 2.4 TAIR FTP site for downloading *Arabidopsis* data 26
Figure 2.5 *Arabidopsis thaliana* Microsoft Access Master Database. 27
Figure 2.6 Image of *Drosophila melanogaster* 28
Figure 2.7 The genome structure of the *Drosophila melanogaster* separated into chromosomes 30
Figure 2.8 Batch download from the FlyBase website for data collection 32
Figure 2.9 MS Access master table containing all length data from publicly available sequencing data .. 33
Figure 3.1 Scatter plots of D_1 verses D_2 of *E. coli* and *H. sapiens* 41
Figure 3.2 The histogram and smoothed density of distance TSS-TLS for *A. thaliana* and *H. sapiens* 46
Figure 3.3 Schematic Representation of Promoter Prediction 47
Figure 3.4 Comparison of probability of prediction of promoter sequences 49
Figure 3.5 The comparison of three methods with D_1 in Group 1 51
Figure 4.1 NCBI RefSeq website Protein table screen. 56
Figure 4.2 Excel file containing length data obtained from NCBI RefSeq tables 57
Figure 4.3 - Mean number of proteins per Mb over all chromosomes 58
Figure 4.4 Density of proteins per Mb across 15 different genomes of eukaryotes 61
Figure 4.5 Median length (bp) across chromosomes of 15 species 64
Figure 4.6 Total percentage of protein coding sequences (CDS) 71
Figure 5.1 Length distributions (bp) of noncoding and coding sequences. 80
Figure 5.2 Median length values (bp) of each gene region for *Drosophila melanogaster* 82
Figure 5.3 Median length values (bp) of each gene region for *Arabidopsis thaliana* 82
Figure 5.4 Nonlinear functional relationship between d_1 and d_2^* 84
Figure 5.5 R-squared values based on nonlinear functional relationship for *Arabidopsis thaliana*. 85
Figure 5.6 R-squared values based on nonlinear functional relationship for *Drosophila melanogaster* 86
Figure 5.7COG database FTP file format 89
Figure 5.8 Protein category classifications for *Arabidopsis thaliana* 91
Figure 5.9 Protein category classifications for *Drosophila melanogaster* 92
Figure 6.1 Scatter plot of the gene expression average intensity and 3'UTR length 103
Figure 6.2 Total intron length of *Arabidopsis thaliana* 104
Figure 6.3 *Arabidopsis thaliana* median length distributions for 5' UTR and gene expression 106

Figure 6.4 *Arabidopsis thaliana* median length distributions for CDS and gene expression 107
Figure 6.5 *Arabidopsis thaliana* mean length distributions for 3' UTR and gene expression 109
Figure 6.6 *Arabidopsis thaliana* average intensity (gene expression) within protein categories 110
Figure 7.1 Representation of Canonical Variates – linear combinations of variables............................. 115
Figure 7.2 Mean length (base pairs) of coding & noncoding sequences for *Drosophila melanogaster* .. 117
Figure 8.1 The quantile curves of the average gene expression intensity for *Arabidopsis thaliana*........ 126
Figure 8.2 18,445 genes in *Arabidopsis thaliana* for the 5' UTR length, excluding introns. 132
Figure 8.3 Relationship of gene expression in *Arabidopsis thaliana* 135
Figure 8.4 Relationship of gene expression in *Drosophila melanogaster*... 136
Figure 8.5 Quantile regression plot for *Arabidopsis thaliana* of 5' UTR.. 138
Figure 8.6 Quantile regression plot for *Arabidopsis thaliana* of CDS 139
Figure 8.7 Quantile regression plot for *Arabidopsis thaliana* of 3' UTR.. 140
Figure 8.8 Quantile regression plot for *Drosophila melanogaster* of 5' UTR.. 141
Figure 8.9 Quantile regression plot for *Drosophila melanogaster* of CDS... 142
Figure 8.10 Quantile regression plot for *Drosophila melanogaster* of 3' UTR 143
Figure 8.11 Quantile regression plot for *Arabidopsis thaliana* of 5' UTR.. 144
Figure 8.12 Quantile regression plot for *Arabidopsis thaliana* of CDS 145
Figure 8.13 Quantile regression plot for *Arabidopsis thaliana* of 3' UTR.. 146
Figure 8.14 Quantile regression plot for Drosophila melanogaster of 5' UTR 147
Figure 8.15 Quantile regression plot for Drosophila melanogaster of CDS.. 148
Figure 8.16 Quantile regression plot for Drosophila melanogaster of 3' UTR 149
Figure 9.1 Workflow summary of the main research points conducted in this thesis 155

List of Tables

Table 1.1 Ribosome-Binding site sequence of prokaryotic mRNAs 7

Table 1.2 Comparison of the translation initiation process in prokaryotes and eukaryotes 9

Table 1.3 The URLs of databases and other tools used by molecular biologists 14

Table 2.1 NCBI RefSeq table for each chromosome. The data contains a list of protein gene information 31

Table 3.1 Statistics of the distances (bp) of D_1 and D_2..... 40

Table 3.2 Statistics of TSS-TLS distances (D_2) given D_1 in different ranges..... 42

Table 4.1 Species of eukaryotes sequences downloaded in 2007 from the RefSeq collection NCBI website for study 55

Table 4.2 Summary of ANOVA analysis for each eukaryote species..... 65

Table 5.1 Summary of analysis based on the nonlinear model for Arabidopsis thaliana on Chromosome 4. Data was subset into 12 categories based on Log(d_2) values. 85

Table 5.2 Summary of analysis based on the nonlinear model for Drosophila melanogaster on Chromosome 2b. Data was subset into 4 categories based on Log(d_2) values. 86

Table 5.3 COG Functional Protein Classification [http://www.ncbi.nlm.nih.gov.COG/]. The classification was divided into 4 main categories 1) Information storage and processing; 2) Cellular processes and signalling; 3) Metabolism; and 4) Poorly Characterised. 90

Table 5.4 ANOVA analysis on Arabidopsis thaliana between the length of the coding (with and without introns - D1 / d1) and noncoding (with and without introns - D2, D3 / d2, d3) gene regions in relation to protein function. Protein function was divided into four (4) main categories based on the COG classification. 92

Table 5.5 ANOVA analysis on Drosophila melanogaster between the length of the coding (D1 / d1) and noncoding (D2, D3 / d2, d3) gene regions in relation to protein function. Protein function was divided into four (4) main categories based on the COG classification. 93

Table 6.1 Coding sequence of Arabidopsis thaliana without introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 99

Table 6.2 Coding sequence of Arabidopsis thaliana with introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 100

Table 6.3 5’ Un-translated region (UTR) of Arabidopsis thaliana without introns. Average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 100

Table 6.4 5’ Un-translated region (UTR) of Arabidopsis thaliana with introns. Average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 101

Table 6.5 3’ Un-translated region (UTR) of Arabidopsis thaliana without introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 102

Table 6.6 3’ Un-translated region (UTR) of Arabidopsis thaliana with introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions...... 102

Table 7.1 Pearson’s Correlation Analysis with Drosophila melanogaster relating to the length distributions of coding and noncoding sequences to gene expression under all environmental conditions. Sample size of data N=4,841. . 116

Table 7.2 Canonical Correlation Analysis with Drosophila melanogaster relating to the length distributions of coding and noncoding sequences to gene expression under various environmental conditions. * Represents significance at α 0.05..... 118

Table 7.3 Canonical loadings for the First Canonical Function 119

Table 8.1 Quantile regression analysis results on Arabidopsis thaliana between the log of 5’ UTR sequence length and the log of gene expression (GSE31488 gene expression experiment data) 138

Table 8.2 Quantile regression analysis results on Arabidopsis thaliana between the log of CDS sequence length and the log of gene expression (GSE31488 gene expression experiment data) 139

Table 8.3 Quantile regression analysis results on Arabidopsis thaliana between the log of 3’ UTR sequence length and the log of gene expression (GSE31488 gene expression experiment data) 140

Table 8.4 Quantile regression analysis results on Drosophila melanogaster between the log of 5’ UTR sequence length and the log of gene expression (GSE42255 gene expression experiment data) 141

Table 8.5 Quantile regression analysis results on Drosophila melanogaster between the log of CDS sequence length and the log of gene expression (GSE42255 gene expression experiment data) 142

Table 8.6 Quantile regression analysis results on Drosophila melanogaster between the log of 3’ UTR sequence length and the log of gene expression (GSE42255 gene expression experiment data) 143

Table 8.7 Quantile regression analysis results on Arabidopsis thaliana between the log of 5’ UTR sequence length and the log of gene expression (GDS3933 gene expression experiment data) 144

Table 8.8 Quantile regression analysis results on Arabidopsis thaliana between the log of CDS sequence length and the log of gene expression (GDS3933 gene expression experiment data) 145

Table 8.9 Quantile regression analysis results on Arabidopsis thaliana between the log of 3’ UTR sequence length and the log of gene expression (GDS3933 gene expression experiment data) 146

Table 8.10 Quantile regression analysis results on Drosophila melanogaster between the log of 5’ UTR sequence length and the log of gene expression (GSE36507 gene expression experiment data)... 147

Table 8.11 Quantile regression analysis results on Drosophila melanogaster between the log of CDS sequence length and the log of gene expression (GSE36507 gene expression experiment data) 148

Table 8.12 Quantile regression analysis results on Drosophila melanogaster between the log of 3’ UTR sequence length and the log of gene expression 149

Chapter 1 - Introduction

1 Introduction

Mendel, known as the father of genetics, built the fundamental laws of inheritance through experiments and statistical analysis on the garden pea. Genetics then established itself as a core discipline at the beginning of the 20th century, opening a whole new world of science (Fairbanks and Rytting, 2001). Through the development of sequencing technology over the past forty years, a tremendous amount of genomic sequencing data has already been collected, with a flood of such data increasing even more rapidly in the coming years. As a result, a better understanding and insight into the mystery of gene architecture and its associated mechanisms will be possible. Genes are functional units of genetic material. Among many attributes a gene possess, its length is fundamental in the gene's architecture, which is related to function. To understand the relationship between gene lengths, its associated architecture and further gene function will help uncover the sophisticated gene regulation enigma.

… *"for most of the genes that we identify, we have no idea of their biological functions. They are like words in a foreign language, waiting to be deciphered."* Iddo Friedberg, computational biologist at Miami University (33rd_Square)

1.1 Background

1.1.1 Gene Structure

Genes contained in DNA, compose only a small portion of the genome. For example in the human genome, only a small percentage of the total DNA in the genome is made up of protein coding genes. A gene, which Mendel called factors, is the basic functional unit of genetic materials of all living organisms. Any fragment of DNA which can be transcribed into RNA within a cell is called a gene. A structural gene (referred to as a gene for this thesis) contains a sequence that code for proteins. Each protein is produced from the genetic code within the DNA. This type of gene consists of a protein coding region between the translation start (TLS) and stop (STOP), and its 5' upstream, and 3' downstream noncoding regions. For RNA synthesis, at least one transcription start site (TSS) and terminator site (TTS) are located in the upstream and downstream regions, respectively (Figure 1.1). The whole region between TSS and TTS is transcribed into RNA in prokaryotes, however in eukaryotes, there may be certain parts (introns) which do not transcribe.

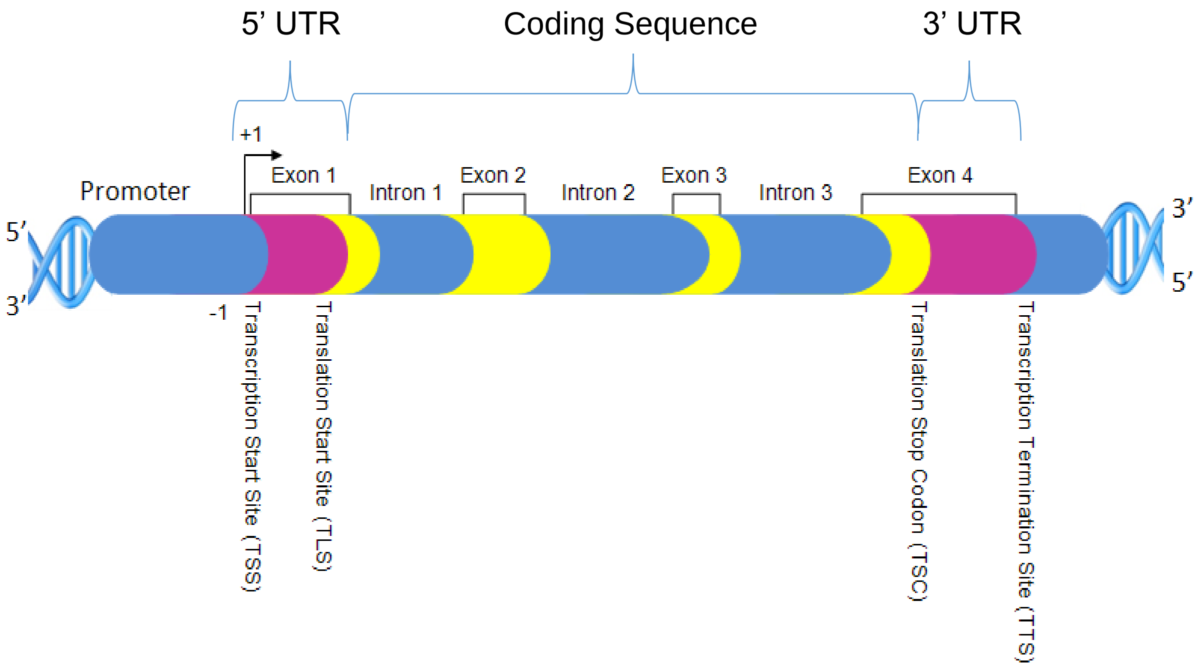

5' UTR
Coding Sequence
3' UTR
+1
Promoter
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Intron 3
Exon 4
5'
3'
3'
5'
-1
Transcription Start Site (TSS)
Translation Start Site (TLS)
Translation Stop Codon (TSC)
Transcription Termination Site (TTS)

Figure 1.1 Diagrammatic illustration of a structural gene, and areas of interest in this study. The gene consists of two untranslated regions 5' and 3' which flanks the coding sequence. The coding sequence is transcribed and translated into proteins from the DNA.

1.1.2 Gene Expression

Gene expression is a tightly regulated and complex process, consisting of two major stages – transcription and translation. It can be described as the process of genetic transcription from the base sequence on DNA, and genetic translation for the production of proteins. Every living organism depends on genes and gene expression to produce proteins that play many critical roles. Proteins not only build structural components but can also determine how food is metabolised or how the organism can fight infections (Villarreal, et al., 2014). For example *Arabidopsis* Receptor-like proteins (RLPs) (Wang, et al., 2008) have been identified as playing significant roles in meristem and organ development (Jeong, et al., 1999).

1.1.3 Transcription

The first step in gene expression is transcription, the process of copying DNA into messenger RNA (mRNA). The mechanisms involved in transcription include the promoter sequences: transcription start site, the TATA box, and sequences bound by transcriptional regulators, the enzyme RNA Polymerase (Pol), and regulatory factors (Hahn, 2004). Transcription factors assemble at the promoter region of a gene, obtaining the RNA polymerase enzyme to form the transcription initiation complex. The transcription mechanism is much more complex in eukaryotes (Lee and Young, 2000), using three nuclear enzymes (Pol I-III) compared to bacteria and archaea which only have one, however the principle of transcription and its regulation is still preserved between these sequences.

1.1.4 The Function of Exons and Introns

Spliceosomal introns are a ubiquitous feature of eukaryote genomes, however are absent from the bacterial and archaeal genomes. In eukaryotic organisms, the coding portion of the gene is called an exon and is usually flanked by sequences called introns. When the gene is transcribed into messenger RNA (mRNA) it still includes both the exon and un-translated introns. This sequence is called the pre-mRNA, and the removal of introns from the pre-mRNA is completed before the mature mRNA is translated into proteins. Figure 1.2 offers a generalized view on the formation of the pre-mRNA and the removal of introns before the polypeptide is translated and produced. Splicing of introns occurs in complexes called spliceosomes (Nilsen, 2003) which occurs in the nucleus of the cell. The pre-mRNA 5' splice junction binds to small nuclear ribonucleoproteins particles, known as snRNPs or snurps.

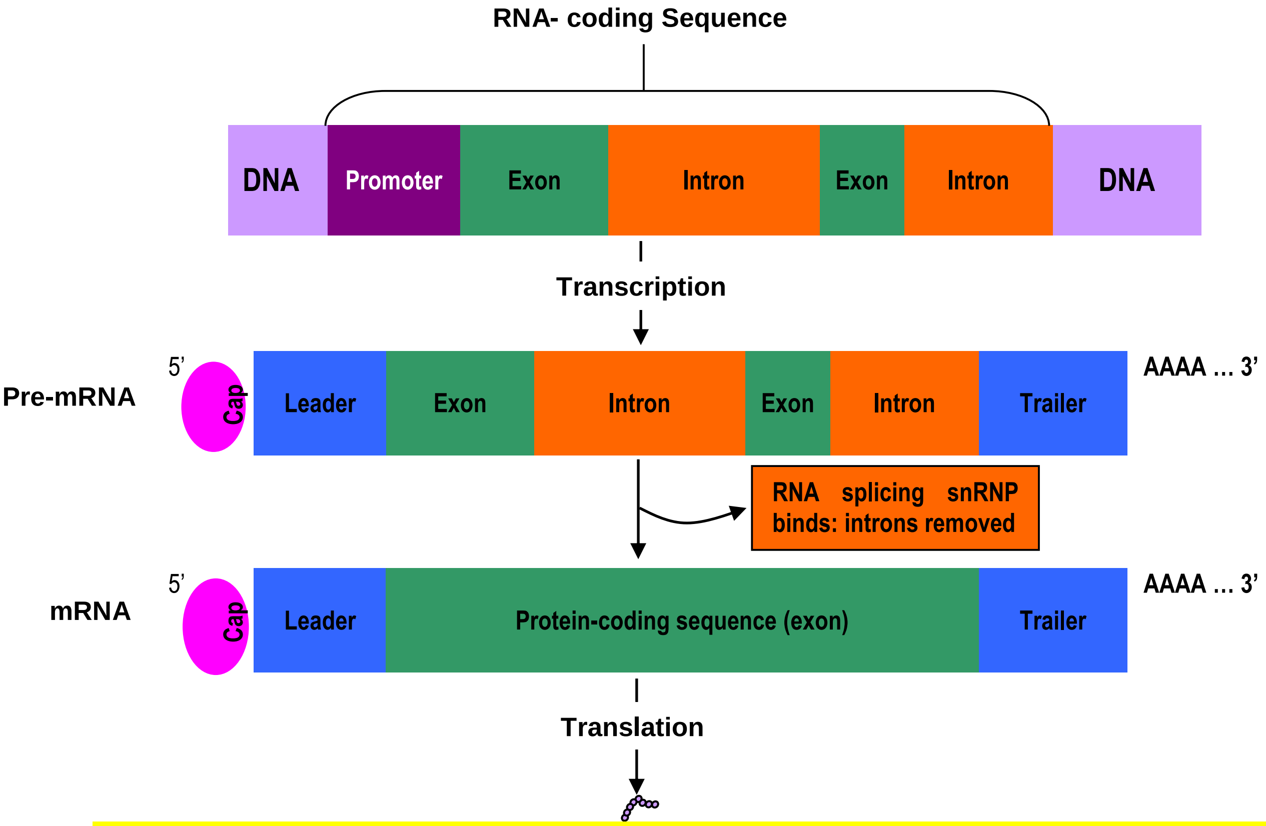


Figure 1.2 General sequence of steps in the formation of eukaryotic mRNA. The coding sequence is transcribed into a pre-mRNA, where the introns are spliced from the sequence to form the mature mRNA. This sequence contains the exons which are translated into proteins.

(LÓPEZ-LASTRA, et al., 2005)

This process of splicing incurs a further cost to the organism in energy and time during replication and transcription (Duret, 2001). Therefore, why do eukaryote organisms have these sequences if they are spliced out of the mRNA? The debate is still continuing in this area, however identification of various models may divulge whether there is a selective advantage on having these noncoding sequences present. Duret (2001) outlines several theories that may clarify why introns have selective advantage, albeit the high energy cost it has on eukaryote organisms. Firstly, alternative splicing produces many proteins from one gene. It is estimated that 60% of all human genes undergo alternative splicing (Bracco and Kearsey, 2003), which could be beneficial in a high source of functional diversity. Secondly, introns may contain regulatory elements, alternative promoters or antisense promoters that aid in the production of proteins. Thirdly, introns may contain genes that produce miRNAs and snoRNAs. Other investigations have inferred that exons, introns and intergenic regions[1] are not random and contribute to the design and architecture of the genome, with length of introns on each chromosome showing a strong relationship to chromosome size (Sakharkar, et al., 2005).

1.1.5 Promoters

Promoter regions are important sequences that starts the process of transcription. A typical promoter sequence is thought to comprise some sequence motifs surrounding transcription start sites (TSSs) (Kanhere and Bansal, 2005). The properties of these regions differs from the genomic regions with structural features being one of the distinguishing features of these regions (Zeng, et al., 2009). Differences occur between the promoter sequences of prokaryotes and eukaryotes with prokaryotic sequences having a relatively short upstream region compared to eukaryotic sequences where they seem to have larger upstream regions (Kanhere and Bansal, 2005). Since 1997 design and implementation of promoter recognition algorithms and software has progressed rapidly (Zeng, et al., 2009). Promoter prediction is an important tool in understanding genomes and gene regulation (Gan, et al., 2012).

1.1.6 Translation

The second step in gene expression, and the production of proteins is translation. The mRNA interacts with a specialised complex known as a ribosome that reads the

[1] An Intergenic region (IGR) is a stretch of DNA sequences located between genes. Intergenic regions are a subset of Noncoding DNA

sequence of the mRNA bases. Each sequence has three bases called a codon, which code for one particular amino acid.

Translation initiation is an important regulatory process in gene expression of all living organisms and was poorly understood until the mid-1970s where studies by Shine and Dalgarno identified consensus sequences relating to levels of gene expression (Fuglsang, 2005). The process in which proteins are synthesized has been explored extensively in bacteria, in particular, *E. coli*. This has enhanced the understanding of the translation initiation process for the production of proteins in both prokaryotes and eukaryotes. The initiation phase governs the regulation of protein synthesis which has made it an important step (Kozak, 2005; Ma, et al., 2002).

The process of translation initiation within prokaryotes involves three monomeric protein initiation factors, IF1, IF2 and IF3 (Londei, 2005), and GTP that bind to a 30S ribosomal subunit (Figure 1.3) (Kozak, 1983). This ribosomal subunit is used as part of the recognition process that identifies the region on the mRNA to start the initiation process. The widely held theory has been that there is a sequence upstream from the initiation code (AUG) – at the 5’ end. This sequence is known as the Shine-Dalgarno sequence, after the two researchers that first identified it (Shine and Dalgarno, 1974). This sequence pairs with the 3’ end of the 16S rRNA (Figure 1.4). The code, which has been found in *E. coli*, consists of the motif of AGGAGG or similar (Osada, et al., 1999; Russell, 2002). Other consensus ribosome binding site sequences found in prokaryotes include AGGAGGU, UAAGGA, UAAGGAGGU, and extensive experiments on *E. coli* have also established the importance of the Shine-Dalgarno base pairing (Ma, et al., 2002). Most binding sites contain a high portion of purine-rich sequences that are located primarily upstream from the initiation codon (Londei, 2005). The ribosome-binding site location and sequences for prokaryotes may vary, depending on the species and protein, the gene is designed to produce. Table 1.1 outlines several prokaryotic ribosome-binding sites and details the location from the initiation codon and the composition of the sequence.

Table 1.1 Ribosome-Binding site sequence of prokaryotic mRNAs

The binding site sequences represent regions of complimentary base pairing between the mRNA and the 3' end of 16S rRNA.

Phage R17 A protein	UCC	UAG	GAG	GUU	UGA	CCU	AUG	CGA	GCU	UUU
Phage Qβ replicase	UAA	CUA	AGG	AUG	AAA	UGC	AUG	UCU	AAG	ACA
Phase λ Cro	AUG	UAC	UAA	GGA	GGU	UGU	AUG	GAA	CAA	CGC
Phage Φ X174 A	AAU	CUU	GGA	GGC	UUU	UUU	AUG	GUU	CGU	UCU
E. coli trpB	AUA	UUA	AGG	AAA	GGA	ACA	AUG	ACA	ACA	UUA
E. coli lacZ	UUC	ACA	CAG	GAA	ACA	GCU	AUG	ACC	AUG	AUU
E. coli RNA Polymerase β	AGC	GAG	CUG	AGG	ACC	CCU	AUG	GUU	UAC	UCC

🟪 Binding site sequences

🟨 Initiation codon

Another important element in the initiation process for prokaryotes is formylmethionine (fMet). This molecule is brought to the ribosome via a transfer ribonucleic acid (tRNA), where it attaches to the start codon, and contains the anticodon sequence, 3' – UAC – 5'. At this point in the initiation process the initiation complex contains the mRNA, 30S subunit, fMET-tRNA and the two remaining initiation factors as well as the Guanosine-5' Triphosphate (GTP) molecule. AUG start sites in prokaryotic mRNAs appear to be more common, which may be explained by the stability the codon creates when binding to the fMet-tRNA. However, there are other initiator codons that are used within > 10% of bacterial genes, and they include GUG and UUG (Kozak, 2005). Release of the remaining two initiation factors, IF1 and IF2 is obtained by the binding of the 50S ribosomal subunit. This final complex, before elongation of the polypeptide chain transpires is known as the 70S initiation complex, and consists of two binding sites which include a P site (peptidyl) that contains the mRNA and fMet-tRNA, and the A site which is vacant (Kozak, 2005).

Eukaryotic translation initiation entails a more complicated process. What contributes to the complexity of the initiation of translation in eukaryotes can be stipulated by several factors. Eukaryotes mRNA shape is unusual in that it adopts a circular structure due to the interaction between the proteins of the poly(A) tail and the 3′ end containing a number of factors which are used to recognise the cap at the 5′ end (Londei, 2005). The poly(A) plays an important role in the initiation of translation, bringing together the 3′

end of the mRNA to the 5′ end, stimulating initiation. Another factor that contributes to the complexity is that the process requires over 10 factors that assist in the initiation process (Londei, 2005; Preiss and Hentze, 2003). The factors are also an important characteristic, because unlike the prokaryotes, there is no defined interaction with the ribosome, therefore the mRNA and many different factors are involved in this interaction. In addition, the factors also aid in the preliminary unwinding of the secondary structures in the mRNA (Londei, 2005).

The current theory on the initiation of translation within eukaryotes involves four subsequent steps. The first step involves the eIF-4F initiator factors together with the cap-binding protein (CBP) binding to the 5′ end cap of the mRNA (Figure 1.5). Secondly, a 43S initiation complex is created from a 40S ribosomal subunit, a Met-tRNA initiator and several eIF protein initiation factors, together with GTP. The initiation complex then binds to the 5′ mRNA where in the third step "scans" the 5′ untranslated region (UTR) of the mRNA until recognition occurs with the initiator AUG start codon. The complex distinguishes this codon as the initiator codon as it sits in a short sequence known as the Kozak sequence (Kozak, 1987), and is virtually the first AUG codon from the 5′ end of the mRNA. Finally, the 43S complex binds to the AUG codon and a 60S subunit joins it, creating a large 80S ribosome initiation complex. In this step, the eIFs are released and the Met-tRNA initiator locates itself with the P site, which is a similar method found in the prokaryotes (Preiss and Hentze, 2003) (Figure 1.5).

1.1.7 Comparison between Prokaryotes and Eukaryotes

Eukaryotes have more elaborate translational regulation mechanisms in comparison to the prokaryotes (Table 1.2). In the example from bacteria, the presence of the Shine-Dalgarno sequence allows for a more rapid decoding process due to the mRNAs being largely polycistronic[2]. This recognition mechanism of the ribosome and the mRNA is sufficient for polycistronic mRNA. However, eukaryotes are more sophisticated and require a higher level of translational regulation involving a considerable number of initiation factors, which may be redundant for prokaryotes (Londei, 2005).

[2] Single mRNA which can code for several genes

Table 1.2 Comparison of the translation initiation process in prokaryotes and eukaryotes

(Kozak, 2005; Londei, 2005; Pestova, et al., 2001; Preiss and Hentze, 2003)

Organism	Initiation Factors	Ribosomal Subunit	Initiation complex	Selection of Start sites	Final initiation Complex
Prokaryote	IF-1, IF-2, IF-3, GTP and magnesium ions.	30S ribosomal subunit containing all initiation factors. Binds to mRNA around AUG initiation codon region.	30S initiation complex, which consists of mRNA, 30S subunit, fMET-tRNA and the two remaining initiation factors as well as the GTP molecule	Start sites consists predominately of AUG but can consist of GUG and UUG (>10% of bacterial genes).	70S initiation complex incorporates the mRNA, fMet-tRNA, 50S and 30S ribosomal subunits, a P site which contains the fMet-tRNA and an A site which is vacant.
Eukaryote	Over 10 eIFs and GTP. ↓ eIF-1, eIF-1A, eIF-2, eIF-2B, eIF-3, eIF-4E, eIF-4G, eIF-4A, eIF-4b, eIF-5	40S ribosomal subunit containing initiation factors. Binds to the 5' cap of the mRNA.	43S initiation complex is created from a 40S ribosomal subunit, a Met-tRNA initiator and several eIF protein initiation factors, together with GTP.	Start sites consist of AUG and a "Scanning" mechanism is used to find the first AUG start codon from the 5' end of the mRNA.	80S initiation complex includes the mRNA, Met-tRNA, 40S and 60S ribosomal subunits and a P site which contains the Met-tRNA.

1.1.8 Protein Function

The shape and function of the protein is determined from this code, which enumerates the number of amino acids and order in which to place them. Proteins are long chains of polypeptides, as many as 20 different kinds of amino acids linked in a characteristic sequence. The proteins produced in an organism, have important applications for the living organism. A cell can accommodate thousands of different proteins, which all have essential functions within a cell (Buxbaum, 2007). The protein functions includes enzymes for making new molecules and catalysing all chemical processes in a cell; they can also give the cells their structural shape (de Lanerolle and Cole, 2002); hormones for signalling (Adams, et al., 2000; Rosenbaum, et al., 2009); antibodies for recognizing foreign molecules and combating disease (Westergard, et al., 2007); as well as transport molecules (Ehrnstorfer, et al., 2014; Terwilliger, 1998).

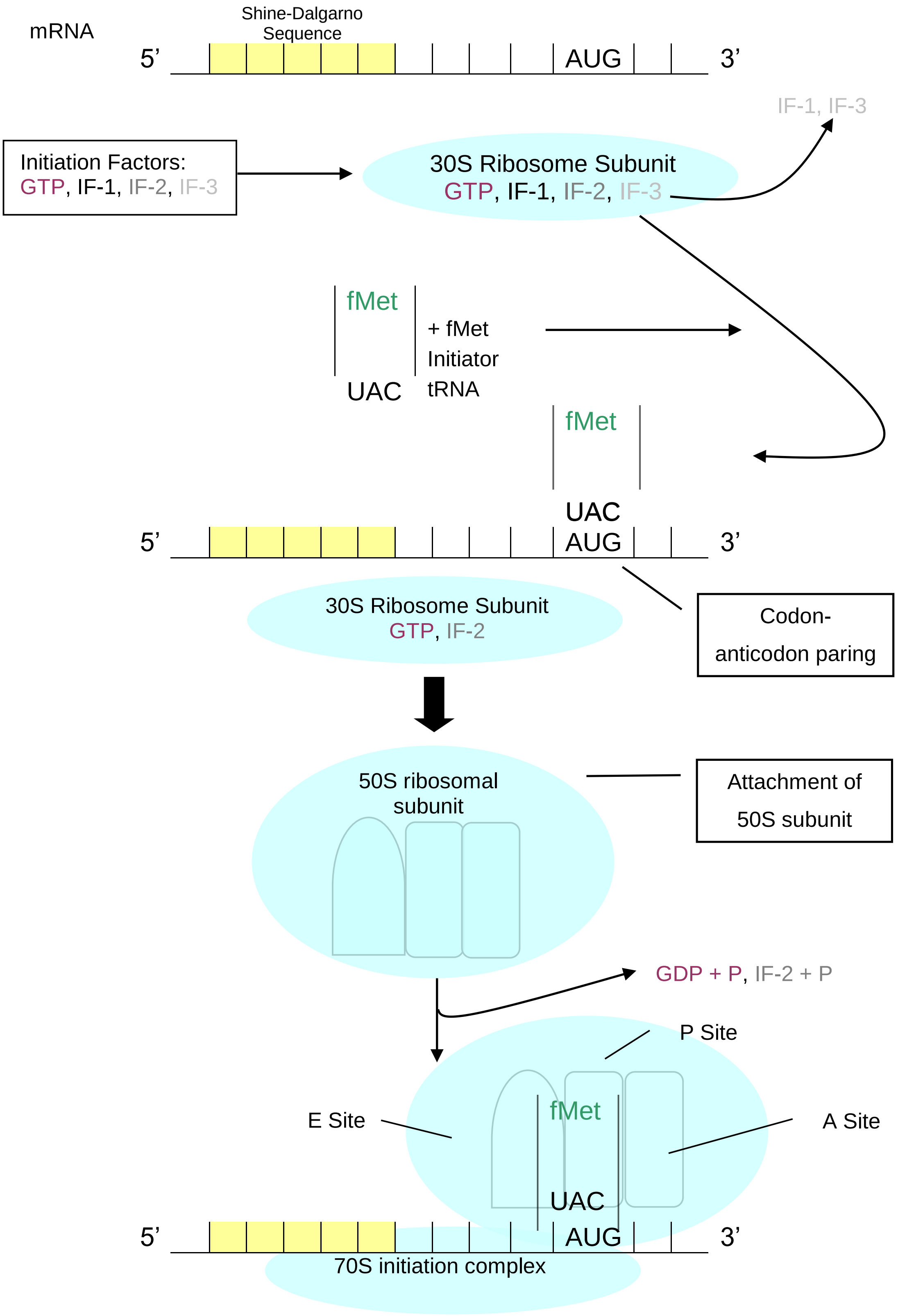


Figure 1.3 Translation initiation in prokaryotes

A 30S ribosomal subunit which is bound by initiation factors IF1, IF2, IF3, GTP and magnesium ions binds to a mRNA in the region of the AUG initiation codon. fMet-tRNA also binds to the mRNA at which point the IF1 is released and forms a more stable 30S initiation complex. The formation of the final 70S initiation complex is instigated by the binding of the 50S ribosomal subunit, where the remaining initiation factors are released and GTP is hydrolysed and released.

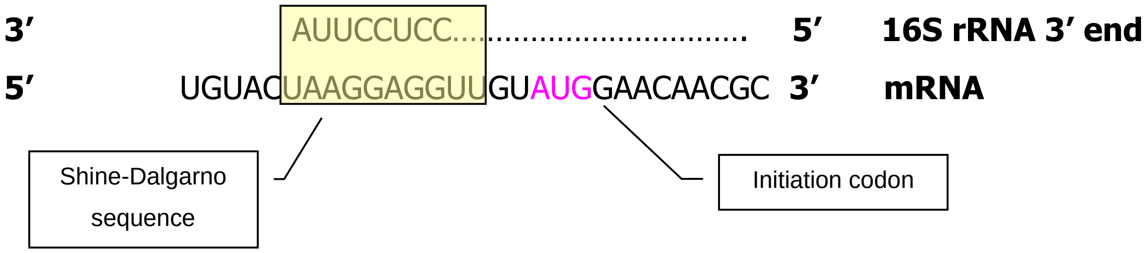

3'
AUUCCUCC.....................................
5'
16S rRNA 3' end
5'
UGUACUAAGGAGGUUGUAUGGAACAACGC
3'
mRNA
Shine-Dalgarno sequence
Initiation codon

Figure 1.4 DNA Sequences on the 16S rRNA

DNA Sequences on the 16S rRNA that are compatible with the DNA sequences upstream from the initiation codon (AUG). This sequence is known as the Shine-Dalgarno sequence in prokaryotes.

1.1.9 Model Organisms

Model organisms are widely used to understand a range of biological phenomena in order to apply generalised theories and principles to more complex organisms. The organisms are not only used for the convenience of maintaining and breeding in a laboratory environment, but there is also a large collection of data readily available that is publicly accessible (Twyman, 2002). Model organisms emerged in the early 1900s in three stages (Davis, 2004) revitalising the age of comparative genomics. The most widely used species include the mouse (*Mus musculus*), rat (*Rattus rattus*), zebrafish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), and thale cress (*Arabidopsis thaliana*). The data for model organisms has also been used extensively in many studies, and the integrity of the data has been proven already in peer reviewed publications and supported websites and databases.

There are three major types of model organisms:

- Genetic model organisms (used for genetic analysis);
- Experimental model organisms (experimental advantages); and
- Genomic model organisms (occupy a position in the evolutionary tree).

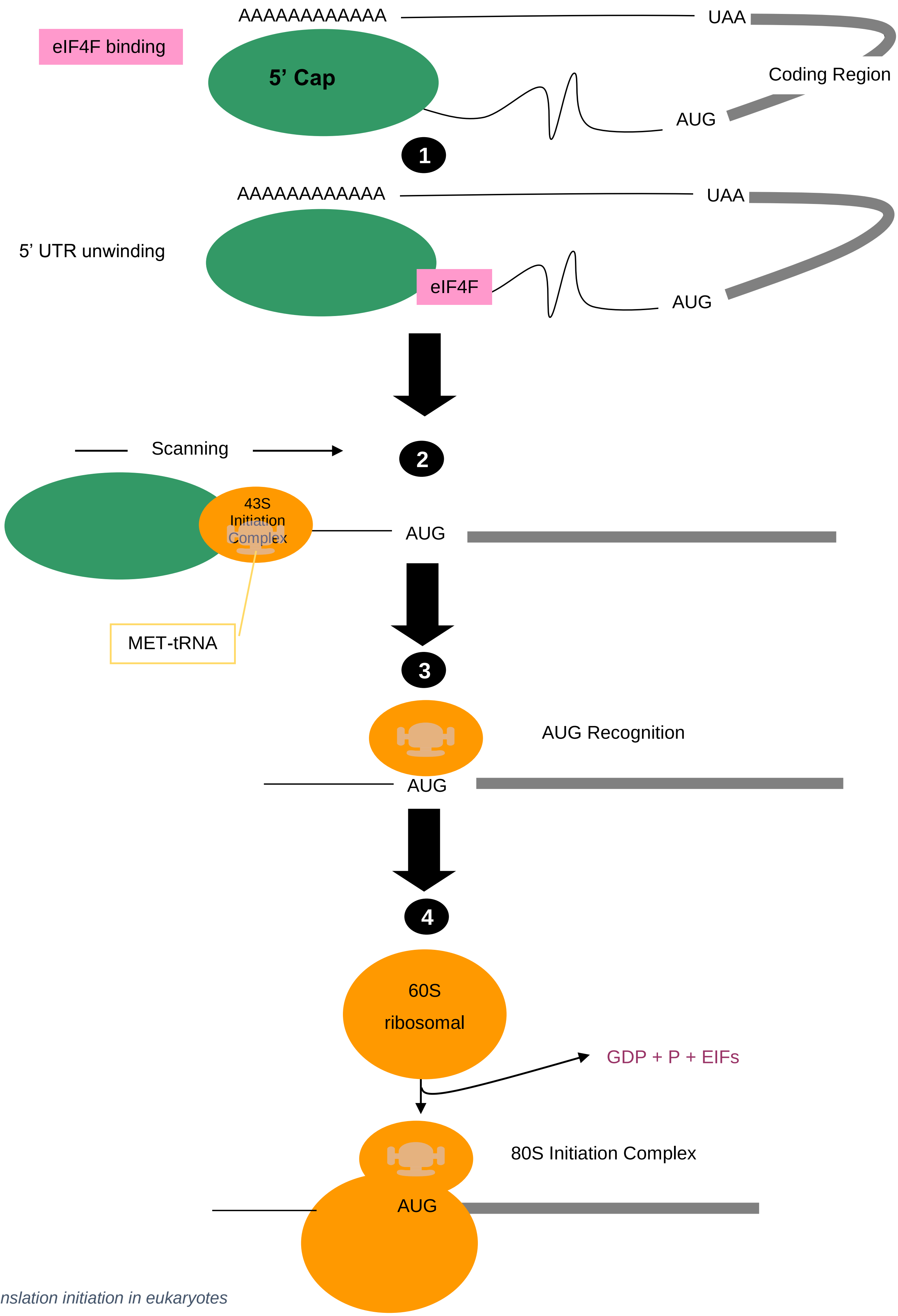


Figure 1.5 Translation initiation in eukaryotes

A 40S ribosomal subunit which is bound by several initiation factors and GTP bind to an mRNA in the 5’ cap region. Met-tRNA also binds to the mRNA which forms the 43S initiation complex. The complex scans the mRNA to find the initiator codon AUG. The formation of the final 80S initiation complex is instigated by the binding of the 60S ribosomal subunit, where GTP is hydrolyzed and released (Jackson, et al., 2010; Pestova, et al., 2001; Preiss and Hentze, 2003).

Model organisms have improved understanding in ovarian cancer metastasis (Naora and Montell, 2005), human disease studies (Chintapalli, et al., 2007) as well crop improvement (Bressan, et al., 2009). Model insects such as the *Drosophila melanogaster* have improved the understanding of behaviour and environmental interaction (Jasny, et al., 2008; Robinson, et al., 2008), as well as determining the basic rules of circadian clocks which has led to discoveries in sleep deprivation, obesity, diabetes, depression and other human health conditions (Panda, et al., 2002). These examples are just a small fraction of what is currently being investigated, and as more data becomes available for other organisms the list of model organisms will grow.

1.1.10 Sequence Databases and Tools

With the commercial introduction of the internet in the early 1990's, the scope and expanse of the "World Wide Web" could not have been foreseen with such a dramatic impact on culture, commerce and molecular biology research. After the introduction of the internet, thousands of web sites across the world have been created relevant to biology. Walter Gilbert (Gilbert, 1991) urged molecular biologists to cultivate their computer literacy skills to start a worldwide communication network. The Internet has benefited the science community with data published and available virtually instantaneously, and allows users to exchange views and ideas, and access a network of tools for biological research. For biologists, the use of the internet has allowed access to a wide range of up-to-date information without leaving their laboratory (Recipon and Makalowski, 1997).

Each year the number of web sites, tools and databases available for researchers increases considerably. Additional to these sites, researchers can also download from FTP sites, view journals on line, and join news groups. Table 1.3 outlines a few relevant molecular biology sites currently available from the Internet. These are only a few web sites out of hundreds that are available on the Internet.

Nucleotide sequence databases require unique identifiers for each item and are known as the Accession Number. This number never changes, and therefore can be quoted in scientific literature (Apweiler, Bairoch et al. 2004). These databases have improved connections to a wide range of data and allowed for greater comparative analyses.

Table 1.3 The URLs of databases and other tools used by molecular biologists

Database or Site	URL	Description
123 Genomes	http://www.123genomics.com/	A Genomics, Proteomics and Bioinformatics Knowledge Base.
*COG Database	http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups of proteins (COGs).
*EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/	Europe's primary nucleotide sequence resource (Stoesser, et al., 2003).
Expression Atlas	http://www.ebi.ac.uk/gxa/home	The Expression Atlas provides information on gene expression patterns under different biological conditions.
GEO DataSets	http://www.ncbi.nlm.nih.gov/gds	This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository.
KEGG	http://www.genome.jp/kegg/	KEGG **(**Kyoto Encyclopaedia of Genes and Genomes**)** is a bioinformatics resource for linking genomes to life and the environment.
*NCBI	http://www.ncbi.nlm.nih.gov/	A national resource for molecular biology information.
Pfam	http://pfam.sanger.ac.uk/	The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
UniProt	http://www.ebi.ac.uk/uniprot/	High quality and freely accessible resource of protein sequence and functional information.
PACdb	http://harlequin.jax.org/pacdb/	PACdb is a database of mRNA three prime (3') processing sites.
DBTSS	http://dbtss.hgc.jp/	DBTSS is a database of transcriptional start sites, based on our unique collection of precise, experimentally-determined 5'-end sequences of full-length cDNAs.
*Flybase	http://flybase.org/	A database of *Drosophila* genes and genomes.
*TAIR	http://www.arabidopsis.org/	The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*.
UTRdb	http://utrdb.ba.itb.cnr.it/	UTRdb is a curated database of 5' and 3' untranslated sequences of eukaryotic mRNAs, derived from several sources of primary data.
Virtual Library - Biosciences	http://vlib.org/Biosciences	The ultimate bioscience jump-station, with links to just about anything you want to know about biology.
Wormbase	http://www.wormbase.org/#01-23-6	WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of C. elegans and related nematodes (Harris, et al., 2010).

* Data accessed for the research in this thesis

1.2 From Sequence to Discovery – review of length distribution studies

The primary structure of DNA and proteins has been predominantly the focus of sequence analysis. However, other attributes, such as sequence length are also important. We set the stage for this thesis by presenting the current understanding and research in the domain of genome size, protein length and length distributions, with reference to gene expression. The journey of gene length research commences with Zhang (2000) who conducted an investigation on protein length for three domains of life. Protein length was found to be 40-60% greater in eukaryotes than in prokaryotes (Zhang, 2000). This finding was substantiated by Xu (2006), which found that the mean length of genic coding sequences is highly conserved in prokaryotes and eukaryotes but diverges between the two kingdoms (Wang, 2005; Xu, et al., 2006). They reported that the coding sequence length is on average 445 bp longer in eukaryotes than in prokaryotes (Xu, et al., 2006). These findings still hold true today.

Zhang's research also suggests that the differences in the length is not random, but has some biological significance (Zhang, 2000). This has led to research focusing on eukaryote protein size, conservation, complexity, and compactness. Proteins evolve under a variety of constraints and include links to specific function, GC content of DNA, and protein length. Wang (2005) discovered that among eukaryotes, comparison of protein sizes vary between the younger and older proteins. They found that the younger proteins are significantly longer than old proteins, by approximately 22% (Wang, 2005). There are several advantages of producing shorter proteins, which include regulation of innate immunity; protection against pathogens; cell communication and homeostasis as ligands and hormones; signal transduction; and metabolism (Frith, 2006). This research suggests that protein size is an important factor in the management of biological processes, particularly in eukaryotes, and protein size influences these processes. Function was also attributed to protein size and conservation. When associating protein size to conservation, it was found that poorly conserved proteins are, on average, shorter than the highly conserved proteins (Lipman, et al., 2002).

Protein length is also a contributing factor to the complexity of eukaryotes, in addition to regulation and structure. Eukaryotic genes are distinctively more complex than prokaryotes (He and Zhang, 2005; Huang, et al., 1999; Zhang, 2000) and protein length appears to be a significant mechanism in influencing complexity (Brocchieri and Karlin, 2005; He and Zhang, 2005; Tan, et al., 2005).

Furthermore, research has determined when investigating the length of introns, UTRs and the coding sequences that specific genes, notably housekeeping genes[3] are more compact than other tissue-specific genes (Eisenberg and Levanon, 2003). Vinogradov (2004) identified that more tissue-specific genes are longer than the housekeeping genes due to more functional domains (Vinogradov, 2004). However, the latest research investigating housekeeping genes found that the genes are less compact and older that the tissue-specific genes (Zhu, 2008). It was also found in *E. coli* studies that the variance of the length distribution for essential genes is found to be smaller than for non-essential genes, implying that these distributions are intentional (Ribeiro, et al., 2012).

The abundance of microarray and sequencing data has also extended understanding in the areas of classification, composition and evolution (Akashi, 2001; Lin and Chien, 2009; Raghava and Han, 2005). The next transition in understanding the effects of length was to incorporate gene expression level data into the investigations. Gene expression is a fundamental process to all living organisms and involves stringent levels of control at the transcriptional and translation initiation stages. A large amount of work has been conducted in this area. Focusing mainly on protein length, a noteworthy study was published in 1999. When Duret and Mouchiroud sought to examine expression levels in association with selection on codon usage, for three model organisms, *D. melanogaster*, *C. elegans*, and *A. thaliana* (Duret and Mouchiroud, 1999). This opened the debate on the correlations between length and gene expression. This research concluded that there was a strong negative correlation between coding usage[4] and protein length (Duret and Mouchiroud, 1999). The *R* values obtained from all three organisms were negative and averaging around -0.42 for moderate expression and -0.46 for high expression (Duret and Mouchiroud, 1999).

In 2006 Ren et al reported that in both monocot rice and dicot *Arabidopsis* plants, highly expressed genes are less compact than lower expressed genes (Ren, et al., 2006). The research found when considering the full length per gene, the sequence is larger in higher expressed genes than in lower expressed genes. These results were influenced by the

[3] Any of the genes that are constitutively expressed at a relatively constant level across many or all known conditions.

[4] Assumption is that all genomes have uniform codon usage meaning that synonymous codons are used with equal frequency.

higher number of introns, in spite of that, the average exon length was negatively correlated with the expression level (Ren, et al., 2006). However, in a study in 2009, the research found highly expressed genes are compact, particularly in the noncoding regions for rice and *Arabidopsis* plant species (Yang, 2009). This research indicates that the noncoding regions have importance in the regulation of gene expression, and that longer UTRs may contain regulatory motifs that have the potential to produce complex temporal and spatial translational programmes (Doran, 2008). It has been shown that 3' UTRs are significantly longer than 5' UTRs, with 3' UTR sequences changing over time, contributing to organism complexity (Mazumder, et al., 2003). UTR length has also been attributed to cellular proliferation, with shorter UTRs observed in cell lines and tumor cells relative to untransformed tissue (Doran, 2008). The regulation of gene products is an important facet of an organism and has been associated with different regions of the gene, including 5' and 3' un-translated regions. Variations in coding and noncoding sequence length, intron number and size differ significantly among living organisms. It would be beneficial to identify additional examples of the noncoding regions influence, in a diverse range of model organisms, to extend the current understanding.

1.3 Motivation and thesis outline

Bioinformatics has become a major discipline not just a “tool kit”, in the post-genomic era. The need for computational methods, statistics, data storage, data mining and analysis after the Human Genome project to deal with the large influx of sequence data, drove the formation of this discipline. Since then, scientists have been able to answer many fundamental biological questions, not just from a biology standpoint, but view the data from a mathematical computer analysis perspective (Webb, 2011).

This thesis presents length data and statistical methodology generally on two model organisms, (one plant and one animal) addressing the following questions:

- Is there a relationship between the distributions of coding and noncoding regions of protein coding genes?
- Is there a relationship between the length distribution of each gene region in a protein coding gene, in relation to protein function and gene expression?

- How can mathematical modelling, statistics and computer algorithms help us to observe patterns and trends that we can associate to biological and functional processes?

Many studies have only looked at protein length or UTRs, and there is a great deal of contention between results. Little research to date has combined the coding and noncoding regions in comparative studies among animal and plant species, to either confirm or refute previous research. Upstream regions in a gene have been an important part in the initiation of translation for gene expression. Little research has focused on the interrelationships between these regulatory elements, with most research focusing on the elements themselves.

As genome sequence data becomes readily available for different living organisms, and the explosion of data from biological experiments, there is a greater need for automated tools to classify and analysis this data, as well as increasing the scale and sophistication of the information technology, in order to draw conclusions from the data and to formulate new directions for research. The main aim of this thesis is to explore and understand, using statistics and mathematical modelling, the length distribution relationship between the coding and noncoding regions of protein coding genes.

The gene length of protein coding genes were divided into three sections, and data was collected for each region including or excluding introns. The distances were measured in base pairs (bp) of the nucleotide sequence. As shown in Figure 1.6, the first region is situated between the Translation Start Site (TLS) and the Translation Stop Codon (TSC). This region will be referred to as D_1, or coding region length (TLS-TSC distance) for the rest of this thesis. The second region encompasses the +1 position after the promoter (the Transcription Start Site (TSS)) to the last nucleotide before the TLS. This region will be referred to as D_2 (TSS-TLS distance). The third region is situated between the translation stop codon (TSC) and the Transcription Termination Site (TTS), and will be referred to as D_3 (TSC-TTS distance). The data collected without introns will be denoted as d_1, d_2 and d_3 respectively.

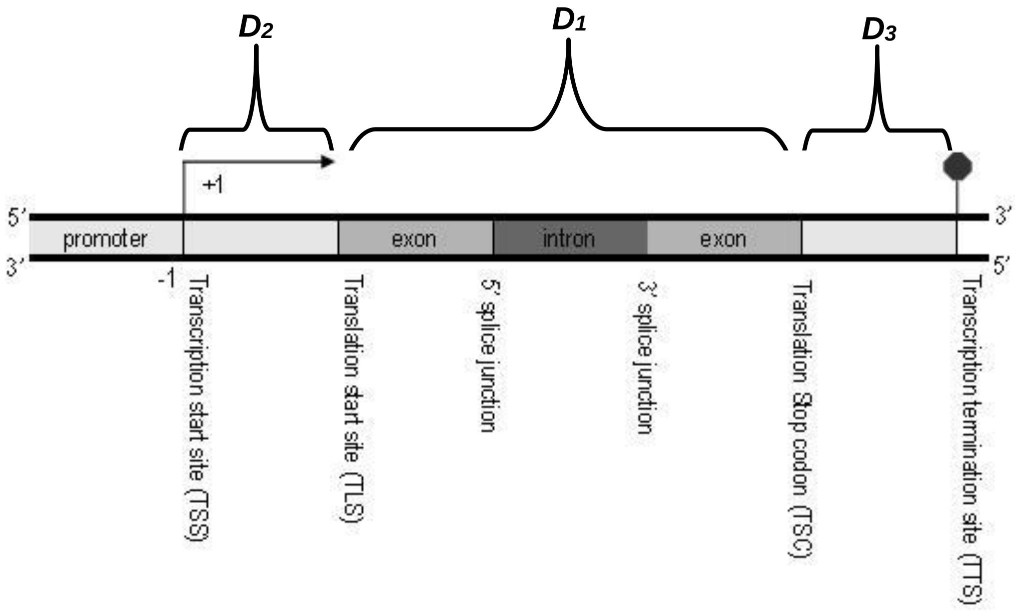

D2
D1
D3
+1
5'
3'
promoter
exon
intron
exon
3'
5'
-1
Transcription start site (TSS)
Translation start site (TLS)
5' splice junction
3' splice junction
Translation Stop codon (TSC)
Transcription termination site (TTS)

Figure 1.6 Diagrammatic illustration of a structural gene, including introns, and areas of interest in this study.

The project involved data acquirement from the internet, data formatting and creation of a database for the research, pattern search for target DNA elements, followed by the examination of the interrelationships between these regulatory elements. Specifically, in Chapter 5 we introduce a nonlinear model to investigate the relationships between the coding and noncoding regions with two model organisms, *Arabidopsis* and *Drosophila* including protein function. Chapter 6 we introduce gene expression data into the analysis and use *Arabidopsis* as a case study. Chapter 7 explores the use of Canonical Correlation Analysis (CCA) using *Drosophila* as a case study. And finally, Chapter 8 progresses into more complex analysis, introducing quantile regression analysis, with the aim at comparing the animal and plant species in relation to length and gene expression. These chapters include published work that have been peer reviewed. The chapter format will include a brief introduction with information not included in the main introduction, statistical analysis that is not outlined in the data collection chapter, with a results and discussion section. The conclusions of all the results will be discussed in Chapter 9. The outcome of this project will not only offer a better understanding of the correlation of gene expression / function and gene architecture with regards to the length distribution in the coding and noncoding regions of a protein coding gene, but also help develop better tools to analyse the data.

It is important to note that this research project was conducted part-time over a 9 year period. Data was collected annually, as new releases of data occurred on a regular basis. Data was collected at the beginning of each year (from 2007-2015), before more analysis was conducted. A history of the data was stored on external drives as a reference point. The early chapters were written at the primary stages of this thesis and reflect the available data at the time.

Chapter 2 - Data Collection of Organisms Studied

2 Data Collection of Organisms Studied

A great deal of data has been collected, documented and published on numerous organisms, making them “Model Organisms”. Model organisms can be used to gain information indirectly about other species that may be difficult and time consuming to study. Many organisms are listed as model organisms, and cover the 3 kingdoms (fungi, plants, animals). Most of the organisms listed as “model” have extensive genomic research data available and have been studied for many years.

For the initial study, an understanding and confirmation of previous research on the coding sequence was conducted on fifteen organisms (Figure 2.1) which were used to compare the coding sequence data with and without introns.

The reasoning behind selecting these organisms was the availability of data in the early stage of this thesis. CDS data was easily obtainable, for these organisms, however limited data on the UTRs restricted the number of organisms selected. For the majority of research in this thesis, two major model organisms were examined for several reasons:

- They cover a good range in the evolutionary tree;
- The Genome sequence has been completed many years ago;
- Large amounts of data is available publicly from the World Wide Web;
- Many studies have already been completed, including comparative studies for these organisms and have their data verified in peer reviewed publications;
- Data for protein function, CDS, 5’ and 3’ UTR and gene expression data was readily available.

The organisms selected for extensive study included:

- *Arabidopsis thaliana* (Thale Cress); and
- *Drosophila melanogaster* (Fruit Fly).

Data for this project was obtained over many different databases, imported from FTP sites from various research centres. By merging these data together I have contributed to the bioinformatics topic by automating the cleaning process, and the ability to analysis data that had not previously been combined. Researchers that are wanting to study the coding and untranslated regions of protein coding genes would be able to use the model organism databases created for this thesis.

The organisms selected for this research cover two major branches of the eukarya domain, from a simple plant species to the higher animal species.

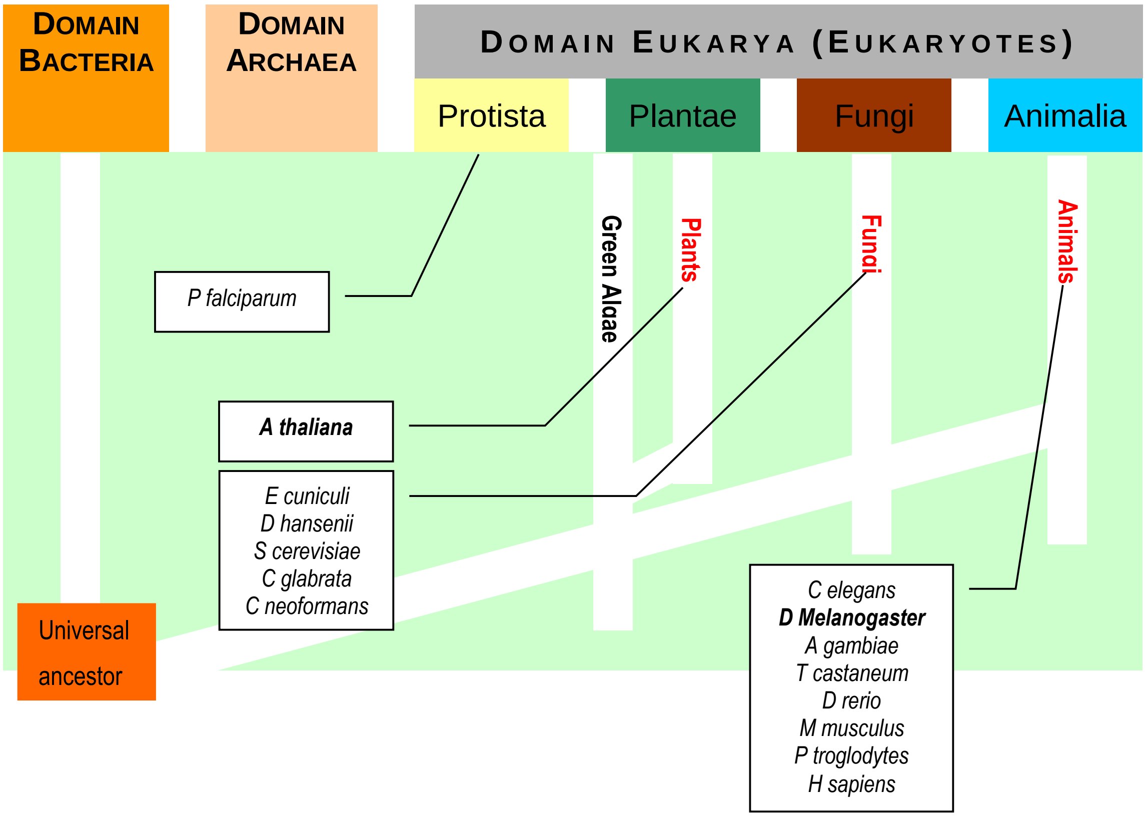
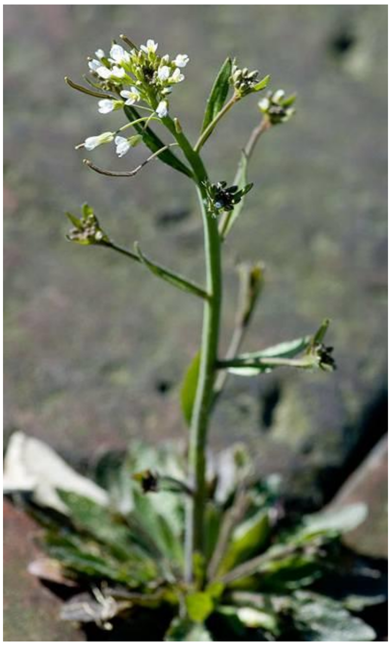


Figure 2.1 Phylogeny of Eukaryotes

Three domains of life, Bacteria, Archaea and Eukarya from one universal ancestor (Keeling, et al., 2009).

2.1 *Arabidopsis thaliana*



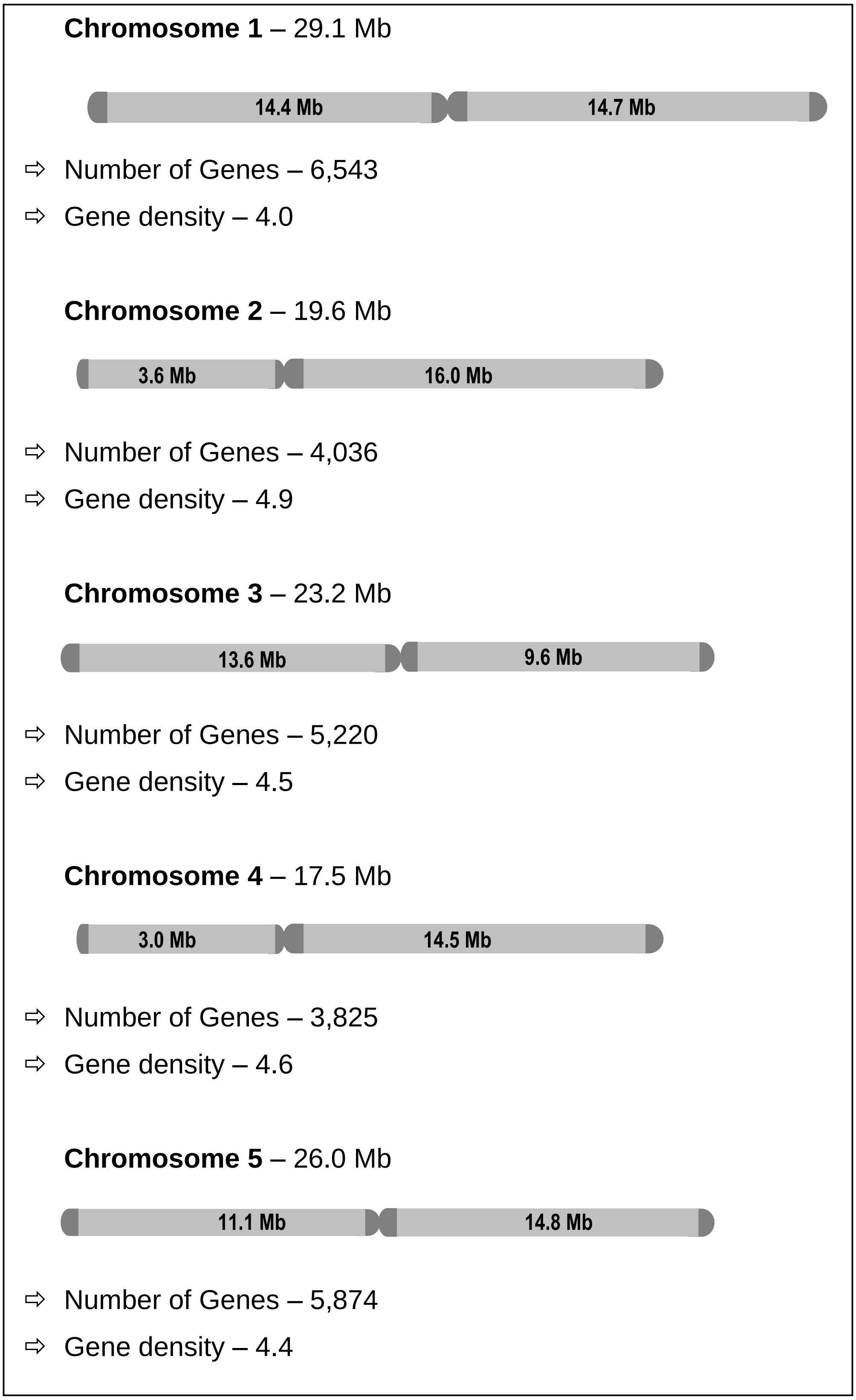
Kingdom:	Plantae
(unranked):	Angiosperms
(unranked):	Eudicots
(unranked):	Rosids
Order:	Brassicales
Family:	Brassicaceae
Genus:	Arabidopsis
Species:	A. thaliana

Figure 2.2 Image of Arabidopsis thaliana

Arabidopsis thaliana (Figure 2.2) known as thale cress, or mouse-ear cress, is a small flowering plant which is a member of the mustard family and native to Europe, Asia and north western Africa. It was the first plant genome to be sequenced, and has been studied extensively. Research with this species has involved many plant biology and genetic studies, making it a perfect model organism for multiple disciplines (Meinke, et al., 1998). The plant's rapid life cycle, and relatively small genome has also made this a popular organism for study. The information gained from the sequencing data has contributed to a generalized view on plant genes, and understanding of the molecular biology of many plant traits, including plant development (Takano, et al., 2006; Vanneste and Friml, 2009) and light sensing (Cheng, et al., 2003). This unprecedented resource has accelerated not only plant research but has had beneficial effects on health science research (Jones, et al., 2008; Xu and Møller, 2011) and agriculture and crop development (Ferrier, et al., 2011; Gonzalez, et al., 2009)

2.1.1 *Arabidopsis thaliana* Genome

The *Arabidopsis thaliana* genome consists of 5 chromosomes (Figure 2.3), with the sequence region spanning ~115.4 megabases (Mb). In 2000 the genome contained 25,498 genes encoding proteins from 11,000 families (Initiative, 2000) with several releases after this first count. The protein functional classification range from cellular metabolism to protein synthesis, and is similar to the functional diversity found in the *Drosophila* species.



Chromosome 1 – 29.1 Mb

- ⇨ Number of Genes – 6,543
- ⇨ Gene density – 4.0

Chromosome 2 – 19.6 Mb

- ⇨ Number of Genes – 4,036
- ⇨ Gene density – 4.9

Chromosome 3 – 23.2 Mb

- ⇨ Number of Genes – 5,220
- ⇨ Gene density – 4.5

Chromosome 4 – 17.5 Mb

- ⇨ Number of Genes – 3,825
- ⇨ Gene density – 4.6

Chromosome 5 – 26.0 Mb

- ⇨ Number of Genes – 5,874
- ⇨ Gene density – 4.4

Figure 2.3 The genome structure of the Arabidopsis thaliana *separated into chromosomes*

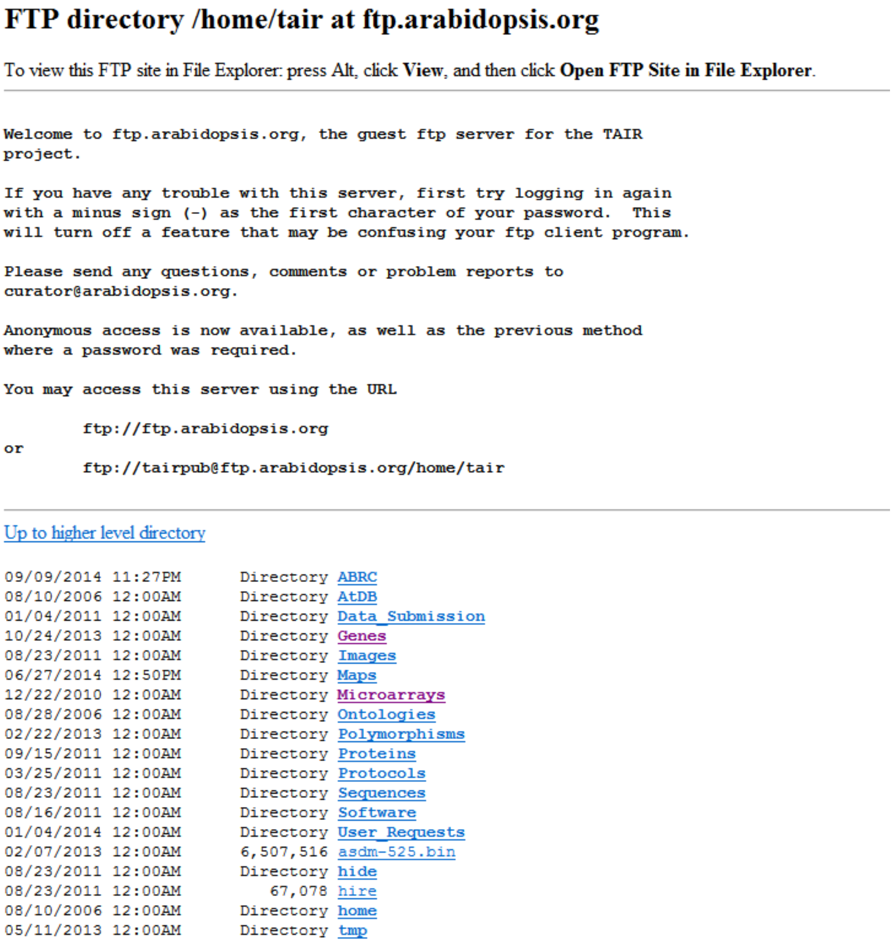
The genome consists of 5 chromosomes (1 to 5). The numbers given correspond to their lengths in megabases (Mb) (Initiative, 2000)

2.1.2 Gene Number

The TAIR Consortium (Rhee, et al., 2003) current data release is version 10 and contains 27,416 protein coding genes. The number of genes used in the research conducted in this thesis differs to this number as a result of available data for the untranslated regions, protein function, and gene expression data.

2.1.3 Coding and Untranslated regions

Arabidopsis coding, untranslated regions and gene expression data was downloaded from the TAIR FTP site: ftp://ftp.arabidopsis.org/home/tair (Figure 2.4)



FTP directory /home/tair at ftp.arabidopsis.org

To view this FTP site in File Explorer: press Alt, click **View**, and then click **Open FTP Site in File Explorer**.

```
Welcome to ftp.arabidopsis.org, the guest ftp server for the TAIR
project.

If you have any trouble with this server, first try logging in again
with a minus sign (-) as the first character of your password.  This
will turn off a feature that may be confusing your ftp client program.

Please send any questions, comments or problem reports to
curator@arabidopsis.org.

Anonymous access is now available, as well as the previous method
where a password was required.

You may access this server using the URL

        ftp://ftp.arabidopsis.org
or
        ftp://tairpub@ftp.arabidopsis.org/home/tair
```

Up to higher level directory

```
09/09/2014 11:27PM      Directory ABRC
08/10/2006 12:00AM      Directory AtDB
01/04/2011 12:00AM      Directory Data_Submission
10/24/2013 12:00AM      Directory Genes
08/23/2011 12:00AM      Directory Images
06/27/2014 12:50PM      Directory Maps
12/22/2010 12:00AM      Directory Microarrays
08/28/2006 12:00AM      Directory Ontologies
02/22/2013 12:00AM      Directory Polymorphisms
09/15/2011 12:00AM      Directory Proteins
03/25/2011 12:00AM      Directory Protocols
08/23/2011 12:00AM      Directory Sequences
08/16/2011 12:00AM      Directory Software
01/04/2014 12:00AM      Directory User_Requests
02/07/2013 12:00AM      6,507,516 asdm-525.bin
08/23/2011 12:00AM      Directory hide
08/23/2011 12:00AM         67,078 hire
08/10/2006 12:00AM      Directory home
05/11/2013 12:00AM      Directory tmp
```

Figure 2.4 TAIR FTP site for downloading Arabidopsis *data*

All data was downloaded as text (.txt) files and cleaned by running a visual basic script. See appendix A for script details.

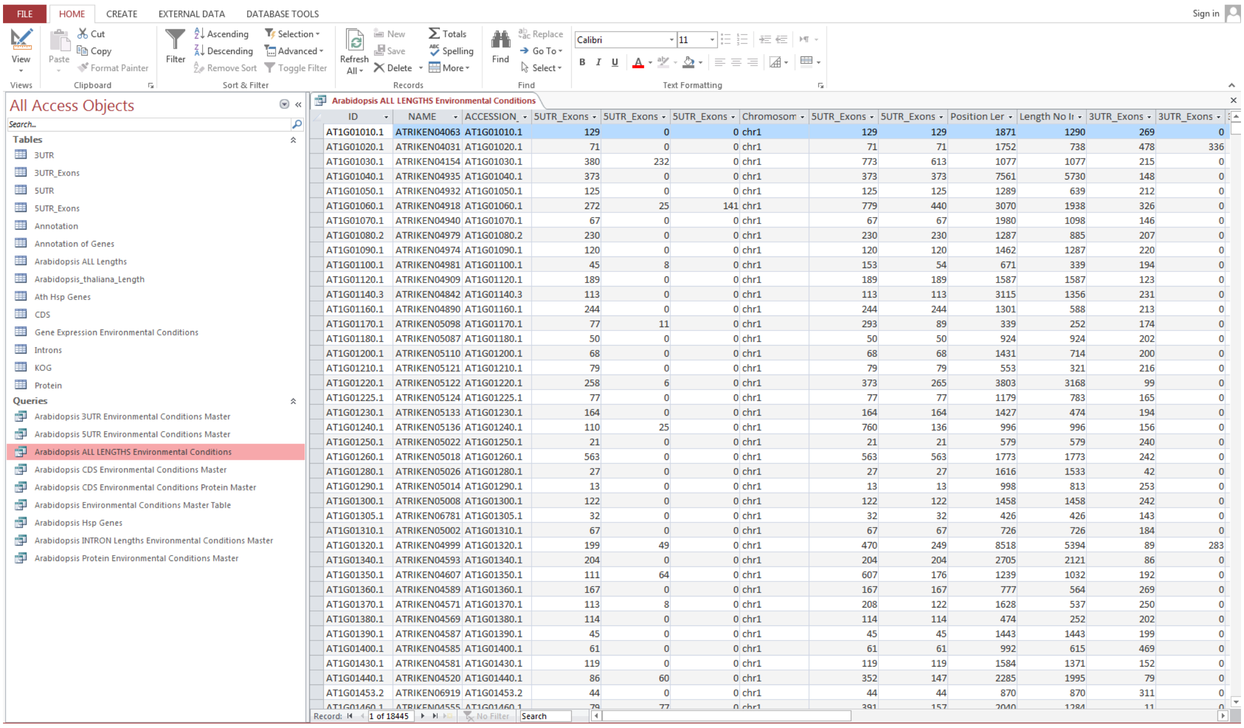
2.1.4 Gene Expression Data

Gene expression data was downloaded from two online databases. The first expression set was downloaded from the TAIR FTP site which was an average of all Arabidopsis

Functional Genomics Consortium (AFGC) microarray experiments. The average intensity values represented in this dataset was a large range of conditions and tissue types. To focus on environmental conditions and a control sample, a set of gene expression data was downloaded from the NCBI GEO Datasets database (series GSE 34188) including the annotation files (Hanada, et al., 2013). Other gene expression data was downloaded from the NCBI GEO Datasets database, and is outlined in Chapter 8.

As data is constantly being renewed on these databases, review and modification of the files was performed on a regular basis to keep up to date with the current sequencing data. Data was accessed on a yearly basis and updated usually coinciding with new analysis techniques and hypothesis testing. It is important to note that sample sizes may vary throughout the course of this thesis, due to the time of download and the analysis conducted.

The tables from all data sources were linked with the Accession number to merge all the data into one master table in Microsoft Access. An example of the MS Access master database is shown in Figure 2.5 (Powell, et al., 2010). The database contained tables and queries and can be used to extract information from the databases as new hypotheses and statistical tools are formulated.



Arabidopsis ALL LENGTHS Environmental Conditions

ID	NAME	ACCESSION_	5UTR_Exons	5UTR_Exons	5UTR_Exons	Chromosom	5UTR_Exons	5UTR_Exons	Position Ler	Length No Ir	3UTR_Exons	3UTR_Exons
AT1G01010.1	ATRIKEN04063	AT1G01010.1	129	0	0	chr1	129	129	1871	1290	269	0
AT1G01020.1	ATRIKEN04031	AT1G01020.1	71	0	0	chr1	71	71	1752	738	478	336
AT1G01030.1	ATRIKEN04154	AT1G01030.1	380	232	0	chr1	773	613	1077	1077	215	0
AT1G01040.1	ATRIKEN04935	AT1G01040.1	373	0	0	chr1	373	373	7561	5730	148	0
AT1G01050.1	ATRIKEN04932	AT1G01050.1	125	0	0	chr1	125	125	1289	639	212	0
AT1G01060.1	ATRIKEN04918	AT1G01060.1	272	25	141	chr1	779	440	3070	1938	326	0
AT1G01070.1	ATRIKEN04940	AT1G01070.1	67	0	0	chr1	67	67	1980	1098	146	0
AT1G01080.2	ATRIKEN04979	AT1G01080.2	230	0	0	chr1	230	230	1287	885	207	0
AT1G01090.1	ATRIKEN04974	AT1G01090.1	120	0	0	chr1	120	120	1462	1287	220	0
AT1G01100.1	ATRIKEN04981	AT1G01100.1	45	8	0	chr1	153	54	671	339	194	0
AT1G01120.1	ATRIKEN04909	AT1G01120.1	189	0	0	chr1	189	189	1587	1587	123	0
AT1G01140.3	ATRIKEN04842	AT1G01140.3	113	0	0	chr1	113	113	3115	1356	231	0
AT1G01160.1	ATRIKEN04890	AT1G01160.1	244	0	0	chr1	244	244	1301	588	213	0
AT1G01170.1	ATRIKEN05098	AT1G01170.1	77	11	0	chr1	293	89	339	252	174	0
AT1G01180.1	ATRIKEN05087	AT1G01180.1	50	0	0	chr1	50	50	924	924	202	0
AT1G01200.1	ATRIKEN05110	AT1G01200.1	68	0	0	chr1	68	68	1431	714	200	0
AT1G01210.1	ATRIKEN05121	AT1G01210.1	79	0	0	chr1	79	79	553	321	216	0
AT1G01220.1	ATRIKEN05122	AT1G01220.1	258	6	0	chr1	373	265	3803	3168	99	0
AT1G01225.1	ATRIKEN05124	AT1G01225.1	77	0	0	chr1	77	77	1179	783	165	0
AT1G01230.1	ATRIKEN05133	AT1G01230.1	164	0	0	chr1	164	164	1427	474	194	0
AT1G01240.1	ATRIKEN05136	AT1G01240.1	110	25	0	chr1	760	136	996	996	156	0
AT1G01250.1	ATRIKEN05022	AT1G01250.1	21	0	0	chr1	21	21	579	579	240	0
AT1G01260.1	ATRIKEN05018	AT1G01260.1	563	0	0	chr1	563	563	1773	1773	242	0
AT1G01280.1	ATRIKEN05026	AT1G01280.1	27	0	0	chr1	27	27	1616	1533	42	0
AT1G01290.1	ATRIKEN05014	AT1G01290.1	13	0	0	chr1	13	13	998	813	253	0
AT1G01300.1	ATRIKEN05008	AT1G01300.1	122	0	0	chr1	122	122	1458	1458	242	0
AT1G01305.1	ATRIKEN06781	AT1G01305.1	32	0	0	chr1	32	32	426	426	143	0
AT1G01310.1	ATRIKEN05002	AT1G01310.1	67	0	0	chr1	67	67	726	726	184	0
AT1G01320.1	ATRIKEN04999	AT1G01320.1	199	49	0	chr1	470	249	8518	5394	89	283
AT1G01340.1	ATRIKEN04593	AT1G01340.1	204	0	0	chr1	204	204	2705	2121	86	0
AT1G01350.1	ATRIKEN04607	AT1G01350.1	111	64	0	chr1	607	176	1239	1032	192	0
AT1G01360.1	ATRIKEN04589	AT1G01360.1	167	0	0	chr1	167	167	777	564	269	0
AT1G01370.1	ATRIKEN04571	AT1G01370.1	113	8	0	chr1	208	122	1628	537	250	0
AT1G01380.1	ATRIKEN04569	AT1G01380.1	114	0	0	chr1	114	114	474	252	202	0
AT1G01390.1	ATRIKEN04587	AT1G01390.1	45	0	0	chr1	45	45	1443	1443	199	0
AT1G01400.1	ATRIKEN04585	AT1G01400.1	61	0	0	chr1	61	61	992	615	469	0
AT1G01430.1	ATRIKEN04581	AT1G01430.1	119	0	0	chr1	119	119	1584	1371	152	0
AT1G01440.1	ATRIKEN04520	AT1G01440.1	86	60	0	chr1	352	147	2285	1995	79	0
AT1G01453.2	ATRIKEN06919	AT1G01453.2	44	0	0	chr1	44	44	870	870	311	0
AT1G01460.1	ATRIKEN04555	AT1G01460.1	79	77	0	chr1	391	157	2040	1284	11	0

Figure 2.5 Arabidopsis thaliana *Microsoft Access Master Database.*

The master database was used to merge all length, gene expression and protein data together for easy querying and export to Excel & SPSS for data analysis.

2.2 Drosophila melanogaster

Figure 2.6 Image of Drosophila melanogaster

Kingdom: Animalia

Phylum: Arthropoda

Class: Insecta

Order: Diptera

Family: Drosophilidae

Genus: Drosophila

Subgenus: Sophophora

Species group: melanogaster group

Species subgroup: melanogaster subgroup

Species complex: melanogaster complex

Species: *D. melanogaster*

With the introduction of the *Drosophila melanogaster* by William Castle almost a decade ago, this organism has become one of the most important model organisms studied to date especially in the field of genetics. The completion of the fly genome in 2000 has extended scientists understanding in the study of transcription, protein binding, and genetic variation and illustrates the enormity this data can offer (Celniker and Rubin, 2003). The sequencing of the Drosophila's genome set precedence on the use of the whole-genome shotgun (WGS) sequencing method, which had only been successfully tried on bacterial genomes, not large more complex genomes. Shotgun sequencing is used when large DNA strands are the focus. The fly genome project demonstrated this method in the study of the *Drosophila melanogaster* species (Ashburner and Bergman, 2005).

Drosophila melanogaster is an excellent model system which continues to be used extensively in human health studies. Recent research has included using *Drosophila* as a model in human disease therapeutic drug discovery (Pandey and Nichols, 2011) and pathogenic human viruses (Hughes, et al., 2012), as well as to understand the genetics and pathology of human CoQ deficiencies (Fernández-Ayala, et al., 2014). It has also been used to identify the health benefits of organically grown foods (Chhabra, et al., 2013).

2.2.1 *Drosophila melanogaster* Genome

The genome of the *Drosophila melanogaster* consists of the sex chromosomes X and Y, left and right arms of chromosomes 2 and 3 (2L, 2R, 3L and 3R) and a small 4th chromosome. The size of the genome is approximately 180 megabases (Mb) and segmented by two-thirds euchromatin[5] and one-third heterochromatin[6]. The protein-coding genes are represented in the euchromatin (Celniker and Rubin, 2003). 98% of the protein-coding genes are found in the genome. The genome structure of the *Drosophila melanogaster* is outlined in Figure 2.7 and shows the composition of the chromosomes and lengths of each section in megabases (Celniker and Rubin, 2003).

2.2.2 Gene Number

The sequencing of the *Drosophila melanogaster* genome was published by Celera Genomics and the Berkeley Drosophila Genome Project (BDGP) with several releases of updated data. In 2003 Celniker & Rubin (2003) published the number of genes from this collaboration, which was reported at 13,676.

[5] Euchromatin is a lightly packed form of chromatin that is rich in gene concentration, and is often (but not always) under active transcription. It is found in both eukaryotes and prokaryotes.

[6] Heterochromatin is a tightly packed form of DNA. Its major characteristic is that transcription is limited. As such, it is a means to control gene expression, through regulation of the transcription initiation.

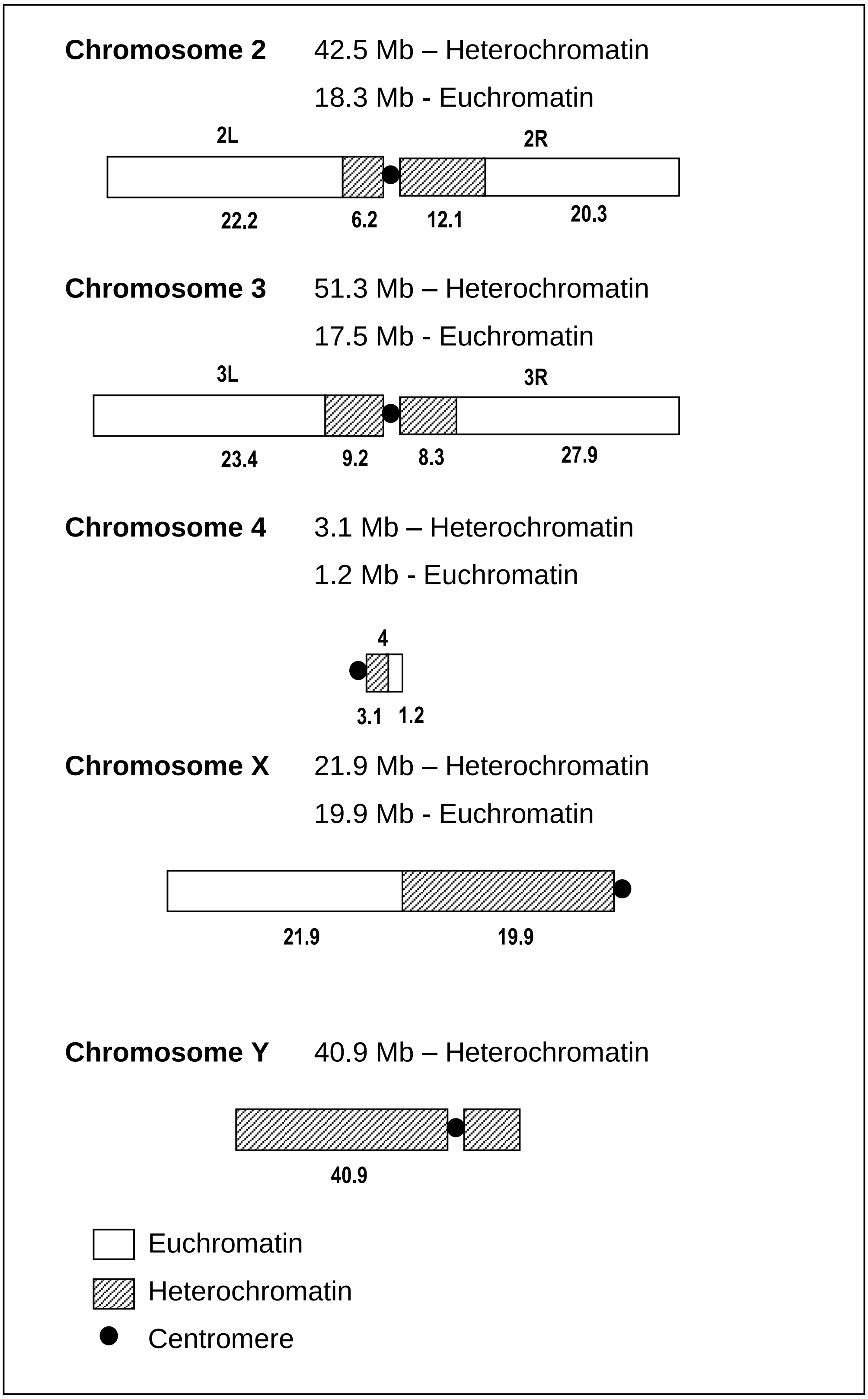


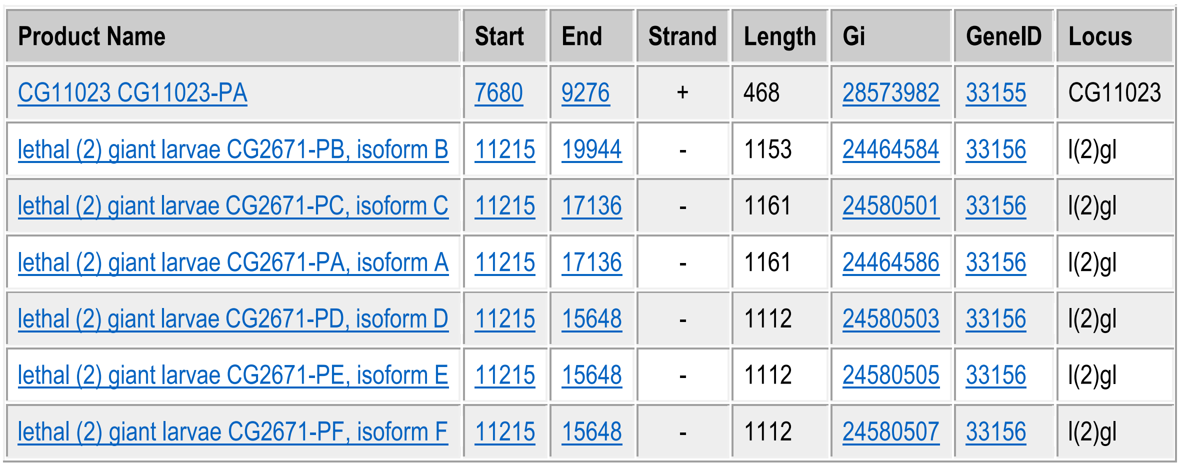
Figure 2.7 The genome structure of the Drosophila melanogaster *separated into chromosomes*

The genome consists of 5 chromosomes, which includes the sex (X & Y) chromosomes, left and right arms of chromosomes 2 and 3, and a small 4th chromosome. The numbers given below the chromosomes correspond to their lengths in megabases (Mb) (Celniker and Rubin, 2003).

Molecular identification of genes on the Y chromosome of *Drosophila melanogaster* is difficult because the entire chromosome is heterochromatic. Approximately 80% of Y chromosome DNA is composed of nine simple repeated sequences, including (AAGAC)n (8 Mb), (AAGAG)n (7 Mb), and (AATAT)n (6 Mb) (102) (Celniker and Rubin, 2003). For this reason, chromosome comparisons for *Drosophila* for chromosome Y are absent from the analysis.

Length data was downloaded from the RefSeq NCBI database, and within the tables exported, contains information pertaining to each protein coding gene such as Start Position of CDS; End Position of CDS; Protein length; Gene Product; and Gene Product ID. This is illustrated in the table below:

Table 2.1 NCBI RefSeq table for each chromosome. The data contains a list of protein gene information



Product Name	Start	End	Strand	Length	Gi	GeneID	Locus
CG11023 CG11023-PA	7680	9276	+	468	28573982	33155	CG11023
lethal (2) giant larvae CG2671-PB, isoform B	11215	19944	-	1153	24464584	33156	l(2)gl
lethal (2) giant larvae CG2671-PC, isoform C	11215	17136	-	1161	24580501	33156	l(2)gl
lethal (2) giant larvae CG2671-PA, isoform A	11215	17136	-	1161	24464586	33156	l(2)gl
lethal (2) giant larvae CG2671-PD, isoform D	11215	15648	-	1112	24580503	33156	l(2)gl
lethal (2) giant larvae CG2671-PE, isoform E	11215	15648	-	1112	24580505	33156	l(2)gl
lethal (2) giant larvae CG2671-PF, isoform F	11215	15648	-	1112	24580507	33156	l(2)gl

The RefSeq table for the *D. melanogaster* was interpreted from the genome data submitted from the FlyBase Consortium.

The data and tables for each chromosome were exported to a Microsoft Excel spreadsheet where it was formatted and cleaned using a macro. The macro script can be found in Appendix A.

The D1 / d1 (coding sequence with and without introns) values were calculated from the data obtained from the RefSeq NCBI website. The calculations were incorporated in the macro and were calculated during the cleaning and formatting on each of the excel files exported. The Excel formula included:

D1 – calculation =   End Position subtracted by Start Position + 1 (with introns)

d1 – calculation =   Protein Length X 3 + 3 (without introns)

2.2.3 Coding and noncoding regions

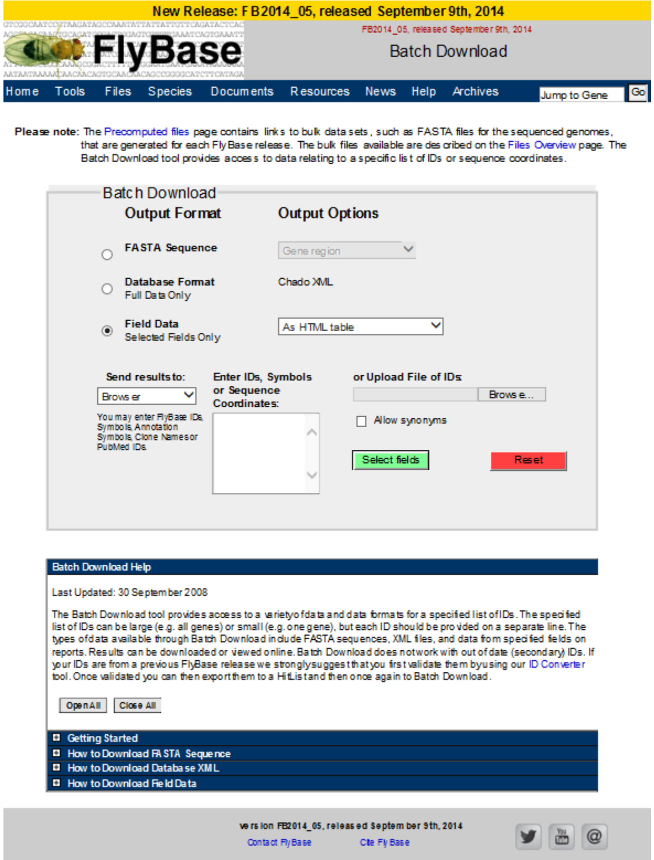
Data for the regions d2 and d3 were collected from the Flybase Consortium (http://flybase.org/). Steps taken to collect this data included:

1. Copy the product ID retrieved from the Refseq data to Excel and extract the product name. For example: CG10417

2. Copy these product names in the “Enter List of IDs:” at the following site:

   http://flybase.bio.indiana.edu/static_pages/downloads/ID.html

FASTA Genome Sequence output format was selected, and gives the researcher options to select the section of the genome of interest, for example 5’ UTR (Figure 2.8).



New Release: FB2014_05, released September 9th, 2014

FB2014_05, released September 9th, 2014

Batch Download

Please note: The Precomputed files page contains links to bulk data sets, such as FASTA files for the sequenced genomes, that are generated for each FlyBase release. The bulk files available are described on the Files Overview page. The Batch Download tool provides access to data relating to a specific list of IDs or sequence coordinates.

Batch Download

Output Format — Output Options

○ FASTA Sequence — Gene region

○ Database Format, Full Data Only — Chado XML

● Field Data, Selected Fields Only — As HTML table

Send results to: Browser

You may enter FlyBase IDs, Symbols, Annotation Symbols, Clone Names or PubMed IDs.

Enter IDs, Symbols or Sequence Coordinates:

or Upload File of IDs:

□ Allow synonyms

Batch Download Help

Last Updated: 30 September 2008

The Batch Download tool provides access to a variety of data and data formats for a specified list of IDs. The specified list of IDs can be large (e.g. all genes) or small (e.g. one gene), but each ID should be provided on a separate line. The types of data available through Batch Download include FASTA sequences, XML files, and data from specified fields on reports. Results can be downloaded or viewed online. Batch Download does not work with out of date (secondary) IDs. If your IDs are from a previous FlyBase release we strongly suggest that you first validate them by using our ID Converter tool. Once validated you can then export them to a HitList and then once again to Batch Download.

⊞ Getting Started
⊞ How to Download FASTA Sequence
⊞ How to Download Database XML
⊞ How to Download Field Data

version FB2014_05, released September 9th, 2014

Figure 2.8 Batch download from the FlyBase website for data collection

3. Data can be saved as a text file once the table is launched in the selected internet browser, with the queries you select in the batch download.

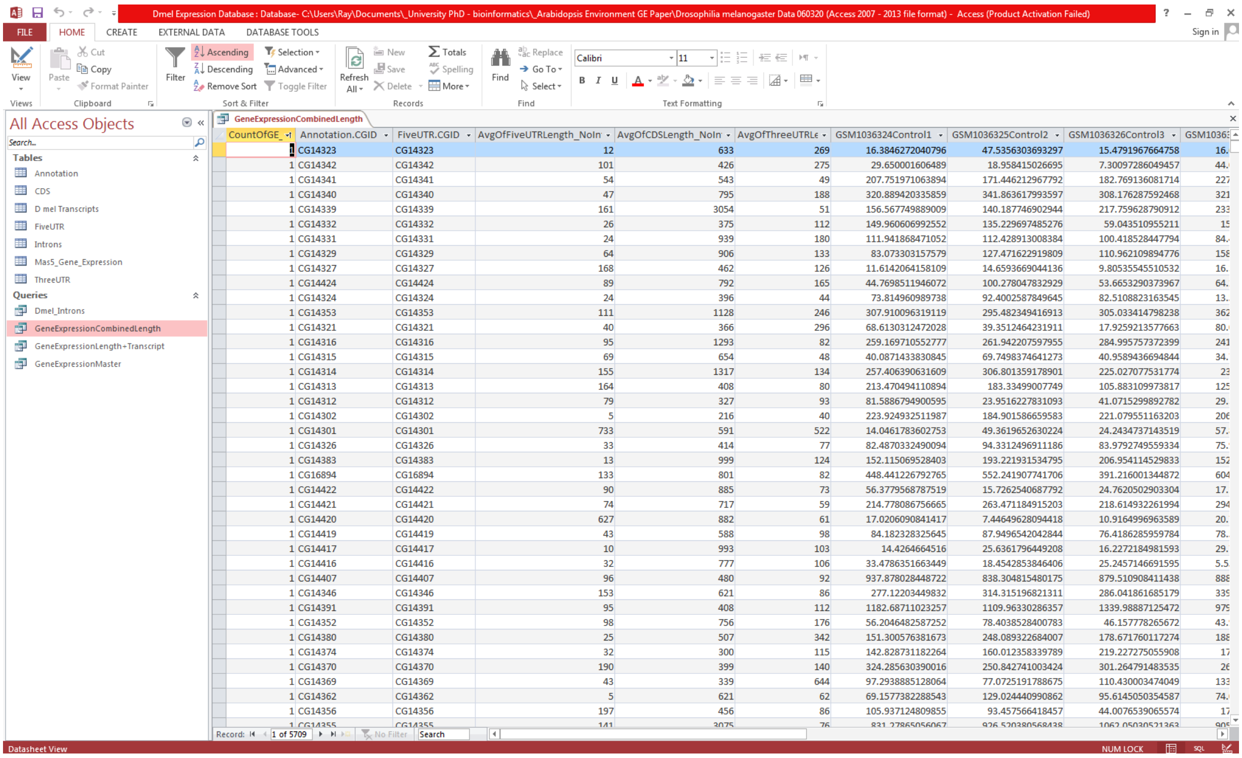
4. The text file can be used and imported into Microsoft Access for manipulation and query purposes.

2.2.4 Genome Sequence data

The Genome Sequence data was downloaded from: http://www.fruitfly.org/sequence/download.html. The file format was a FASTA.gz zip file (na_arm2L_genomic_dmel_RELEASE4.FASTA.gz). This file can be viewed in notepad or MS Excel as a text file. This data was used to reference the cDNA data positions in the genome sequence to confirm and identify the positions of each region of interest.

As data is constantly being updated to these databases, review and modification of the files were performed on a regular basis to keep up to date with the current sequencing and functional data. Variation in sample size was dependent on the year of download and the analysis conducted on the data.

The tables from all data sources were linked with the CG ID to merge all the data into one master database in Microsoft Access. An example of the MS Access master database is shown in Figure 2.9.



CountOfGE_	Annotation.CGID	FiveUTR.CGID	AvgOfFiveUTRLength_NoIn	AvgOfCDSLength_NoIn	AvgOfThreeUTRLe	GSM1036324Control1	GSM1036325Control2	GSM1036326Control3	GSM1036
1	CG14323	CG14323	12	633	269	16.3846272040796	47.5356303693297	15.4791967664758	16.
1	CG14342	CG14342	101	426	275	29.650001606489	18.958415026695	7.30097286049457	44.
1	CG14341	CG14341	54	543	49	207.751971063894	171.446212967792	182.769136081714	227
1	CG14340	CG14340	47	795	188	320.889420335859	341.863617993597	308.176287592468	321
1	CG14339	CG14339	161	3054	51	156.567749889009	140.187746902944	217.759628790912	233
1	CG14332	CG14332	26	375	112	149.960606992552	135.229697485276	59.043510955211	15
1	CG14331	CG14331	24	939	180	111.941868471052	112.428913008384	100.418528447794	84.
1	CG14329	CG14329	64	906	133	83.073303157579	127.471622919809	110.962109894776	158
1	CG14327	CG14327	168	462	126	11.6142064158109	14.6593669044136	9.80535545510532	16.
1	CG14424	CG14424	89	792	165	44.7698511946072	100.278047832929	53.6653290373967	64.
1	CG14324	CG14324	24	396	44	73.814960989738	92.4002587849645	82.5108823163545	13.
1	CG14353	CG14353	111	1128	246	307.910096319119	295.482349416913	305.033414798238	362
1	CG14321	CG14321	40	366	296	68.6130312472028	39.3512464231911	17.9259213577663	80.
1	CG14316	CG14316	95	1293	82	259.169710552777	261.942207597955	284.995757372399	241
1	CG14315	CG14315	69	654	48	40.0871433830845	69.7498374641273	40.9589436694844	34.
1	CG14314	CG14314	155	1317	134	257.406390631609	306.801359178901	225.027077531774	23
1	CG14313	CG14313	164	408	80	213.470494110894	183.33499007749	105.883109973817	125
1	CG14312	CG14312	79	327	93	81.5886794900595	23.9516227831093	41.0715299892782	29.
1	CG14302	CG14302	5	216	40	223.924932511987	184.901586659583	221.079551163203	206
1	CG14301	CG14301	733	591	522	14.0461783602753	49.3619652630224	24.2434737143519	57.
1	CG14326	CG14326	33	414	77	82.4870332490094	94.3312496911186	83.9792749559334	75.
1	CG14383	CG14383	13	999	124	152.115069528403	193.221931534795	206.954114529833	152
1	CG16894	CG16894	133	801	82	448.441226792765	552.241907741706	391.216001344872	604
1	CG14422	CG14422	90	885	73	56.3779568787519	15.7262540687792	24.7620502903304	17.
1	CG14421	CG14421	74	717	59	214.778086756665	263.471184915203	218.614932261994	294
1	CG14420	CG14420	627	882	61	17.0206090841417	7.44649628094418	10.9164996963589	20.
1	CG14419	CG14419	43	588	98	84.182328325645	87.9496542042844	76.4186285959784	78.
1	CG14417	CG14417	10	993	103	14.4264664516	25.6361796449208	16.2272184981593	29.
1	CG14416	CG14416	32	777	106	33.4786351663449	18.4542853846406	25.2457146691595	5.5
1	CG14407	CG14407	96	480	92	937.878028448722	838.304815480175	879.510908411438	888
1	CG14346	CG14346	153	621	86	277.12203449832	314.315196821311	286.041861685179	339
1	CG14391	CG14391	95	408	112	1182.68711023257	1109.96330286357	1339.98887125472	979
1	CG14352	CG14352	98	756	176	56.2046482587252	78.4038528400783	46.157778265672	43.
1	CG14380	CG14380	25	507	342	151.300576381673	248.089322684007	178.671760117274	188
1	CG14374	CG14374	32	300	115	142.828731182264	160.012358339789	219.227275055908	17
1	CG14370	CG14370	190	399	140	324.285630390016	250.842741003424	301.264791483535	26
1	CG14369	CG14369	43	339	644	97.2938885128064	77.0725191788675	110.430003474049	133
1	CG14362	CG14362	5	621	62	69.1577382288543	129.024440990862	95.6145050354587	74.
1	CG14356	CG14356	197	456	86	105.937124809855	93.457566418457	44.0076539065574	17
1	CG14355	CG14355	141	3075	76	831.27865056067	926.520380568438	1062.05030521363	905

Figure 2.9 MS Access master table containing all length data from publicly available sequencing data

Data was used throughout this thesis from databases and research organisations that had verified data. If the data integrity was questioned during my analysis, I had regular dialogue with the researchers from the primary sources of the data.

Chapter 3 – Promoter Prediction in Relation to Coding and Noncoding Sequences

3 Promoter Prediction in Relation to Coding and Noncoding Sequences

This chapter is slightly modified from the paper:

Caldwell, R., Dai, Y., Srivastava, S., Lin, Y., and Zhang, R. (2008) Improving neural network promoter prediction by exploiting the lengths of coding and noncoding sequences, Chapter: Advances of Computational Intelligence in Industrial Systems (Studies in Computational Intelligence) edited by Ying Liu, Aixin Sun, Han Tong Loh, Wen Feng and Ee-Peng Lim, Springer, 213-230.
doi:10.1007/978-3-540-78297-1_10.

3.1 Introduction

Much attention within computational biology research has focused on identifying gene products and locations from experimentally obtained DNA sequences. The use of promoter sequence prediction and positions of the transcription start sites can inevitably facilitate the process of gene finding in DNA sequences. This can be more beneficial if the organisms of interest are higher eukaryotes, where the coding regions of the genes are situated in an expanse of noncoding DNA.

With the genomes of numerous organisms now completely sequenced, there is a potential to gain invaluable biological information from these sequences. Computational prediction of promoters from the nucleotide sequences is one of the most attractive topics in sequence analysis today. Current promoter prediction algorithms employ several gene features for prediction. These attributes include homology with known promoters, the presence of particular motifs within the sequence, DNA structural characteristics and the relative signatures of different regions in the sequence.

3.1.1 Currently Used Algorithms

Different algorithms have been developed which vary in performance and can be categorized into two main groups. The first depends upon recognition of conserved signals such as the TATA box and the CCAAT box as well as the spacing between patterns. This approach uses either the neural network genetic algorithm or the weight matrix methodology. The second relies on identification of promoters within a sequence that may contain the elements. This approach is content-based and distinguishes differences such as triplet base-pair preferences around the transcription start site (TSS), and hexamer frequencies in consecutive 100-bp upstream regions (Qiu, 2003). There are also techniques that combine both these methods, which look for signals and for regions of specific compositions (Ohler and Niemann, 2001).

Many promoter prediction programs are readily available to the scientific community to utilize and explore. The programs that presented relatively high accuracy in their results include the GeneID / Promoter 1.0, TSSW, PromoterInspector and the Neural Network for Promoter Prediction (NNPP) (Burden, et al., 2005; Fickett and Hatzigeorgiou, 1997).

Currently, the Neural Network algorithm is probably the most widely used program in promoter prediction [http://www.fruitfly.org/seq_tools/promoter.html]. It is based on a time-delay neural network (TDNN) architecture that originated from speech recognition sequence patterns in time series. This method corresponds to how the brain's learning process operates. What makes this system unique is that it has the advantage of learning to recognize the degenerate patterns that characterize promoter motifs. The algorithm was initially designed for predicting promoters in the *Drosophila* genome and it has been developed to be a common method used to find both eukaryotic and prokaryotic promoters. The NNPP 2.2 algorithm recognizes only the presence and relative location of patterns and motifs within a promoter. It predicts the probability that a tested sequence position s ±3 base pairs (bp) contains a true TSS denoted by $P(s \in S)$, where S is the class of the true TSS positions (Burden, et al., 2005).

The popularity of NNPP has also been supported by comparative studies. An investigation by Fickett & Hatzigeorgiou recognized 13 of the 24 promoters (54%) in the test data set by NNPP and 31 false positives (1/1068 bp) were reported. These were significantly better than the outcomes of GeneID / Promoter 1.0 which identified 42% of the promoters and 51 false positives (1/649 bp) and the TSSW program (42% of true promoters and 42 false positives (1/789 bp)). Reese found similar results on the *Drosophila* genome, with a rate of 75% (69/92) of recognition and a rate of 1/547 bases of false positives (Reese, 2001).

3.1.2 Further Improvements in Promoter Prediction

Current algorithms to predict promoters are still far from satisfactory. The challenge that occurs in proposing a high level of prediction of promoters, with a reasonable percentage predicted, is that the level of falsely predicted promoters, known as false positives (FPs), is also high when a large percentage of predictions are met.

Another challenge faced which makes prediction difficult is that promoters are very diverse, and even some well-known signals such as TATA box and CCAAT box are not always conserved in all promoters. The TATA box can only be found in ~75% of vertebrates RNA Pol-II promoters and the CCAAT box is only found in half of vertebrate promoters (Qiu, 2003). Detectable motifs that exist within promoters can also occur randomly throughout the genome creating additional complications (Burden, et al., 2005). Promoters are defined based on functionality rather than structure, causing major impediments in creating near perfect predictions (Pandey and Krishnamachari, 2005). The promoter recognition systems for large-scale screening require acceptable ratios of true positives (TPs) and false positive predictions (i.e. those that maximize the TP recognition while minimize the FP recognition).

What currently is required out of these algorithms is the reduction in false positives in respect to promoter prediction. To achieve this it is possible to develop powerful computational methods and to replace current computational promoter prediction procedures. These approaches can be beneficial in increasing the accuracy of promoter prediction, and these changes are not restricted to just computational modifications. One approach in addressing these limitations is to investigate if the outcome of promoter prediction based on current techniques can be improved by incorporating additional information, such as the 5’ UTR sequence from the underlying DNA sequence.

The influx of DNA sequences, now publicly available, has allowed more and more information to be extracted. This has given computer and mathematical scientists the opportunity to run statistical analysis on this added information. The information gained will increase the understanding of the statistical behaviour of promoter positions for different genes across species. While much information can be integrated into any computational promoter prediction algorithms, our approach has been to exploit the distance information between gene elements. The study on *E. coli* (Burden, et al., 2005) was the first to investigate the use of the distance between TSS and TLS to improve the NNPP2.2 promoter prediction accuracy rate. Analysis and information retrieval performed by computers, particularly when dealing with large data sets has been an important tool for biologists. The information gained by these computations can guide biologists more efficiently in identifying areas of the DNA sequence experimentally infeasible without this data (Bajic, et al., 2004).

This chapter will summarize the TLS-NNPP approach and further extend the basic idea of the TLS-NNPP to more general circumstances with our more recent research results. The aim of this chapter is to firstly demonstrate why and how some measurements in DNA sequences can be used to significantly improve computational promoter prediction. And secondly it is intended to bring researcher’s attention to the DNA sequence information which is released through DNA sequence quantitative measurements instead of DNA sequence pattern information. For simplicity reasons, the research will only focus on the NNPP computational method as a reference method and demonstrate how DNA sequence quantitative measurements can be used to improve the promoter prediction of NNPP2.2. The technique discussed in this paper can be easily integrated with other computational promoter prediction algorithms by some minor modifications.

3.2 Gene Expression

In the process of transcription initiation, sets of genes can be turned on or off, determining each cell type, in response to different internal and external cues. The importance of transcriptional control is also associated with all forms of diseases, including cancer which is the improper regulation of the transcription of genes involved in cell growth (Hughes, 2006; Pedersen, et al., 1999; Qiu, 2003). Therefore, accurate prediction methods and understanding of these regions can be beneficial in human health in addition to computational biology.

The regulation of gene expression involves a complex molecular network with DNA-binding transcription factors (TFs) being an important element in this network. Most prokaryotes are unicellular organisms and promoters are recognized directly by RNA Polymerases, however eukaryotic organisms are more complex with the recognition of promoters consisting of large numbers of transcription control elements. One of the most complex processes found in molecular biology is the function of the promoter in transcription initiation. Promoters contain the nucleotide sequences which indicate the starting point for RNA synthesis. The promoter is positioned within the noncoding region upstream from the transcription start site which is referred to as the +1 position.

Apart from regulatory elements, other attributes of a gene such as its nucleotide composition, length, location (proximity to neighbours) and orientation may also play vital roles in gene expression. Genome size contrasts from organism to organism, and it appears that this divergence correlates with gene length variation.

3.3 Statistical Characteristics on Quantitative Measurements

The gene length can be divided into three sections, and for the purpose of this research the introns were included for each section, refer to Figure 1.1 in chapter 1. The distances were measured in base pairs (bp) of the nucleotide sequence. This information is just one of several attributes that could be utilized to improve promoter prediction in a variety of organisms.

The distances, TLS-TSC (*D_1*), TSS-TLS (*D_2*) and TSC-TTS (*D_3*) are varied, and can be considered as random components in gene sequences. The intention of this research is to contend that empirical information of these random components can benefit promoter prediction. Therefore, the aim of this research is to integrate this information with existing computational promoter prediction algorithms, and show that it will provide power to improve the prediction results. To understand why this information might help to improve computational promoter prediction, it is necessary to know the probability structure of these random components and see how much information is involved. Several model species ranging from bacteria to mammals will be used in this research to exploit the statistical information involved in the data. The species involved include *Escherichia coli* and *Bacillus subtilis* (bacterium), *Saccharomyces cerevisiae* (yeast), *Arabidopsis thaliana* (plant), *Mus musculus* and *Homo sapiens* (mammals).

To obtain the TLS-TSC (*D_1*) and TSS-TLS (*D_2*) distances, numerous databases were explored to determine absolute TSS, TLS and TSC positions on the various species genomes. The species chosen represent several model organisms that have been studied extensively, and possess a large amount of experimental data available to the public. The species were also chosen as they characterize a range of different classes, ranging from very simple organisms such as bacteria and yeast to the higher organisms such as the mammals.

TSS information was obtained from various databases, depending on the experimental research that had been conducted for each organism. The TSS information for *E. coli* was obtained from RegulonDB (Salgado, et al., 2006), *B. subtilis* data were obtained from the DBTBS database (Makita, et al., 2004), SCPD for *S. cerevisiae* (Zhu and Zhang, 1999), TAIR for *A. thaliana* (Garcia-Hernandez, et al., 2002) and DBTSS version 5.1.0 for both *M musculus* and *H. sapiens* (Suzuki, et al., 2004) Each of the TSS positions was considered to be positioned at multiple locations in a gene, thereby allowing multiple TSS-TLS distances to be generated. The *D_1* data was extracted from protein table files from the NCBI database.

In prokaryotes, the existence of operons is highly common. Therefore, in cases such as this, we regard the genes that are organized in one operon and controlled by the same promoter as separate gene units. Thus a single TSS-TLS distance may correspond to more than one coding region.

The statistical summary on the distances given by the six species was produced by the statistical package SPSS 12.0 / 15.0 and are presented in Table 3.1. The mean and median of each species was calculated for the distances between the TSS-TLS and TLS-TSC. The median was used for its simplicity and is not severely affected by extreme values (outliers) as is the mean value. Since the TSS positions have not been experimentally verified for all genes in an organism's genome, the sample size of *D_2* is relatively smaller as compared to *D_1*.

Table 3.1 Statistics of the distances (bp) of D_1 and D_2

Species	TLS-TSC distance (*D_1*)			TSS-TLS distance (*D_2*)		
	Sample Size	Median	Mean	Sample Size	Median	Mean
E. coli	4237	846	954	1017	66	164
B. subtilis	4015	771	896	483	67	93
S. cerevisiae	5850	1233	1503	202	68	110
A. thaliana	30480	1623	1939	20560	112	213
M. musculus	27132	10054	34552	14520	378	10913
H. sapiens	14796	16339	45445	14588	809	15291

The summary shows that the means of *D_1* and *D_2* are increasing as the species moves from a relatively simple organism to a more complex organism. The distance between mean and median is also increasing as the species becomes more complex. This denotes that the distribution of both *D_1* and *D_2* are skewed to the left and exhibits a very long right tail and is shown in Figure 3.1a for *H. sapiens*. Positive skewness was obtained from the data analysis (skewness = 6.605 for *H. sapiens*). Accordingly, the data indicates that in the simple organisms such as bacteria, there is a higher likelihood that they have short *D_1* and *D_2* distances than in the more complex species. It is important to note that even in different species within the eukaryotic and prokaryotic kingdoms there could be differences in the probability distributions for the distance components. To test for statistical significance between organisms, an Independent-Samples Kruskal-Wallis test was performed on the D1 and D2 data. Significance was $P< 0.000$ for both D1 and D2, at a level of 0.05, indicating that the distribution of D1 and D2 between organisms is different.

Considering the joint relationship of *D_1* and *D_2* for the six species, there is more information to be gained. The following two-dimensional scatter plots (Figure 3.1b) of *D_1* verses *D_2* for *E. coli* and *H. sapiens* shows that the correlation between these distances is varied from species to species. The bacterium species *D_1* value tended to be smaller and appears that compared with the D_2 value would not change to a great extent. However the plot for the *H. sapiens* illustrates different trends. The D_1 value declined in a different region on the plot and therefore made the distribution of D_2 look different. According to the data of the six species, the research found the more complex a species, the stronger the correlation between D_1 and D_2. The understanding of this relationship guided our research to explore this correlation further with more complex organisms and is outlined in chapter 5 of this thesis.

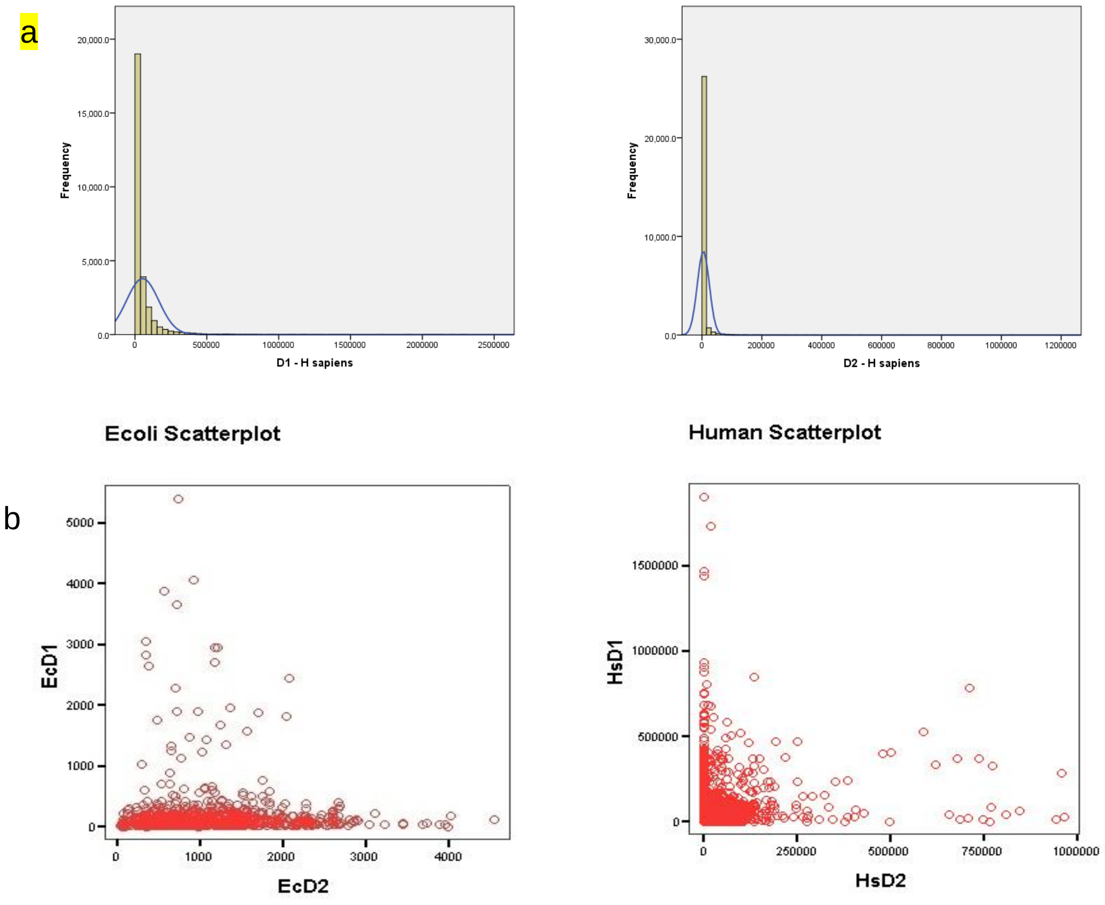


Figure 3.1 a) Frequency histograms of D1 and D2 for H. sapiens showing the positive skewness in the data b) Scatter plots of D_1 verses D_2 of E. coli and H. sapiens. Significance of correlation between the presented variables are statistically attested.

To explore the relationship between the TSS-TLS and TLS-TSC distances, and to ascertain whether there is a certain level of impact from the D_1 value on the probability distribution of D_2, the complete *H. sapiens* and *M. musculus* data sets were used. The higher organisms were chosen due to the higher correlation between these components found in the comparison above. The data was divided into four groups based on the quartiles of the D_1 values. The first group consisted of all D_1 values to the first quartile, the second of all D_1 values from the first quartile to the median, the third was made up of all *D_1* values from the median to the third quartile, and finally the last group consisted of all *D_1* values from the third quartile to the maximum *D_1* value.

To characterize the location and variability of a data set, the skewness and kurtosis can be used for statistical analysis purposes. Skewness measures the lack of symmetry in a distribution, where the kurtosis describes the data as either peaked or flat relative to a normal distribution.

Table 3.2 Statistics of TSS-TLS distances (D_2) given D_1 in different ranges

H. sapiens	Group 1	Group 2	Group 3	Group 4
Sample Size	2789	2789	2789	2788
Mean	8424.28	9984.15	13497.72	28400.62
Median	758	600	816	1194.5
Std. Deviation	40953.08	48848.37	45191.05	73803.25
Skewness	18.955	12.743	10.812	6.277
Std. Error of Skewness	0.0464	0.0464	0.0464	0.0464
Kurtosis	470.920	191.664	162.138	54.413
Std. Error of Kurtosis	0.093	0.093	0.0923	0.093
Minimum	1	1	1	1
Maximum	1261540	945008	967810	963680
Pearson Correlation	-0.040594	0.0231855	0.01539775	0.15475841
M. musculus	Group 1	Group 2	Group 3	Group 4
Sample Size	4717	2579	1652	5640
Mean	6238.84	6250.67	8012.67	13302.45
Median	422	267	385.5	399.5
Std. Deviation	36156.00	31896.66	40082.79	41833.31
Skewness	18.029	15.958	17.154	9.962
Std. Error of Skewness	0.036	0.048	0.060	0.036
Kurtosis	396.617	333.241	346.126	150.971
Std. Error of Kurtosis	0.071	0.096	0.120	0.071
Minimum	4	1	1	1
Maximum	973006	845821	906370	973292
Pearson Correlation	-0.03395	0.007623	-0.03082	0.239054

Table 3.2 clearly shows that, given *D_1* declining into a different region, the associated random component *D_2* had significantly different probability distribution. To test for statistical significance between each quartile group for D1, an Independent-Samples Kruskal-Wallis test was performed on the D2 data. Significance was *P*< 0.000 for D2, at a level of 0.05, indicating that the distribution of D2 across categories of quartile of D1 is different. Therefore, the larger the value of *D_1* the higher the correlation between *D_1*

given D_2. Since it is relatively easy to identify D_1 distances from the DNA sequence, with this information, the random component D_2 might show a different portion of information about DNA sequences.

Statistically, there is a great deal of potential to extract information from the D_1 and D_2 data however this research will not delve into every aspect. The purpose of the research is to highlight that different species might have different probability structure on their random components. Therefore, the information of D_1 and D_2 which is related to the distance of the TSS-TLS and could be referenced to the promoter position. Currently many computational promoter predictor algorithms do not take into account the information of D_1 and D_2. This information can be utilized to improve computational promoter prediction results and is discussed below.

3.4 The Algorithms for TLS-NNPP and TSC-TSS-NNPP

In this section, two algorithms using the information of D_1 and D_2 will be used to demonstrate that these random components can improve the NNPP performance. The first modification will incorporate the D_2 distance values and is called the TLS-NNPP algorithm (Burden, et al., 2005; Dai, et al., 2006). The other algorithm is known as the TSC-TSS-NNPP algorithm and uses both D_1 and D_2 values.

Reviews conducted on the NNPP algorithm illustrates that it is a competitive tool against several of the other programs available for promoter prediction. However, as with the other programs, this algorithm also suffers from a high instance of false positives. Currently used algorithms are not able to provide highly accurate predictions and the correct prediction promoter rate is only between 13-54%. It has been a research challenge to reduce the level of false positive recognition through modifying mathematical modeling and algorithms. Transcription is a complicated process which involves the interactions of promoter *cis*-elements with multiple *trans*-protein factors. The specific interactions rely not only on the specific sequence recognition between the *cis*- and *trans*-factors but also on some spatial arrangement of them in a complex. Hence, the distance between the TSS and TLS has and can be utilized in promoter prediction.

There are several reasons why the distances between the TSS and TLS (D_2) can be used to improve promoter prediction. For one, the promoter regions are closely associated to the location which in turn will assist in correctly predicting the position of the TSS and will lead to precisely estimating associated promoter regions. Secondly, numerous TSS and TLS experimental data is now accessible by researchers for different species, therefore the empirical probability distribution of TLS-TSS can be obtained. The information of the TLS position can be easily extrapolated from the gene coding region sequence, as it corresponds to the first nucleotide of the coding region. As a result, given a TLS position, and knowing the empirical distribution of the TLS-TSS, the distribution of the TSS can be determined from this distribution. Consequently, improving promoter prediction can be achieved by incorporating this information in the standard NNPP algorithm.

Given a whole DNA sequence of a species, *S* denotes the set of TSS positions in gene sequences. If position *s* is a true TSS position in a gene, it will be denoted by $s \in S$; if a range $[s-a_1, s+a_2] \subset S$ is used, it means the range $[s-a_1, s+a_2]$ covers at least one position which is a real TSS position of the gene. NNPP2.2 will give the probability $P([s-3, s+3] \subset S)$, sometimes, simply denoted by $P(s \in S)$. The NNPP algorithm has a high instance of false positives, which is due to the estimation of $P(s \in S)$. This probability is not accurate and sometimes overestimates the probability. Therefore, in this chapter, we will discuss how to use the information of D_1 and D_2 to adjust the probability given by NNPP.

Two scenarios will be discussed, in the first scenario, only the information of D_2 is considered. But in the second scenario, both information of D_1 and D_2 are take into account for promoter prediction. In the following examples, it always assumes that the position of the TLS and TSC can be easily identified from any given tested gene sequence.

3.4.1 Scenario 1 – TLS-NNPP Algorithm

In this scenario, it supposes that the NNPP2.2 software has been applied to a tested gene sequence and identified a position s in the sequence with probability of $P(s \in S)$. The NNPP2.2 algorithm is based on the nucleotide sequence and recognizes only the presence and relative location of patterns and motifs within a promoter, rather than the location of promoter motifs relative to the TLS. It predicts the probability that a tested sequence ± 3 bp (denoted by s) belongs to the class of true promoters (S). Given the position s and the tested gene sequence, the distance between s and TLS can be accurately identified. In this circumstance, the probability $P(s \in S, D_2(s) \in [d-a, d+a])$, that is, the probability that s is a TSS position and the distance s and its TLS is between *d-a* and *d+a*, is used to measure the likelihood of the s being a true TSS position. The higher the probability is the more likely s is a TSS position. In this paper we chose *a=3* which is the same value as NNPP employs.

The probability $P(s \in S, D_2(s) \in [d-a, d+a])$ can be evaluated by using the following formula:

$$P(s \in S, D_2(s) \in [d-a, d+a]) = P(s \in S)P(D_2(s) \in [d-a, d+a] \mid s \in S) \quad (1)$$

Formula (1) is used to adjust the value $P(s \in S)$ given by NNPP2.2. In the formula, $P(D_2(s) \in [d-a, d+a] \mid s \in S)$the information is ignored by NNPP2.2. To evaluate the probability the following steps are required:

(1) Collect the position information of the true TSS and its associated TLS for tested species. The larger the sample sizes the superior the output.

(2) Use statistical software to produce the empirical cumulative distribution function $F_{d_2}(d^*)$ for *D2*, $0 \le d^* < \infty$. Then use a nonparametric method to smooth the empirical cumulative distribution of *D2*. Both the above functions can be found from all common statistical software. The empirical cumulative distribution will give the estimation of$P(D_2(s) \le d^* \mid s \in S)$ for all $0 \le d^* < \infty$.

(3) Estimate $P(D_2(s) \in [d-a, d+a] \mid s \in S)$by $F_{d_2}(d+a) - F_{d_2}(d-a)$ and substitute it to Formula (1) to evaluate the probability $P(D_2(s) \in [d-a, d+a] \mid s \in S)$

The above formula is based on the sample information of *D2* to adjust the probability of *s* given by NNPP2.2. Sometimes we might consider an alternative way to adjust$P(s \in S)$. From research conducted by Dai *et al* (2006) it was found that all the density functions of *D2* are positively skewed. For Example, considering the histogram plots (Figure 3.2), of the *A. thaliana* and *H. sapiens*, the study found when the distance TSS-TLS is large beyond a certain point, the value of the probability density function drops sharply to a very small value.

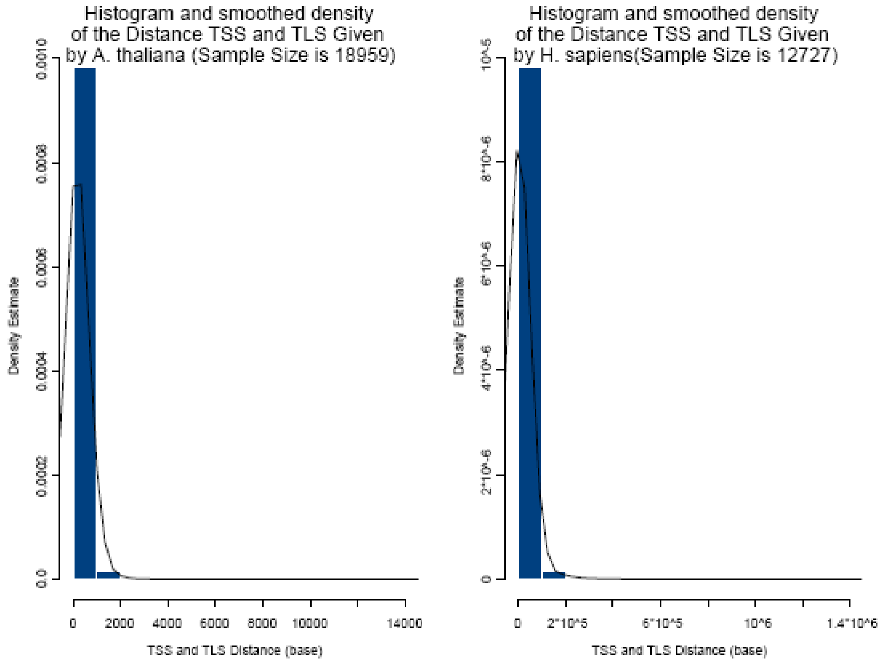


Figure 3.2 The histogram and smoothed density of distance TSS-TLS for A. thaliana *and* H. sapiens

This offers very little information for the position of the TSS when the distance is beyond that point. Therefore, in such situations, it might be worth considering the probability:

$$\begin{aligned} &P(s \in S, D_2(s) \in [d-a, d+a], D_2(s) \le M) = \\ &P(s \in S) P(D_2(s) \in [d-a, d+a], D_2(s) \le M) \mid s \in S) \end{aligned} \qquad (2)$$

instead of Formula (1), while $P(D_2(s) \in [d-a, d+a], D_2(s) \le M) \mid s \in S)$ will be evaluated by the empirical probability distribution determined by the entire sample with $D_2 \le M$. [Dai *et al*., 2006].

3.4.2 Scenario 2 – TSC-TSS-NNPP Algorithm

In this scenario, it is assumed that, the sample information on D_1 and D_2 for tested species is accessible. Under this assumption, given a gene sequence, if the true TSS position is at *s*; the distance between *s* and its TLS is D_2(s) and the distance between its TLS and TSC is D_1(s), the following probability will be worth evaluating:

$$P(s \in S, D_2(s) \in [d-a, d+a], D_1(s) \in [b_1, b_2])$$

where *a*, b_1, b_2 and *d* are positive integers, and *a* is equal to 3 showing that a tested position can differ by plus or minus 3 bp. The probability can be calculated in the following way

$$P(s \in S, D_2(s) \in [d-a, d+a], D_1(s) \in [b_1, b_2]) = P(s \in S) \\ \times P(D_1(s) \in [b_1, b_2] \mid s \in S) P(D_2(s) \in [d-a, d+a] \mid s \in S, D_1(s) \in [b_1, b_2]) \quad (3)$$

To evaluate the above probability, the estimation of $P(s \in S)$ is provided by NNPP2.2; following the similar steps listed in Scenario 1, the estimation of $P(D_1(s) \in [b_1, b_2] \mid s \in S)$ and $P(D_2(s) \in [d-a, d+a] \mid s \in S, D_1(s) \in [b_1, b_2])$ will be given by the empirical distribution of *D1* and the empirical distribution of *D2* given $D_1 \in [b_1, b_2]$ respectively.

However, if TSS positions are only predicted for gene sequences with $D_1 \in [b_1, b_2]$, the above evaluation can be simplified, and evaluate:

$$P(s \in S) P(D_2(s) \in (d-a, d+a) \mid s \in S, D_1 \in (b_1, b_2)) \quad (4)$$

instead of Formula (3). In the next section, we only apply Formula (4) to real data.

Figure 3.3 shows the schematic representation of the algorithms and procedure outlined in this chapter.

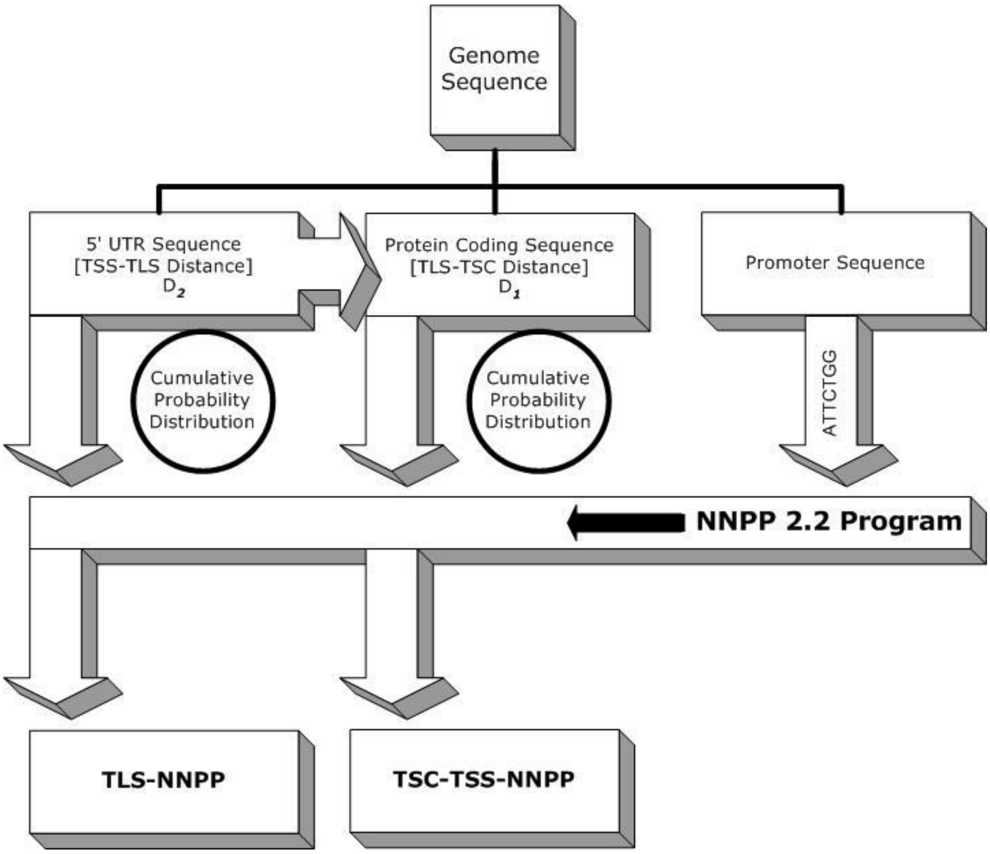

Genome Sequence
5' UTR Sequence [TSS-TLS Distance] D₂
Protein Coding Sequence [TLS-TSC Distance] D₁
Promoter Sequence
Cumulative Probability Distribution
Cumulative Probability Distribution
ATTCTGG
NNPP 2.2 Program
TLS-NNPP
TSC-TSS-NNPP

Figure 3.3 Schematic Representation of Promoter Prediction using TLS-NNPP and TSC-TSS-NNPP Algorithms

3.5 Applications of the Algorithms TLS-NNPP and TSC-TSS-NNPP and the comparisons to NNPP2.2

In this section, two applications are presented and the results of TLS-NNPP and TSC-TSS-NNPP are compared to the relevant results of NNPP2.2. Using the TSC-TSS-NNPP and TSS-NNPP methods to analyze the data, the adjusted score had to be utilized. The NNPP2.2 algorithm generates scores or cutoff values at tenths such as 0.1, 0.2, 0.3, 0.4, 0.5, up to 0.9. To obtain similar values, tenths of the maximum adjusted score were taken to obtain cutoff values for the TSC-TSS-NNPP and TSS-NNPP methods.

We compare the algorithms TLS-NNPP and TSC-TSS-NNPP to NNPP2.2 in term of the probability of correct prediction. For example, the probability of a position which is accepted as TSS position by an algorithm is really a position of TSS.

To save time, the comparison in this paper was done based on a 10% of the gene sample data. This 10% sub-sample is called a testing sample, and is randomly selected from the sample data to reduce the impact of sample error on comparison results. The methods TLS-NNPP, TSC-TSS-NNPP and NNPP2.2 are applied to the sub- sample respectively. Then, for each cut-off value, the total number of predictions and positive predictions in a range greater than each cut-off value were counted and the probability of correct prediction, denoted by *P(Correct Prediction)* will be evaluated for the TSS-NNPP, TSC-TSS-NNPP and NNPP2.2 respectively. The estimations of *P(Correct Prediction)* are the number of positive predictions divided by the total number of predictions.

3.5.1 *E. Coli* Sequence Study Using the TLS-NNPP Algorithm

We firstly used this technique and modification to the NNPP2.2 algorithm on *Escherichia coli* DNA sequences. The process involved in the implementation took several steps. The steps involved creating an empirical distribution for the TSS-TLS distance, next, DNA sequences (500 bp) were run through the NNPP2.2 program and only the true positively predicted TSS positions were used. The Promoters were considered to be correctly predicted when the actual TSS of the promoter fell within ±3 bp of a predicted TSS. The predicted promoters must be in-line with the closest subsequent TLS in the sequences and the TSS-TLS distance.

The research conducted by Burden et al showed that by modifying the NNPP2.2 algorithm program by incorporating addition information, such as the TSS-TLS distance, it greatly improved the prediction of promoters and reduced the incidence of false predictions. Figure 3.4 shows how effective the TLS-NNPP technique was compared to the NNPP2.2 program without the modifications. The number of predictions for this particular species was low due to the training set only containing 293 *E. coli* promoters therefore would not recognize any of the new promoters in the sequences (Burden, et al., 2005).

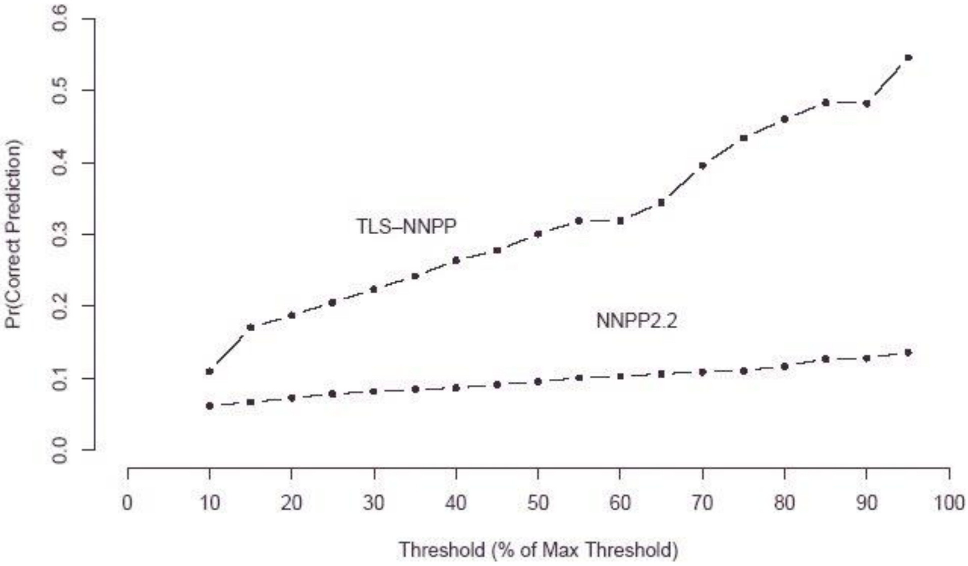


Figure 3.4 Comparison of probability of prediction of promoter sequences at different thresholds for NNPP2.2 and TLS-NNPP
(Burden, et al., 2005)

Further study on a range of species crossing from less complex to more complex organisms also showed that the TLS-NNPP method has power to improve the outcomes of NNPP2.2.

3.5.2 Human Sequence Study Using the TSS-TSC-NNPP Algorithm

As described in the previous section, it is possible to use the TSC-TSS-NNPP approach to improve the performance of NNPP2.2. This is only possible if the data is accessible from databases that could offer large numbers of experimentally defined promoter sequences and start and stop positions for the coding regions and 5' and 3' un-translated regions.

In this section the TSC-TSS-NNPP method is applied to human data. Table 2 in Section 3 shows that, for human data, D_1 dropped into different regions, and might lead to the variation in the probability distribution of D_2. Since the information of D_2 is related to TSS position, it means that the information of the value of D_1 might have certain level of impact on promoter prediction. We adopt the four groups, described in Table 3.2, to group the value of D_1. That is, Group 1 for $D_1 \leq 5583$; Group 1 for $5583<D_1\leq 17466$; Group 3 for $17466 < D_1 \leq 43976$ and group 4 for $43976 < D_1$. The comparisons between the algorithms TSC-TSS-NNPP and NNPP2.2 were done for D_1 in the four groups respectively.

Our results show that in all four groups of the *H. sapiens* data set, the TSC-TSS-NNPP method achieved better results than both NNPP2.2 and TSS-NNPP, particularly for Group 1. Looking at Figure 3.5, 60% seems to be the best cut-off value for the TSC-TSS-NNPP method which has a greater Pr(Correct Prediction) value than the other two methods at this cut-off value. Additionally, within a 10%-60% threshold range for Group 1, this showed that the probability of predicting that a sequence is a promoter is highest for TSC-TSS-NNPP.

As shown in Figure 3.5, the *P(Correct Prediction)* values for TSC-TSS-NNPP and TSS-NNPP dropped down at large threshold values. This is because time constraints did not allow us to examine a large data set for this research and dividing the data into groups extensively reduced its size so much so that there was no data available and information was exhausted at large threshold levels. Therefore, the data should generally be compared within a range of 0 to around 60%.

The TSC-TSS-NNPP method produced better results compared to the NNPP and TSS-NNPP methods for the *H. sapiens* data set for all four groups. We also apply TSC-TSS-NNPP to *M. musculus* data (The results are omitted from this thesis). For the *M. musculus* data however, our study show that the TSC-TSS-NNPP method is the better choice only for Groups 1 and 2, whereas the NNPP2.2 method is better for Groups 3 and 4. It is interesting to note that the TSC-TSS-NNPP method produced extensively better results than TSS-NNPP and NNPP2.2 for small D_1 values (Group 1) in both species. This is a vital merit for the TSC-TSS-NNPP approach, as generally, shown by 3D histograms of all organisms there is a very high proportion of small D_1 values in the complete data set (3D histograms is omitted from this thesis).

Therefore, if the data set consists largely of small D_1 values, this new method will be highly effective in reducing the false positive rate for the NNPP2.2 tool, which will then ensure that each promoter that is predicted is associated with a gene coding region.

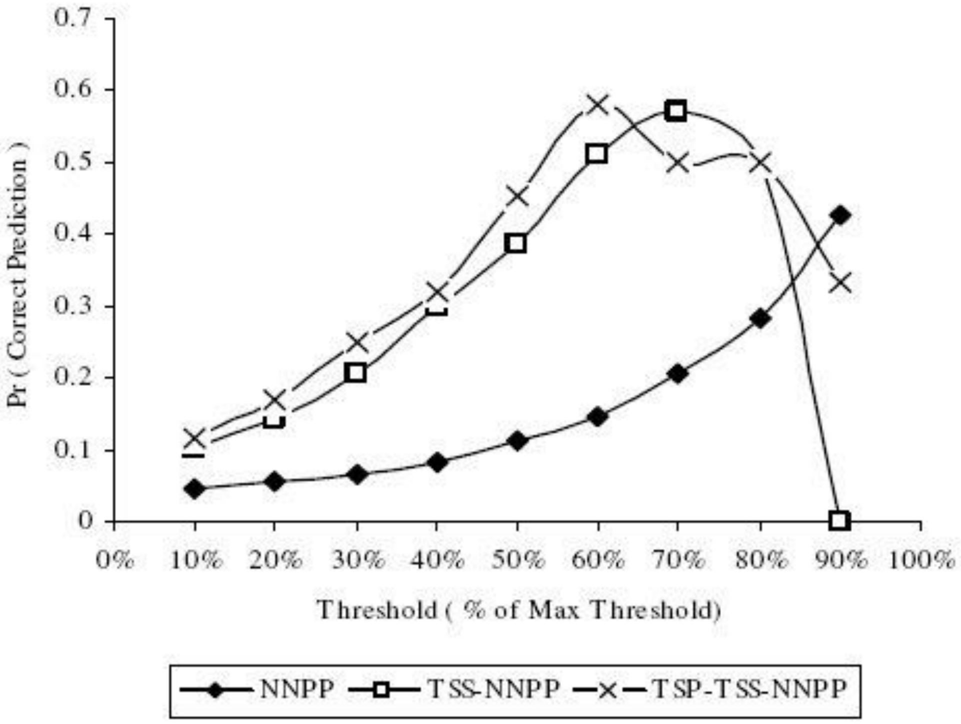


Figure 3.5 The comparison of three methods with D_1 in Group 1

The research in this chapter was beneficial in helping with the understanding of the coding and noncoding sequence length and how this thesis should progress in expanding and creating a better understanding of the length distributions of these regions. To understand the complexity of length it was logical to start with data that was readily available for a variety of organisms, to appreciate the assumptions, limitations and behaviour of the data. The next chapter investigates the coding sequence length among a cross section of organisms.

Chapter 4 – Coding Sequence Length Comparisons

4 Coding Sequence Length Comparisons

4.1 Introduction

Numerous species have now been fully sequenced, including protein coding sequences due to the impressive progress of high-throughput DNA sequencing techniques (Nowrousian, 2010), allowing biologists and statisticians to study and compare various species of prokaryotes and eukaryotes. Up to 2006, when this project started, previous studies on protein lengths had focused on either prokaryotes or eukaryotes, with some research investigating the differences between these organisms, as well as their protein lengths (Wang, 2005; Xu, et al., 2006; Zhang, 2000). Examination of the protein coding sequences had in the past, been limited, particularly on comparing a wide range of eukaryotic organisms in addition to comparisons on their chromosomes and protein numbers.

Chromosomal differences including rearrangements, such as inversions, translocations, and duplication and genetic variation among species have provided fundamental evidence for Darwin’s theory of natural selection (Coghlan, et al., 2005). The study of the chromosomes of *Drosophila melanogaster* and *Drosophila simulans* differ in chromosome III by large inversions, as well as other species of flies. This has initiated many questions of chromosomal structure, including what regions or sites chromosomes are predisposed to change, and how large the DNA segments are inverted, deleted, translocated or duplicated (Eichler and Sankoff, 2003).

Rearrangements in chromosomes can be detected either via a microscope if large, such as deletions, inversions and duplications, or if the rearrangements are fine-scale can be studied through genome sequencing (Coghlan, et al., 2005). Research currently being considered for genome sequencing from The National Human Genome Research Institute and the US Department of Energy comprise ~20 fungal species, ~40 invertebrates and ~25 vertebrates. Since the progression of sequenced data for eukaryotic genomes, information on the smallest of changes for example, single base pair substitutions has become the motivation to further investigate fine-scale changes in chromosomal structures both within and between species (Coghlan, et al., 2005). Information gained through these organisms may have inference about structural and functional genomics (Eichler and Sankoff, 2003).

As a foundation and starting point for this research, the coding sequence, protein number and length of fifteen eukaryotic organisms were examined to understand the complexity of these organisms in relation to their length distributions of the coding sequence, including the investigation of these lengths with individual chromosomes. The coding sequence data was split into chromosome level and data including and excluding introns was explored. Protein information such as protein density per megabase (Mb) for every chromosome was also investigated. Conclusions will be made in regards to the biological processes that may be seen within each organism, and may offer greater insight into the complexity of these organisms.

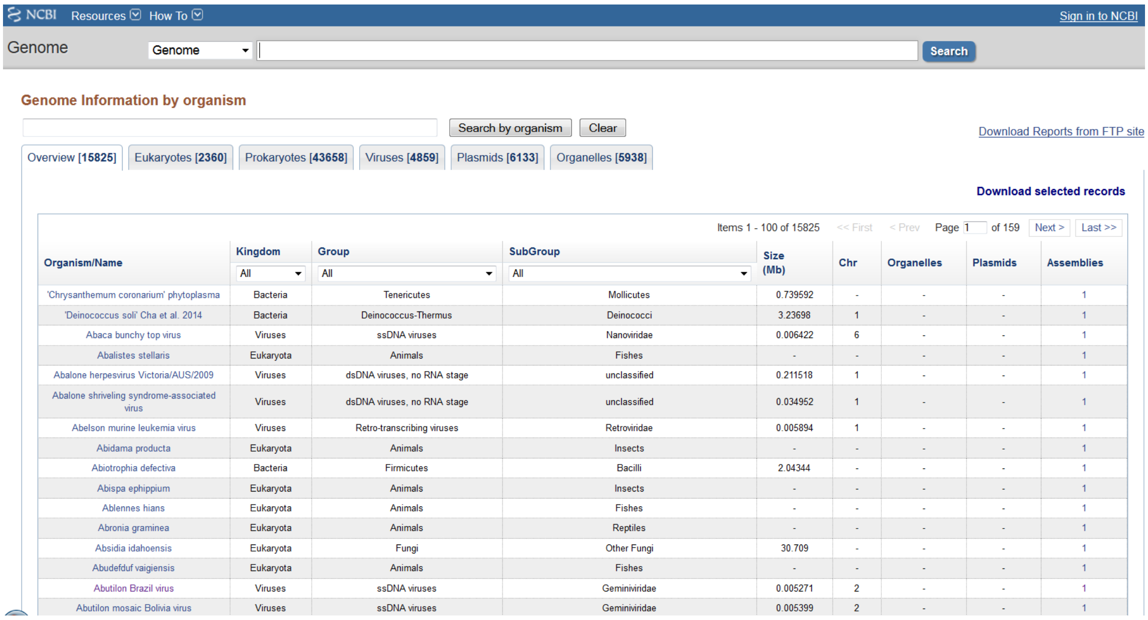
4.2 Data File Construction & Statistical Data Analysis

DNA sequencing data was downloaded from the NCBI Genome web site (http://www.ncbi.nlm.nih.gov/Genomes/) in January 2007 (Figure 4.1). Fifteen complete or assembled sequenced eukaryotic genomes were chosen as part of this study (Table 4.1). All the eukaryotic organisms were selected that contained both start and stop codons and protein lengths. All organism protein tables were downloaded from the Reference Sequence (RefSeq) collection (http://www.ncbi.nlm.nih.gov/RefSeq/) which provides a set of sequences for major research organisms and includes genomic DNA, transcript (RNA) and protein product information (Pruitt, et al., 2014; Pruitt, et al., 2007). This information has been used in a wide range of research, including functional, expression and diversity studies as well as comparative analyses (Fong, et al., 2013; Yi, et al., 2014).

The organisms that were selected included one protist, *Plasmodium falciparum*, one plant species, *Arabidopsis thaliana*, five species of fungi, *Saccharomyces cerevisiae*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi* and eight species of animals, *Anopheles gambiae*, *Tribolium castaneum*, *Caenorhabditis elegans*, *Drosophila Melanogaster*, *Danio rerio*, *Mus musculus*, *Pan troglodytes*, *Homo sapiens*.

Table 4.1 Species of eukaryotes sequences downloaded in 2007 from the RefSeq collection NCBI website for study

Species	Size (Mb)	Number of Chromosomes
Plasmodium falciparum (Gardner, et al., 2002)	27.0235	14
Arabidopsis thaliana	119.668	5
Encephalitozoon cuniculi (Katinka, et al., 2001)	2.49752	11
Debaryomyces hansenii (Dujon, et al., 2004)	12.1819	7
Saccharomyces cerevisiae	14.2673	16
Candida glabrata (Dujon, et al., 2004)	12.338	13
Cryptococcus neoformans (Loftus, et al., 2005)	19.6998	14
Caenorhabditis elegans (Consortium, 1998)	100.286	6
Drosophila melanogaster (Adams, et al., 2000)	164.05	4
Anopheles gambiae	265.027	3
Tribolium castaneum	210.865	10
Danio rerio	1411.76	25
Mus musculus (Consortium, 2002)	2798.79	21
Pan troglodytes	3309.56	24
Homo sapiens (Human Genome Sequencing, 2004)	3256.04	24



Genome: Genome

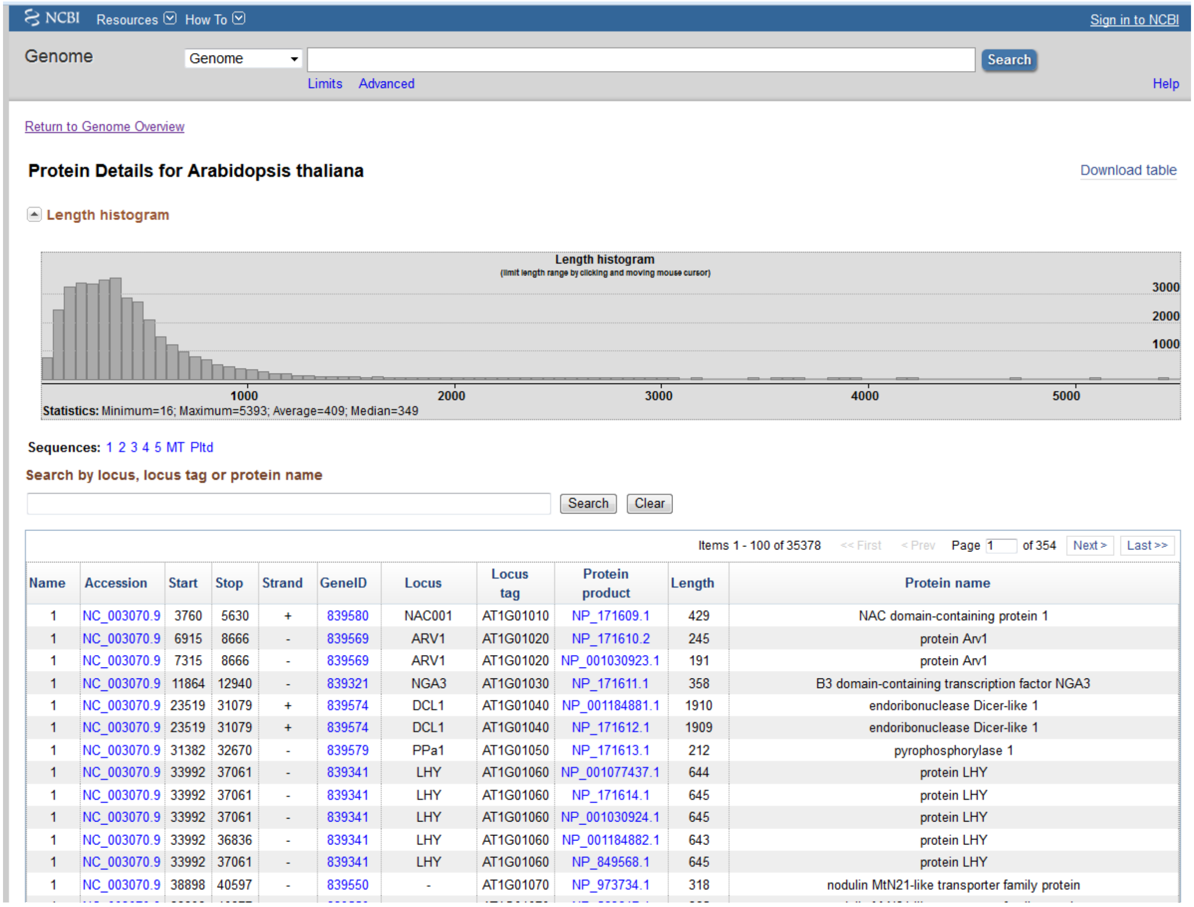
Genome Information by organism

Items 1 - 100 of 15825

Page 1 of 159

Organism/Name	Kingdom	Group	SubGroup	Size (Mb)	Chr	Organelles	Plasmids	Assemblies
	All	All	All					
'Chrysanthemum coronarium' phytoplasma	Bacteria	Tenericutes	Mollicutes	0.739592	-	-	-	1
'Deinococcus soli' Cha et al. 2014	Bacteria	Deinococcus-Thermus	Deinococci	3.23698	1	-	-	1
Abaca bunchy top virus	Viruses	ssDNA viruses	Nanoviridae	0.006422	6	-	-	1
Abalistes stellaris	Eukaryota	Animals	Fishes	-	-	-	-	1
Abalone herpesvirus Victoria/AUS/2009	Viruses	dsDNA viruses, no RNA stage	unclassified	0.211518	1	-	-	1
Abalone shriveling syndrome-associated virus	Viruses	dsDNA viruses, no RNA stage	unclassified	0.034952	1	-	-	1
Abelson murine leukemia virus	Viruses	Retro-transcribing viruses	Retroviridae	0.005894	1	-	-	1
Abidama producta	Eukaryota	Animals	Insects	-	-	-	-	1
Abiotrophia defectiva	Bacteria	Firmicutes	Bacilli	2.04344	-	-	-	1
Abispa ephippium	Eukaryota	Animals	Insects	-	-	-	-	1
Ablennes hians	Eukaryota	Animals	Fishes	-	-	-	-	1
Abronia graminea	Eukaryota	Animals	Reptiles	-	-	-	-	1
Absidia idahoensis	Eukaryota	Fungi	Other Fungi	30.709	-	-	-	1
Abudefduf vaigiensis	Eukaryota	Animals	Fishes	-	-	-	-	1
Abutilon Brazil virus	Viruses	ssDNA viruses	Geminiviridae	0.005271	2	-	-	1
Abutilon mosaic Bolivia virus	Viruses	ssDNA viruses	Geminiviridae	0.005399	2	-	-	1

(a)



Protein Details for Arabidopsis thaliana

Statistics: Minimum=16; Maximum=5393; Average=409; Median=349

Sequences: 1 2 3 4 5 MT Pltd

Search by locus, locus tag or protein name

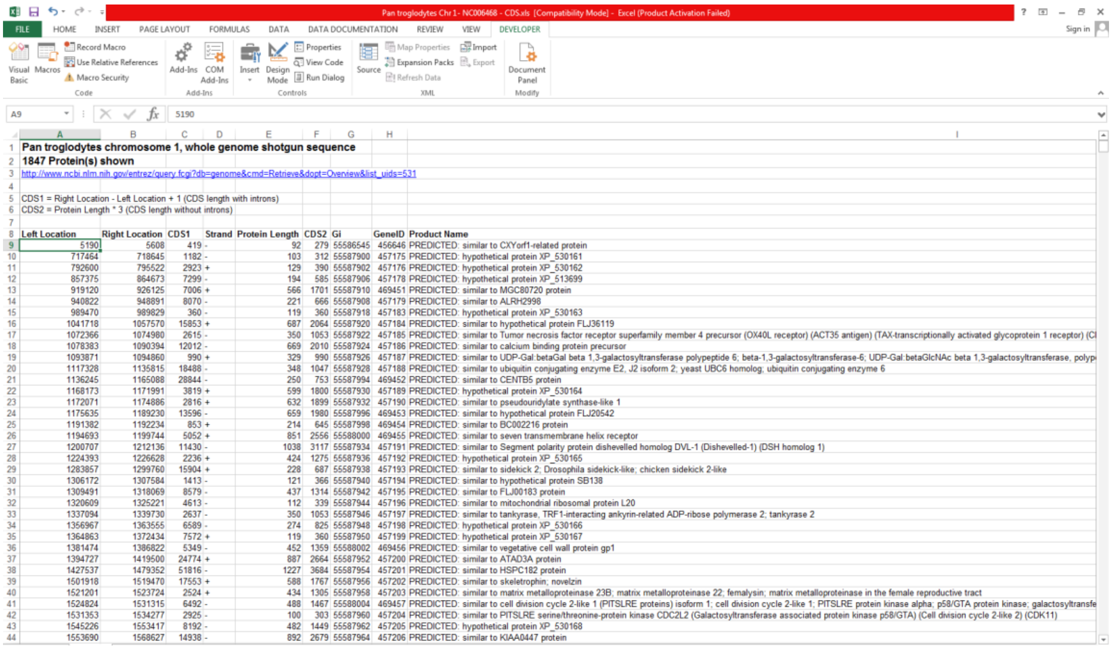
Items 1 - 100 of 35378 Page 1 of 354

Name	Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	Protein name
1	NC_003070.9	3760	5630	+	839580	NAC001	AT1G01010	NP_171609.1	429	NAC domain-containing protein 1
1	NC_003070.9	6915	8666	-	839569	ARV1	AT1G01020	NP_171610.2	245	protein Arv1
1	NC_003070.9	7315	8666	-	839569	ARV1	AT1G01020	NP_001030923.1	191	protein Arv1
1	NC_003070.9	11864	12940	-	839321	NGA3	AT1G01030	NP_171611.1	358	B3 domain-containing transcription factor NGA3
1	NC_003070.9	23519	31079	+	839574	DCL1	AT1G01040	NP_001184881.1	1910	endoribonuclease Dicer-like 1
1	NC_003070.9	23519	31079	+	839574	DCL1	AT1G01040	NP_171612.1	1909	endoribonuclease Dicer-like 1
1	NC_003070.9	31382	32670	-	839579	PPa1	AT1G01050	NP_171613.1	212	pyrophosphorylase 1
1	NC_003070.9	33992	37061	-	839341	LHY	AT1G01060	NP_001077437.1	644	protein LHY
1	NC_003070.9	33992	37061	-	839341	LHY	AT1G01060	NP_171614.1	645	protein LHY
1	NC_003070.9	33992	37061	-	839341	LHY	AT1G01060	NP_001030924.1	645	protein LHY
1	NC_003070.9	33992	36836	-	839341	LHY	AT1G01060	NP_001184882.1	643	protein LHY
1	NC_003070.9	33992	37061	-	839341	LHY	AT1G01060	NP_849568.1	645	protein LHY
1	NC_003070.9	38898	40597	-	839550	-	AT1G01070	NP_973734.1	318	nodulin MtN21-like transporter family protein

(b)

Figure 4.1 NCBI RefSeq website Protein table screen.
Protein tables were downloaded (exported) from the NCBI website and imported into Excel. (a) NCBI browse website (http://www.ncbi.nlm.nih.gov/genome/browse/) to select specific species; (b) protein details, including a length histogram and protein table that can be downloaded to excel.

Protein tables were exported and added into Microsoft Excel files for each organism and arranged into individual chromosome. Two columns were added to each worksheet, CDS 1 which was calculated by subtracting the stop position value from the start position value and adding one (1) for each protein record. This column was then used for information regarding the coding sequence of each protein that contained introns. The second column added was labelled CDS 2 which was calculated by multiplying the protein length for each protein coding gene by 3 and adding three (3). This column was used for information pertaining to the coding sequence of each protein that did not contain introns (Figure 4.2). Each MS Excel table was then imported into one MS Access database for further query construction and statistical analysis. Protein density was calculated by dividing the number of proteins per chromosome by the length of each chromosome.



A9 5190

Pan troglodytes chromosome 1, whole genome shotgun sequence

1847 Protein(s) shown

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome&cmd=Retrieve&dopt=Overview&list_uids=531

CDS1 = Right Location - Left Location + 1 (CDS length with introns)

CDS2 = Protein Length * 3 (CDS length without introns)

Left Location	Right Location	CDS1	Strand	Protein Length	CDS2	Gi	GeneID	Product Name
5190	5608	419	-	92	279	55586545	456646	PREDICTED: similar to CXYorf1-related protein
717464	718645	1182	-	103	312	55587900	457175	PREDICTED: hypothetical protein XP_530161
792600	795522	2923	+	129	390	55587902	457176	PREDICTED: hypothetical protein XP_530162
857375	864673	7299	-	194	585	55587906	457178	PREDICTED: hypothetical protein XP_513699
919120	926125	7006	+	566	1701	55587910	469451	PREDICTED: similar to MGC80720 protein
940822	948891	8070	-	221	666	55587908	457179	PREDICTED: similar to ALRH2998
989470	989829	360	-	119	360	55587918	457183	PREDICTED: hypothetical protein XP_530163
1041718	1057570	15853	+	687	2064	55587920	457184	PREDICTED: similar to hypothetical protein FLJ36119
1072366	1074980	2615	-	350	1053	55587922	457185	PREDICTED: similar to Tumor necrosis factor receptor superfamily member 4 precursor (OX40L receptor) (ACT35 antigen) (TAX-transcriptionally activated glycoprotein 1 receptor) (C
1078383	1090394	12012	-	669	2010	55587924	457186	PREDICTED: similar to calcium binding protein precursor
1093871	1094860	990	+	329	990	55587926	457187	PREDICTED: similar to UDP-Gal:betaGal beta 1,3-galactosyltransferase polypeptide 6; beta-1,3-galactosyltransferase-6; UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polyp
1117328	1135815	18488	-	348	1047	55587928	457188	PREDICTED: similar to ubiquitin conjugating enzyme E2, J2 isoform 2; yeast UBC6 homolog; ubiquitin conjugating enzyme 6
1136245	1165088	28844	-	250	753	55587994	469452	PREDICTED: similar to CENTB5 protein
1168173	1171991	3819	+	599	1800	55587930	457189	PREDICTED: hypothetical protein XP_530164
1172071	1174886	2816	+	632	1899	55587932	457190	PREDICTED: similar to pseudouridylate synthase-like 1
1175635	1189230	13596	-	659	1980	55587996	469453	PREDICTED: similar to hypothetical protein FLJ20542
1191382	1192234	853	+	214	645	55587998	469454	PREDICTED: similar to BC002216 protein
1194693	1199744	5052	+	851	2556	55588000	469455	PREDICTED: similar to seven transmembrane helix receptor
1200707	1212136	11430	-	1038	3117	55587934	457191	PREDICTED: similar to Segment polarity protein dishevelled homolog DVL-1 (Dishevelled-1) (DSH homolog 1)
1224393	1226628	2236	+	424	1275	55587936	457192	PREDICTED: hypothetical protein XP_530165
1283857	1299760	15904	+	228	687	55587938	457193	PREDICTED: similar to sidekick 2; Drosophila sidekick-like; chicken sidekick 2-like
1306172	1307584	1413	-	121	366	55587940	457194	PREDICTED: similar to hypothetical protein SB138
1309491	1318069	8579	-	437	1314	55587942	457195	PREDICTED: similar to FLJ00183 protein
1320609	1325221	4613	-	112	339	55587944	457196	PREDICTED: similar to mitochondrial ribosomal protein L20
1337094	1339730	2637	-	350	1053	55587946	457197	PREDICTED: similar to tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2; tankyrase 2
1356967	1363555	6589	-	274	825	55587948	457198	PREDICTED: hypothetical protein XP_530166
1364863	1372434	7572	+	119	360	55587950	457199	PREDICTED: hypothetical protein XP_530167
1381474	1386822	5349	-	452	1359	55588002	469456	PREDICTED: similar to vegetative cell wall protein gp1
1394727	1419500	24774	+	887	2664	55587952	457200	PREDICTED: similar to ATAD3A protein
1427537	1479352	51816	-	1227	3684	55587954	457201	PREDICTED: similar to HSPC182 protein
1501918	1519470	17553	+	588	1767	55587956	457202	PREDICTED: similar to skeletrophin; novelzin
1521201	1523724	2524	+	434	1305	55587958	457203	PREDICTED: similar to matrix metalloproteinase 23B; matrix metalloproteinase 22; femalysin; matrix metalloproteinase in the female reproductive tract
1524824	1531315	6492	-	488	1467	55588004	469457	PREDICTED: similar to cell division cycle 2-like 1 (PITSLRE proteins) isoform 1; cell division cycle 2-like 1; PITSLRE protein kinase alpha; p58/GTA protein kinase; galactosyltransfe
1531353	1534277	2925	-	100	303	55587960	457204	PREDICTED: similar to PITSLRE serine/threonine-protein kinase CDC2L2 (Galactosyltransferase associated protein kinase p58/GTA) (Cell division cycle 2-like 2) (CDK11)
1545226	1553417	8192	-	482	1449	55587962	457205	PREDICTED: hypothetical protein XP_530168
1553690	1568627	14938	-	892	2679	55587964	457206	PREDICTED: similar to KIAA0447 protein

Figure 4.2 Excel file containing length data obtained from NCBI RefSeq tables

Standard statistical data analysis tools are used in this chapter. Statistical data analysis packages included JMP, SAS and SPSS, to run ANOVA and student *t-test*.

4.3 Empirical & Comparative Study

It was the intention of this part of the study to investigate a wide range of eukaryotes for an initial comparative study. 15 species of eukaryotes were selected, with comparisons on each chromosome. A total of 248,019 protein coding sequences were studied with a focus on three parameters: CDS (without introns), CDS + introns and protein length.

The overall mean value for each species over all the chromosomes shows some notable observations (Figure 4.3). For *E. cuniculi* which has 11 chromosomes in total, the average protein number per Mb for this species is 800. Interestingly, as the organisms become more complex (based on the tree of life), the number of proteins per Mb drops. *H. sapiens*, *M. musculus* and *P. troglodytes* have an average protein number of only 6-9 per Mb. It is worth considering that these organisms have almost twice as many chromosomes than that of *E. cuniculi*. Most of the fungi species have a high number of proteins per Mb than the other organisms (Figure 4.3).

When all the species of eukaryotes are grouped together in their respected categories, specific trends emerge from the data. Observations from the data for the fungi species show to have a large amount of proteins per Mb compared with the animal kingdom, which include the mammals. If placed in order of largest number of proteins per Mb within the total of all chromosomes combined, the fungi group would come first, followed by the plants then closely followed by the protist and finally the animals (Figure 4.3). *T-Test* analysis was performed to test the differences in the means between organisism, and the test was significant (*t* 4.194, *P* = 0.001)

4.3.1 Protein number and density

The plant species, *Arabidopsis thaliana*, contained the highest number of proteins for each chromosome, averaging around 5,800 proteins. By comparison, the fungi species, *Encephalitozoon cuniculi*, had the lowest number of proteins within each of its chromosomes, averaging around 180 proteins (Figure 4.4). *Homo sapiens*, and *Pan troglodytes* species showed large variations in chromosome 1 compared to the Y chromosome in relation to number of proteins (Figure 4.4). For example in the *Homo sapiens* the number of proteins in chromosome 1 is 2,718 compared to Y which has only 104 proteins.

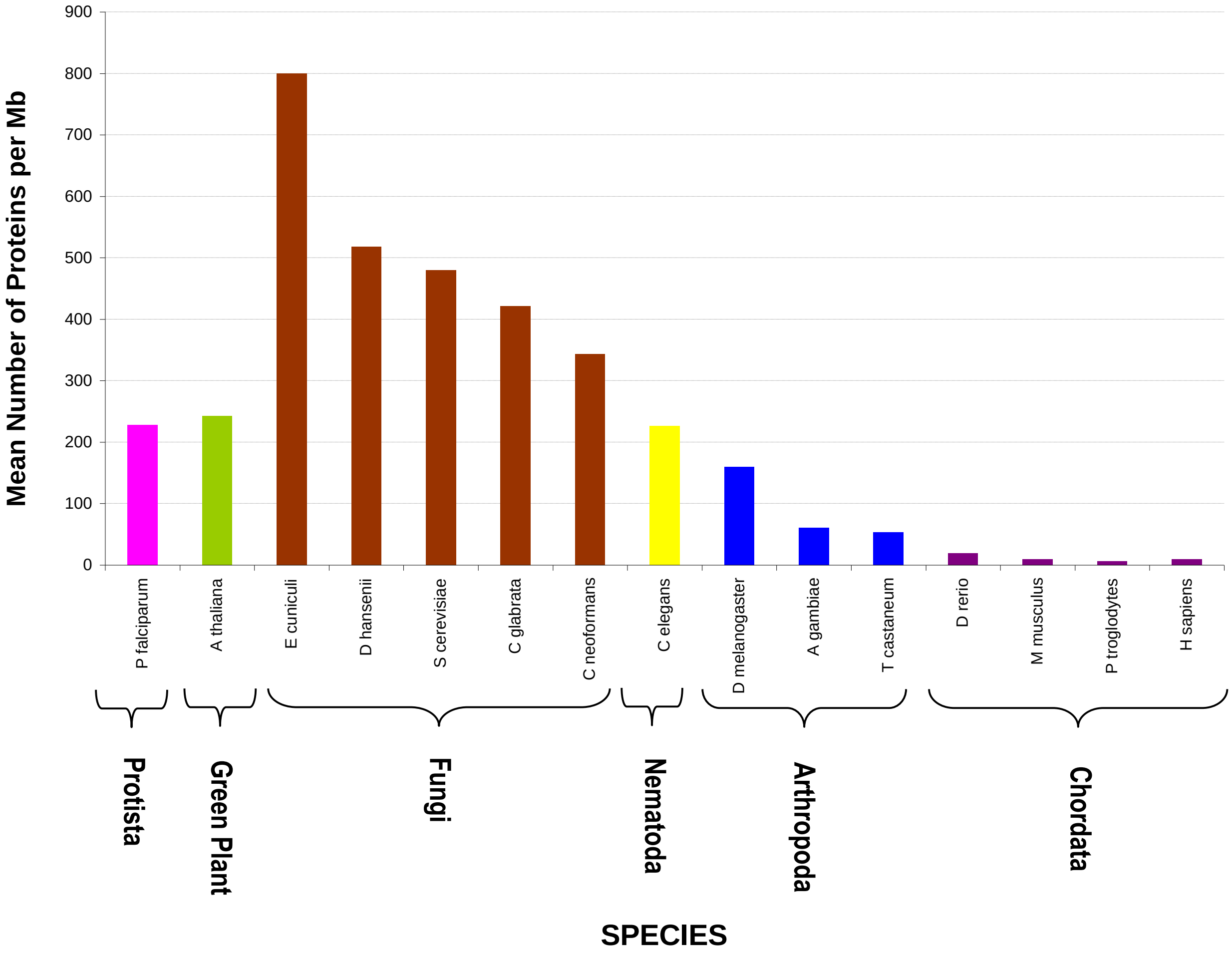
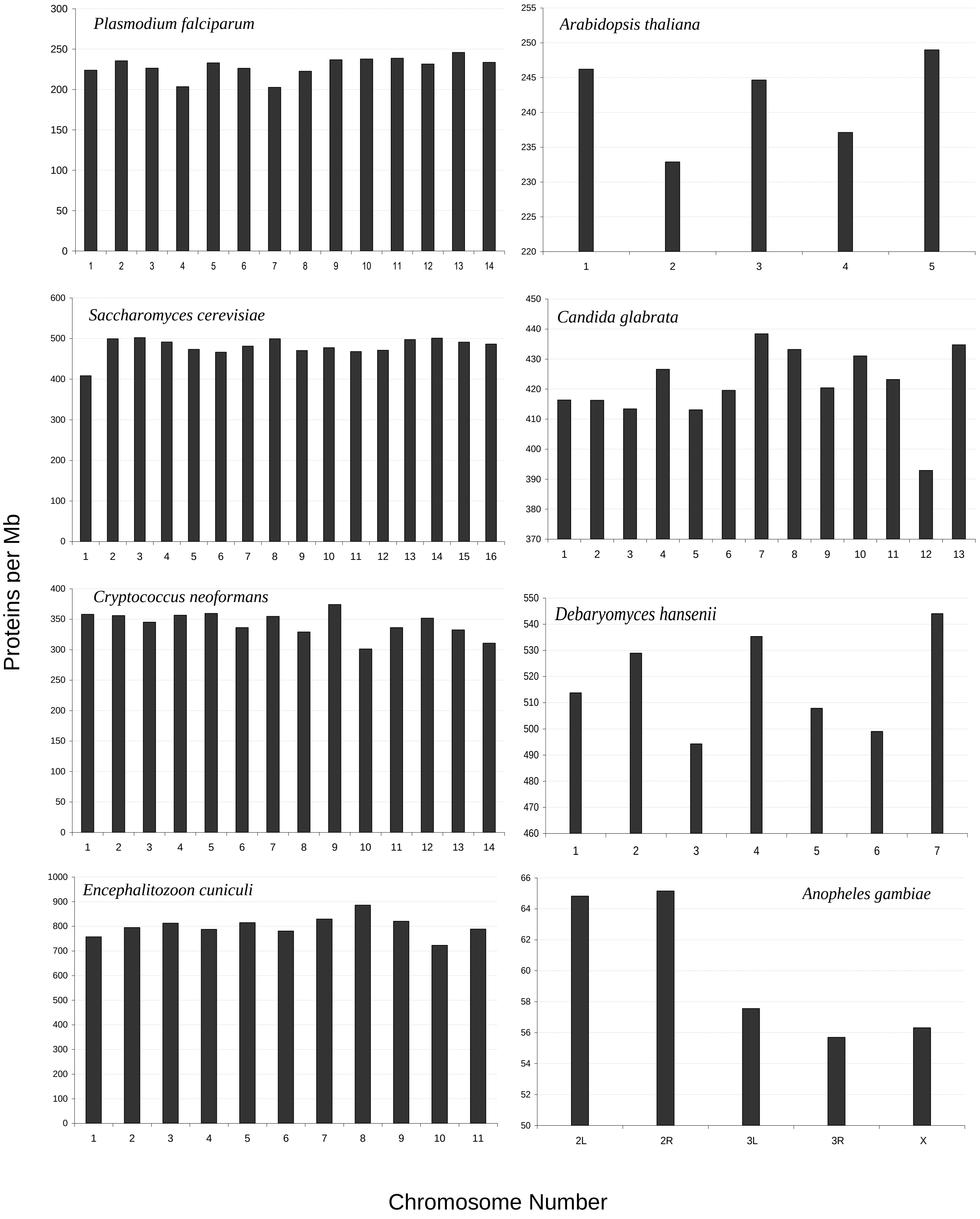


Figure 4.3 - Mean number of proteins per Mb over all chromosomes. 15 species of eukaryotes were studied.

To calculate the density of proteins for each chromosome the number of proteins per chromosome was divided by the length of the chromosome (Mbp). Mean number was calculated from all chromosomes for each species. Data was obtained from the RefSeq proteins tables found at http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi web site.

Observations from the data suggest *A. thaliana* showed higher protein density per Mb in chromosomes, one and five than the other chromosomes (Figure 4.4). Species *C. glabrata* and *D. hansenii* showed large differences within each chromosome. *C. glabrata* showed the lowest amount in chromosome 12 being 390 proteins per Mb, whereas *D. hansenii* had the highest protein number per Mb in chromosome 7 (540 proteins per Mb) and the lowest in chromosome 3 (490 proteins per Mb). *S. cerevisiae* showed a fairly consistent range of densities apart from chromosome 1 which was slightly lower (Figure 4.4). Interestingly, for the higher organisms such as *M. musculus* the range of densities was quite dramatic, with the highest value appearing in chromosome 11. Both *H. sapiens* and *P. troglodytes* had the highest value at chromosome 19, and the lowest at the Y chromosome (Figure 4.4).


Proteins per Mb
Chromosome Number
Plasmodium falciparum
0 50 100 150 200 250 300
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Arabidopsis thaliana
220 225 230 235 240 245 250 255
1 2 3 4 5
Saccharomyces cerevisiae
0 100 200 300 400 500 600
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Candida glabrata
370 380 390 400 410 420 430 440 450
1 2 3 4 5 6 7 8 9 10 11 12 13
Cryptococcus neoformans
0 50 100 150 200 250 300 350 400
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Debaryomyces hansenii
460 470 480 490 500 510 520 530 540 550
1 2 3 4 5 6 7
Encephalitozoon cuniculi
0 100 200 300 400 500 600 700 800 900 1000
1 2 3 4 5 6 7 8 9 10 11
Anopheles gambiae
50 52 54 56 58 60 62 64 66
2L 2R 3L 3R X

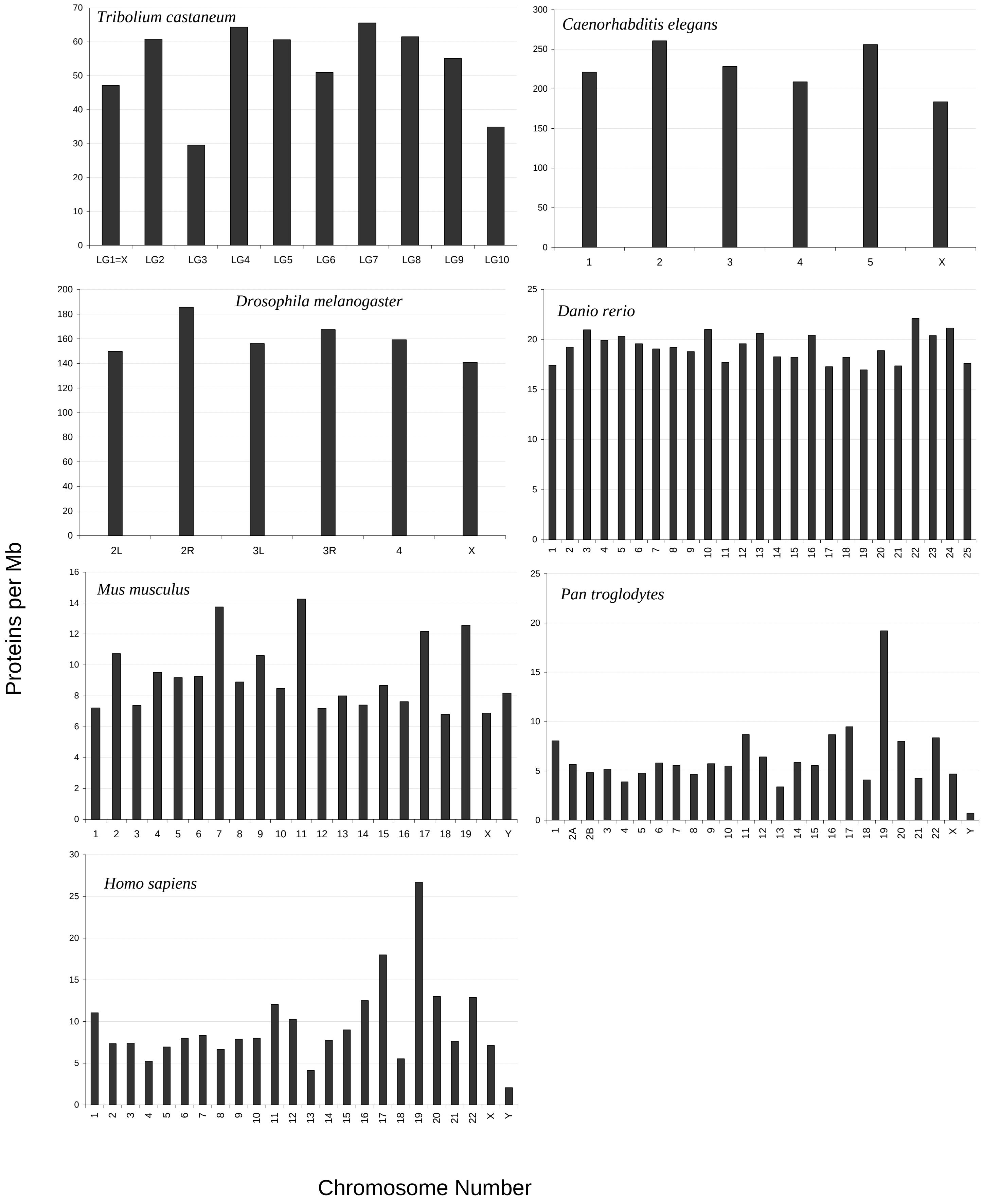
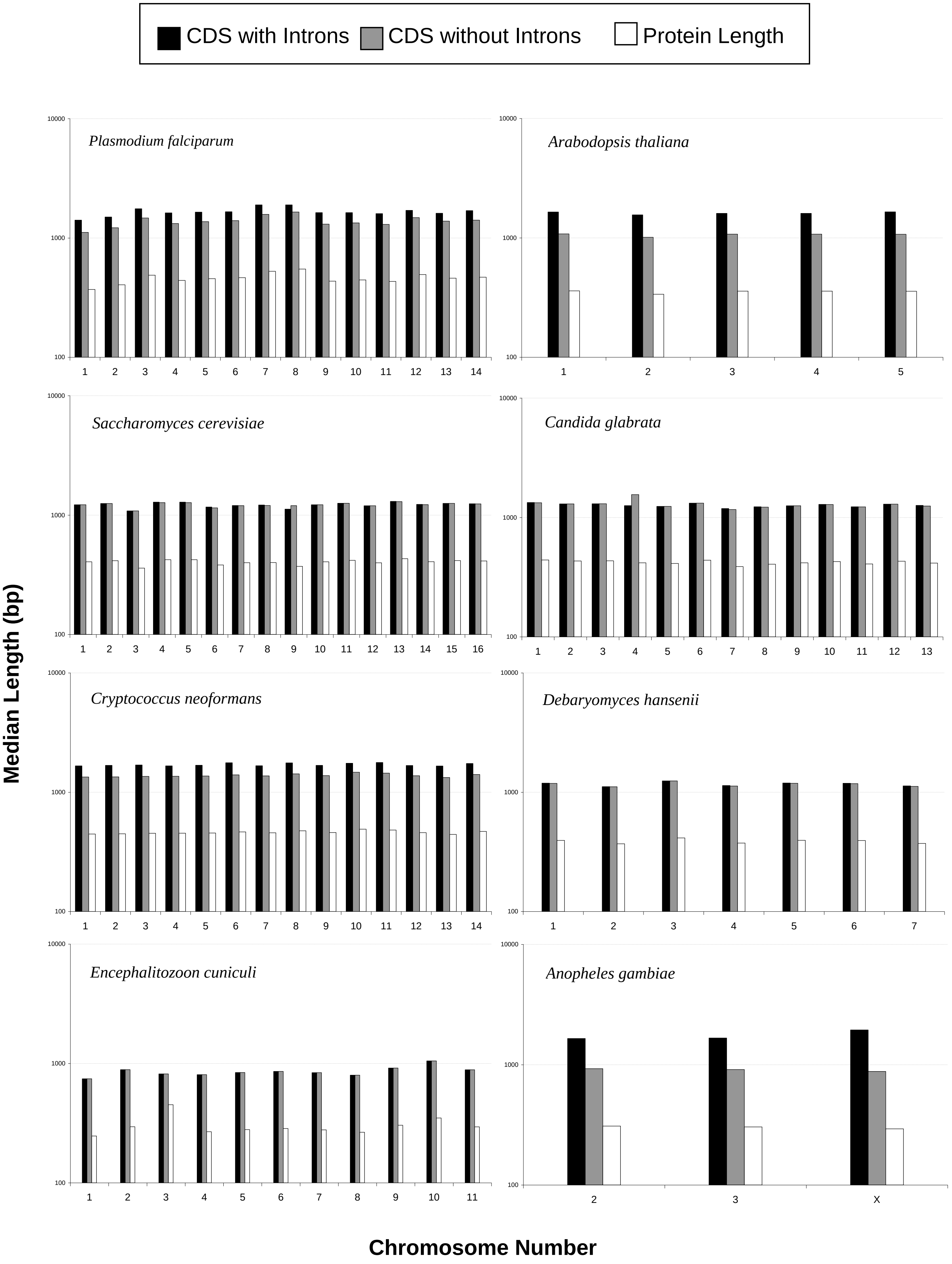


Figure 4.4 Density of proteins per Mb within each chromosome, across 15 different genomes of eukaryotes

4.3.2 Protein coding sequences and chromosomes

The median value was considered in place of the mean values due to the skewness in the data. Log scale was used to display the relative distribution of gene length of each organism due to the values ranging over many orders of magnitude. The lower species, which include the fungi *C. glabrata, C. neoformans, D. hansenii* and *E. cuniculi*, indicated little variation between the CDS that do not contain introns (Figure 4.5). Most values fell in the range of log scale 1000 for both the coding sequence that contained introns and excluded introns. As the species move towards more complex organisms such as insects and animals, the difference between the CDS with and without introns is obvious. The largest differences shown within the species included the higher animals such as the *M. musculus*, *P. troglodytes*, *D. rerio* and *H. sapiens* (Figure 4.5). An example in *H. sapiens* showed most coding sequences that contained introns averaged around the log of 1000, whereas the coding sequences that had the excluded introns were considerably smaller, averaging just over the log of 100 base pairs.

The ANOVA analysis identified significant differences between each chromosome for the three categories, CDS with introns, without introns and protein length. For all the higher organisms, *C. elegans*, *D. Melanogaster*, *D. rerio*, *M. musculus*, *P. troglodytes* and *H. sapiens*, as well as the plant species *A. thaliana*, there were significant differences between chromosomes ($p < 0.05$) (Table 4.2). The fungi species showed no significant differences between chromosomes. Additionally, *T. castaneum*, and *A. gambiae* only showed a significant difference between chromosomes when analysed with the CDS with intron data (Table 4.2). The median protein lengths of all species range from 300 to 450 and are consistent with previously published results on eukaryotes (Brocchieri and Karlin, 2005).


■ CDS with Introns
▨ CDS without Introns
□ Protein Length
Plasmodium falciparum
Arabodopsis thaliana
Saccharomyces cerevisiae
Candida glabrata
Cryptococcus neoformans
Debaryomyces hansenii
Encephalitozoon cuniculi
Anopheles gambiae
10000
1000
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14
1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10 11
2 3 X
Median Length (bp)
Chromosome Number

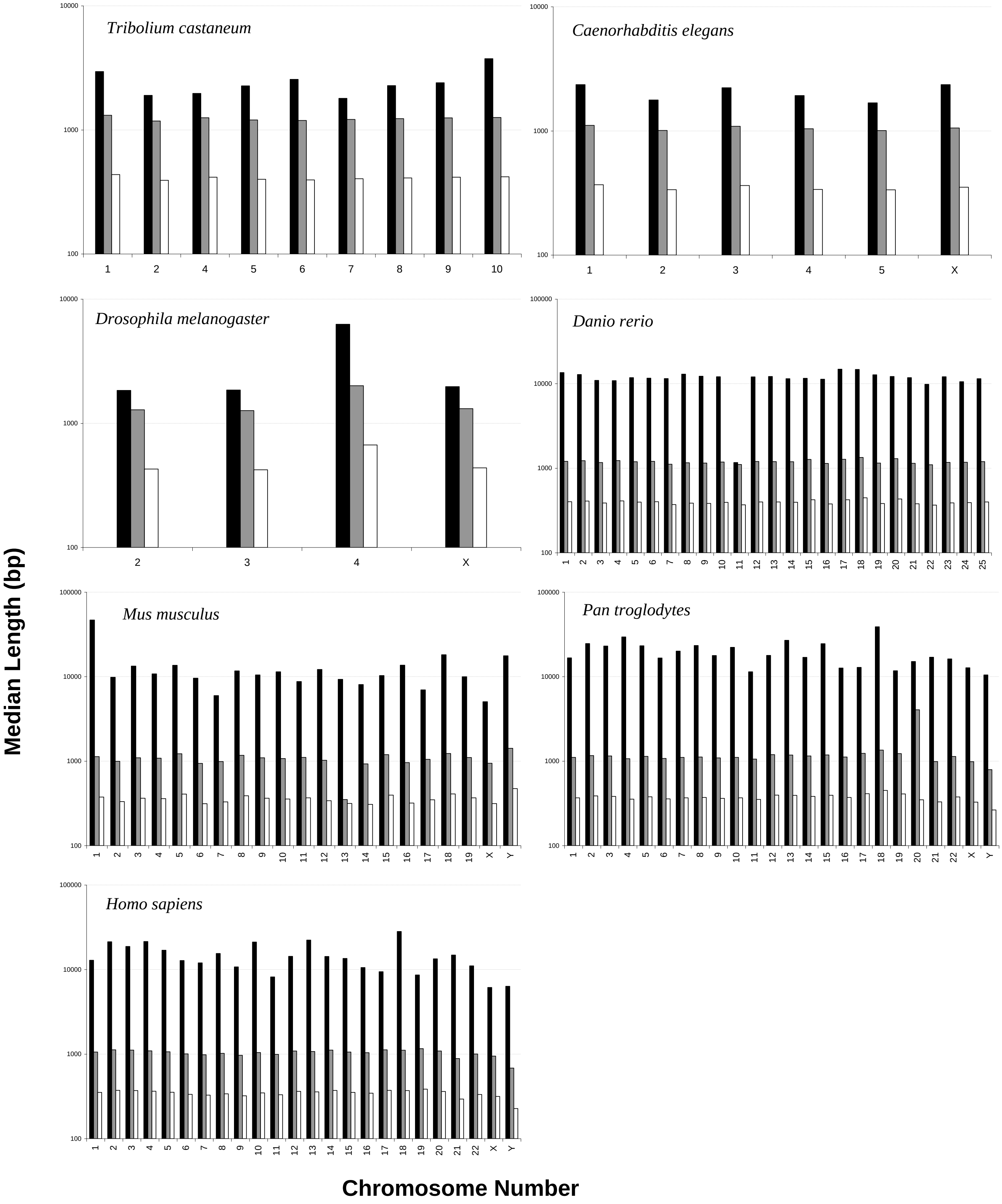


Figure 4.5 Median length (bp) across chromosomes of 15 species

Median length included values covering the coding sequence with introns, coding sequence without introns and protein length. The coding sequence with introns was calculated by subtracting the stop position by the start position plus 1 for each protein entry. The coding sequence without introns was calculated by multiplying the protein length by 3 and adding 3. Protein length was obtained from the protein tables found on the RefSeq proteins tables at http://www.ncbi.nlm.nih.gov/genomes/. Median was calculated by each individual protein entry for the CDS with and without introns in JMP statistical package. Each graph used a log scale for median length to show relative distribution of values.

Table 4.2 Summary of ANOVA analysis and Kruskal-Wallis Test for each eukaryote species

Comparison was made among the chromosomes of each species for the data: CDS with introns, without introns and protein length. The coding sequence with introns was calculated by subtracting the stop position by the start position plus 1 for each protein entry. The coding sequence without introns was calculated by multiplying the protein length by 3 and adding 3. Protein length was obtained from the protein tables found on the RefSeq proteins tables at http://www.ncbi.nlm.nih.gov/genomes. Analysis was conducted through JMP$^{®}$, Version **<9>**. SAS Institute Inc., Cary, NC, 1989-2007 for ANOVA analysis, Kruskal-Wallis hypothesis testing was run on SPSS v24.

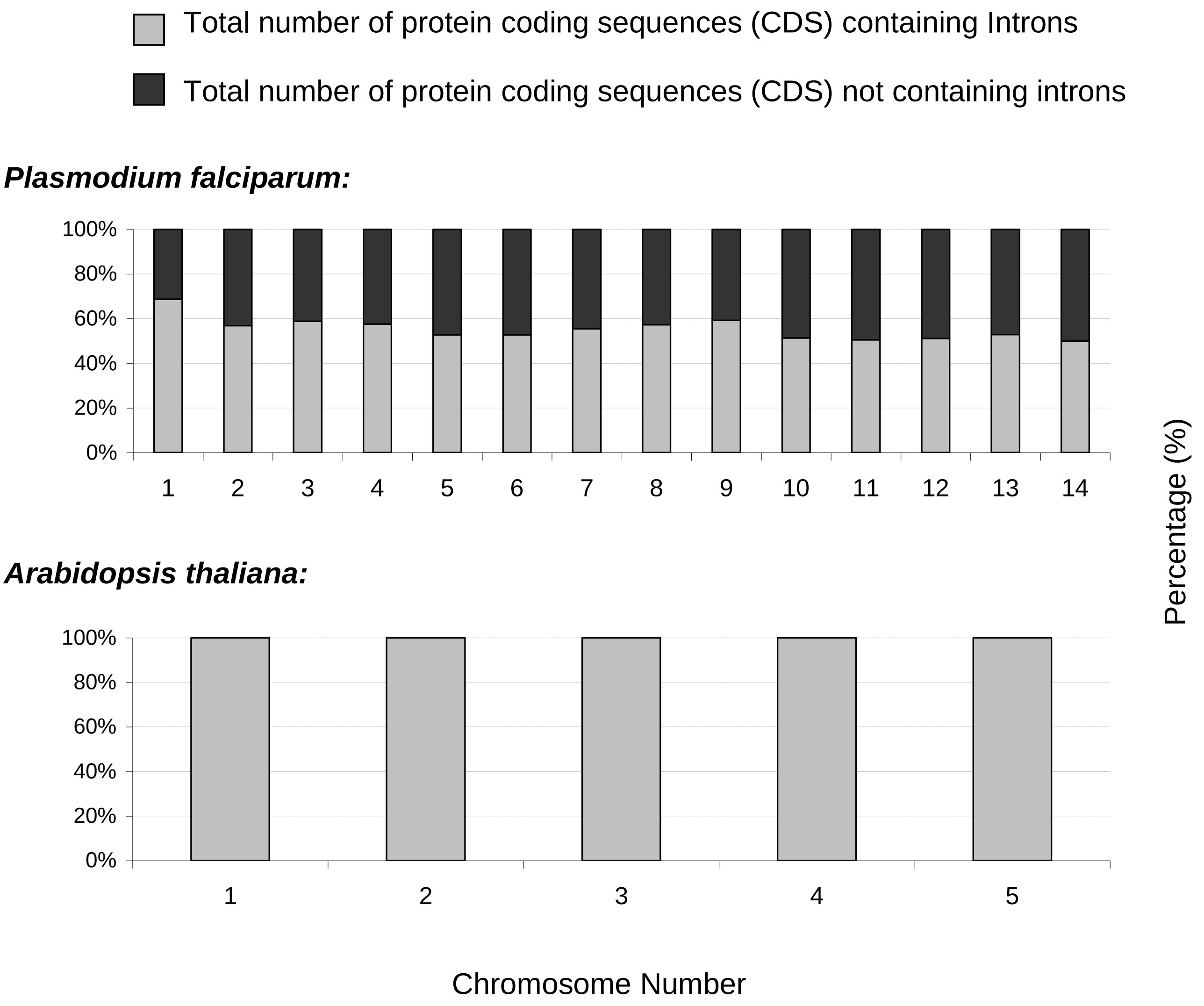
Source	d.f.	*F statistic*	*P*	Kruskal-Wallis Test *P*	Source	d.f.	*F statistic*	*P*	Kruskal-Wallis Test *P*
Plasmodium falciparum					***Tribolium castaneum***				
CDS with Introns	13	1.5525	0.0912	0.118	CDS with Introns	8	3.8063	0.0002*	0.000*
CDS without Introns	13	1.7064	0.0529	0.121	CDS without Introns	8	1.1602	0.3193	0.287
Protein Length	13	1.7064	0.0529	0.121	Protein Length	8	1.1602	0.3193	0.287
Arabidopsis thaliana					***Caenorhabditis elegans***				
CDS with Introns	4	6.9136	< 0.0001*	0.000*	CDS with Introns	5	67.8998	<0.0001*	0.000*
CDS without Introns	4	7.3591	< 0.0001*	0.000*	CDS without Introns	5	16.7514	<0.0001*	0.000*
Protein Length	4	7.3591	< 0.0001*	0.000*	Protein Length	5	16.4552	<0.0001*	0.000*
Saccharomyces cerevisiae					***Drosophila melanogaster***				
CDS with Introns	15	0.4780	0.9527	0.702	CDS with Introns	3	19.3282	<0.0001*	0.000*
CDS without Introns	15	0.4691	0.9565	0.737	CDS without Introns	3	37.7835	<0.0001*	0.000*
Protein Length	15	0.4691	0.9565	0.737	Protein Length	3	37.7835	<0.0001*	0.000*
Candida glabrata					***Danio rerio***				
CDS with Introns	12	0.3361	0.9828	0.923	CDS with Introns	24	2.8066	<0.0001*	0.000*
CDS without Introns	12	0.3580	0.9774	0.883	CDS without Introns	24	3.6758	<0.0001*	0.000*
Protein Length	12	0.3580	0.9774	0.883	Protein Length	24	3.6758	<0.0001*	0.000*
Cryptococcus neoformans					***Mus musculus***				
CDS with Introns	13	0.6110	0.8472	0.894	CDS with Introns	20	9.9967	<0.0001*	0.000*
CDS without Introns	13	0.6873	0.7778	0.822	CDS without Introns	20	6.3825	<0.0001*	0.000*
Protein Length	13	0.6906	0.7745	0.824	Protein Length	20	6.3825	<0.0001*	0.000*
Debaryomyces hansenii					***Pan troglodytes***				
CDS with Introns	6	1.2986	0.2540	0.051	CDS with Introns	23	15.5634	<0.0001*	0.000*
CDS without Introns	6	1.4381	0.1957	0.027*	CDS without Introns	23	2.4144	<0.0001*	0.001*
Protein Length	6	1.4381	0.1957	0.027*	Protein Length	23	2.4144	<0.0001*	0.001*
Encephalitozoon cuniculi					***Homo sapiens***				
CDS with Introns	10	1.3512	0.1974	0.087	CDS with Introns	23	23.2665	<0.0001*	0.000*
CDS without Introns	10	1.3429	0.2016	0.092	CDS without Introns	23	5.0335	<0.0001*	0.000*
Protein Length	10	1.3429	0.2016	0.092	Protein Length	23	5.0335	<0.0001*	0.000*
Anopheles gambiae									
CDS with Introns	2	3.0240	0.0486*	0.004*					
CDS without Introns	2	0.5379	0.5840	0.099					
Protein Length	2	0.5379	0.5840	0.099					

*Significant at α = 0.05 (differences between chromosomes)

Drosophila melanogaster was used as an example to perform a student *t* test to determine which chromosomes varied from each other, found earlier in the ANOVA analysis. The *Drosophila melanogaster* exhibited a difference between chromosome 4, and the other chromosomes, X, 3 and 2, for all categories. The CDS without introns and protein length also displayed differences between chromosome 3 and X. Identification of differences between chromosomes within individual species has been found, particularly within the sex chromosomes during the regulation of transcription (Brown and Bachtrog, 2014).

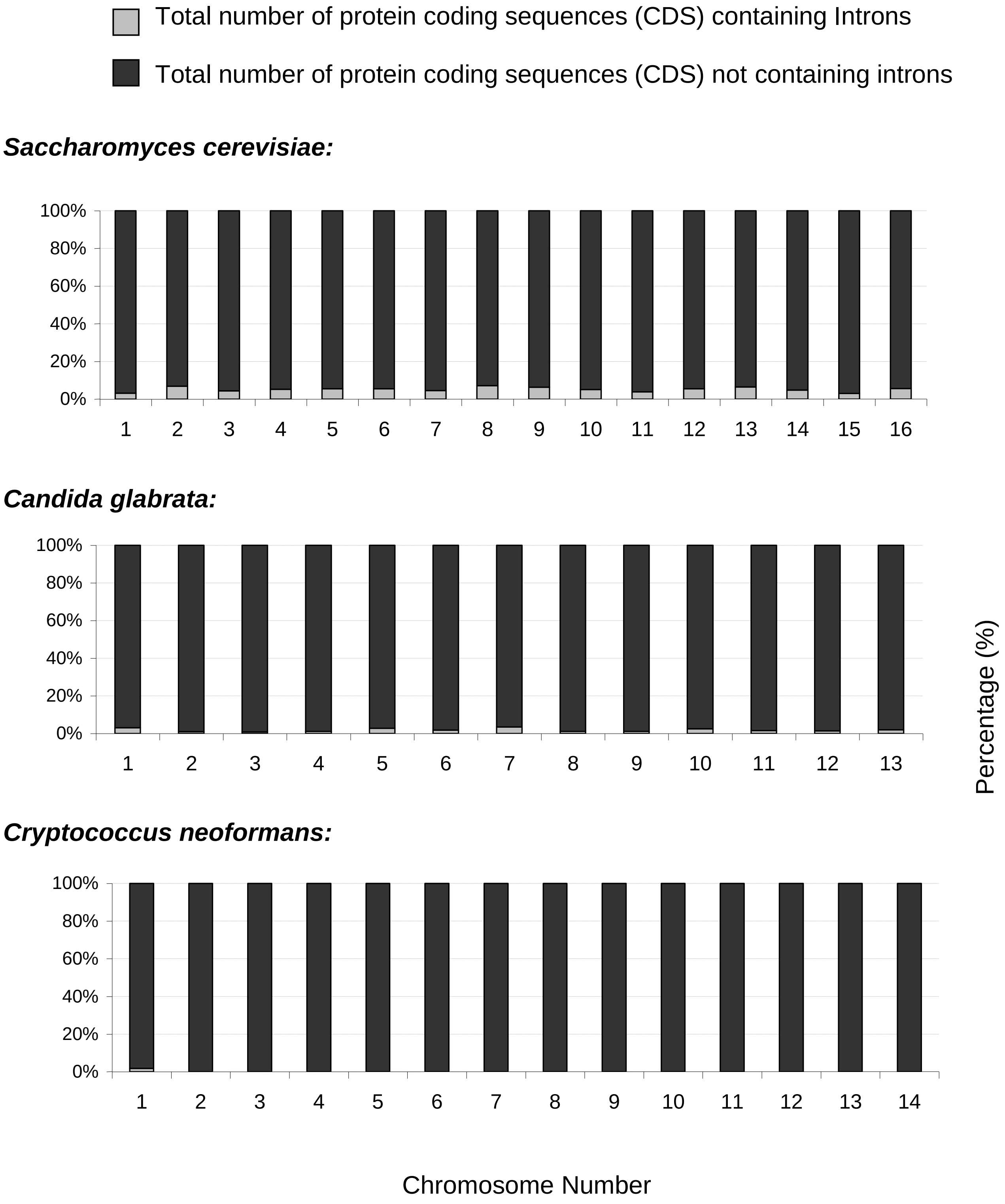
Two species studied, *A. thaliana* and *A. gambiae*, exhibited all chromosomes containing coding sequences with introns (Figure 4.6). *P. falciparum* showed a comparatively even spread over all chromosomes of coding sequences that either contain or lack introns.

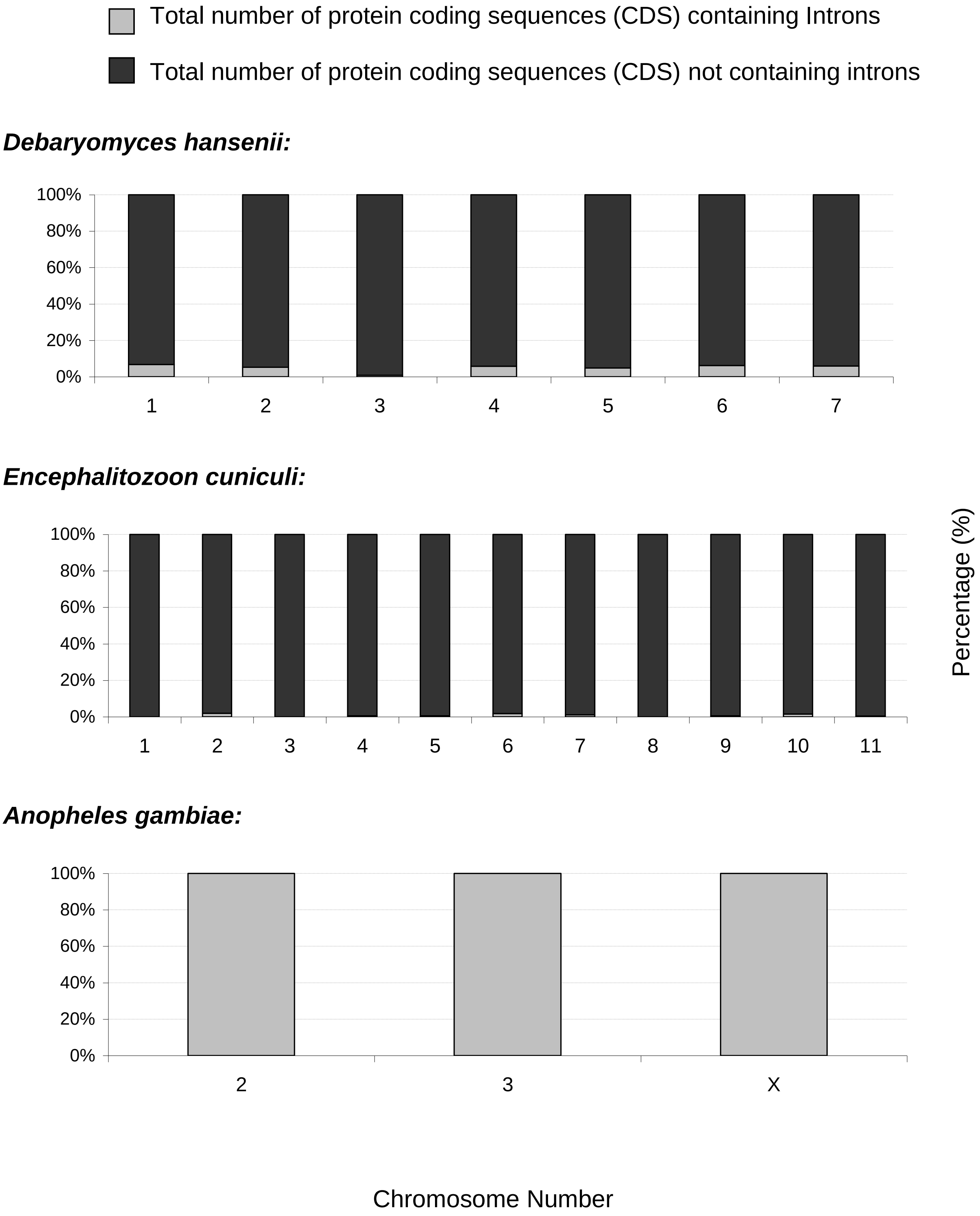
All species of fungi showed little or no presence of introns within the coding sequences of all chromosomes. Within the animal kingdom, most species displayed a large portion of coding sequences with introns among all chromosomes. *M. musculus* had the largest proportion of coding sequences without introns, with the Y chromosome showing the largest percentage. This was also seen in the *H. sapiens*, with chromosomes X and 21 having the largest percentage (Figure 4.6).

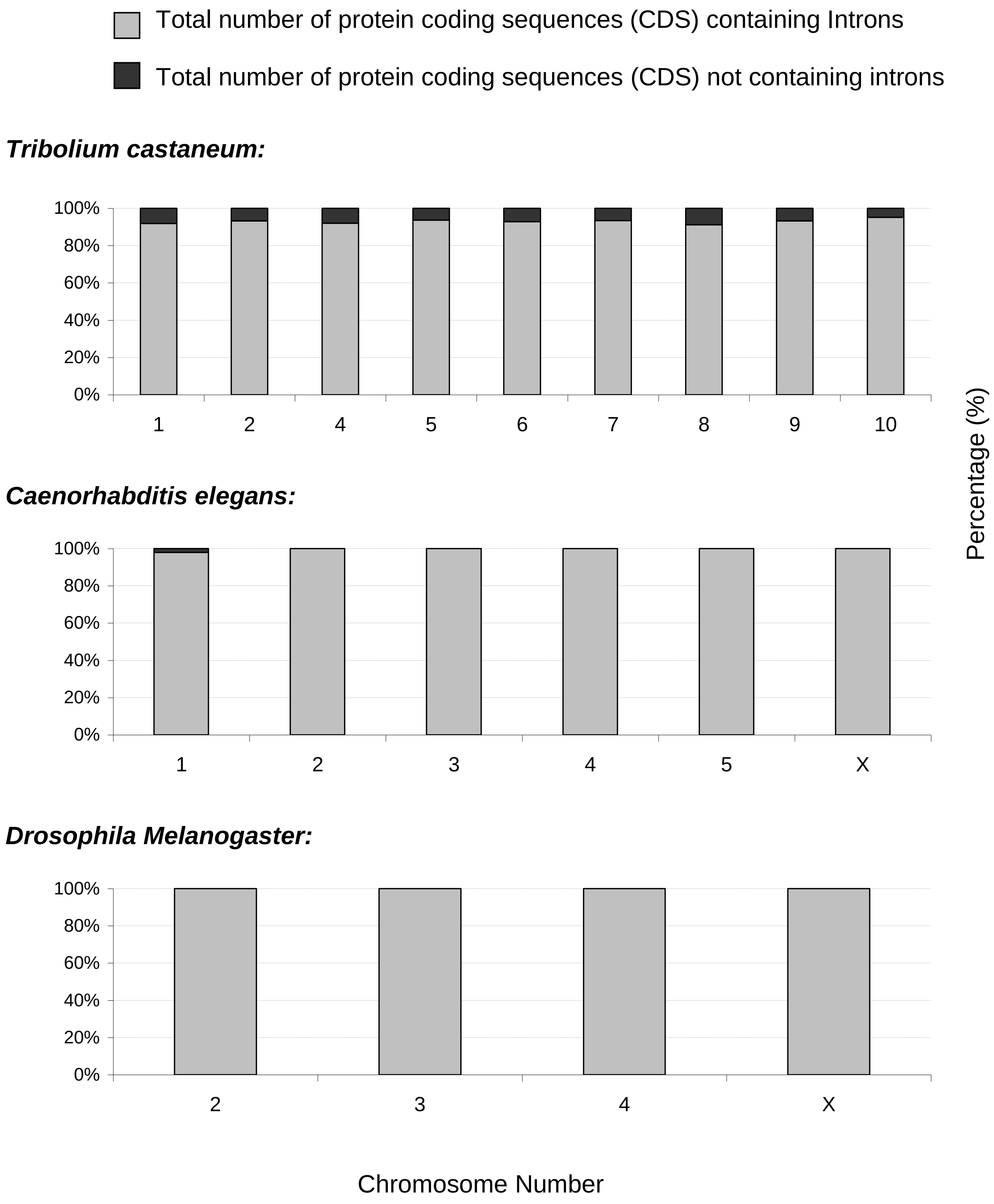


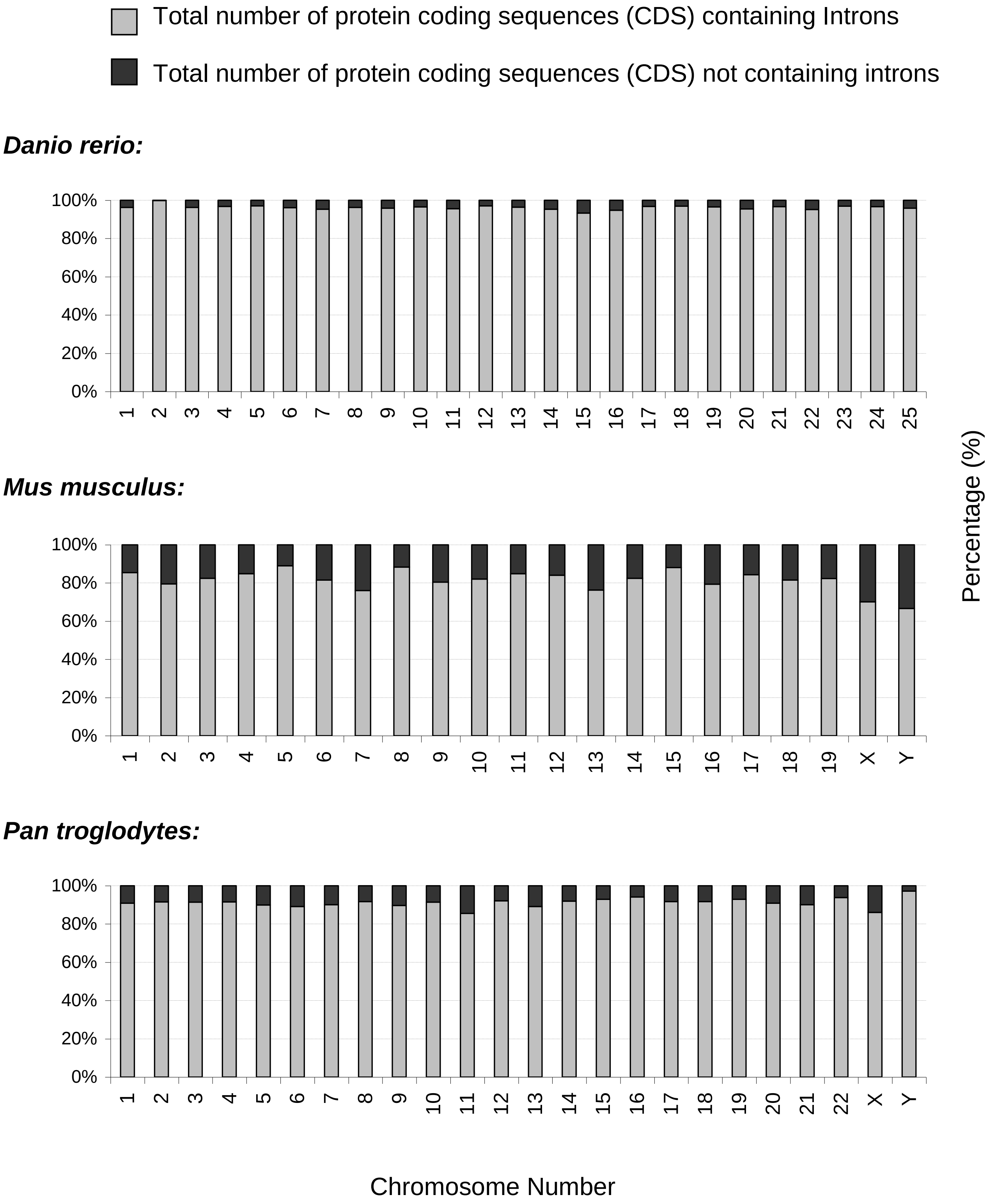
Plasmodium falciparum:

Arabidopsis thaliana:


Total number of protein coding sequences (CDS) containing Introns
Total number of protein coding sequences (CDS) not containing introns
Saccharomyces cerevisiae:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Candida glabrata:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13
Percentage (%)
Cryptococcus neoformans:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Chromosome Number


Total number of protein coding sequences (CDS) containing Introns
Total number of protein coding sequences (CDS) not containing introns
Debaryomyces hansenii:
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
7
Encephalitozoon cuniculi:
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
7
8
9
10
11
Anopheles gambiae:
100%
80%
60%
40%
20%
0%
2
3
X
Percentage (%)
Chromosome Number


Total number of protein coding sequences (CDS) containing Introns
Total number of protein coding sequences (CDS) not containing introns
Tribolium castaneum:
100%
80%
60%
40%
20%
0%
1
2
4
5
6
7
8
9
10
Caenorhabditis elegans:
100%
80%
60%
40%
20%
0%
1
2
3
4
5
X
Drosophila Melanogaster:
100%
80%
60%
40%
20%
0%
2
3
4
X
Percentage (%)
Chromosome Number


Total number of protein coding sequences (CDS) containing Introns
Total number of protein coding sequences (CDS) not containing introns
Danio rerio:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Mus musculus:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y
Pan troglodytes:
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Percentage (%)
Chromosome Number

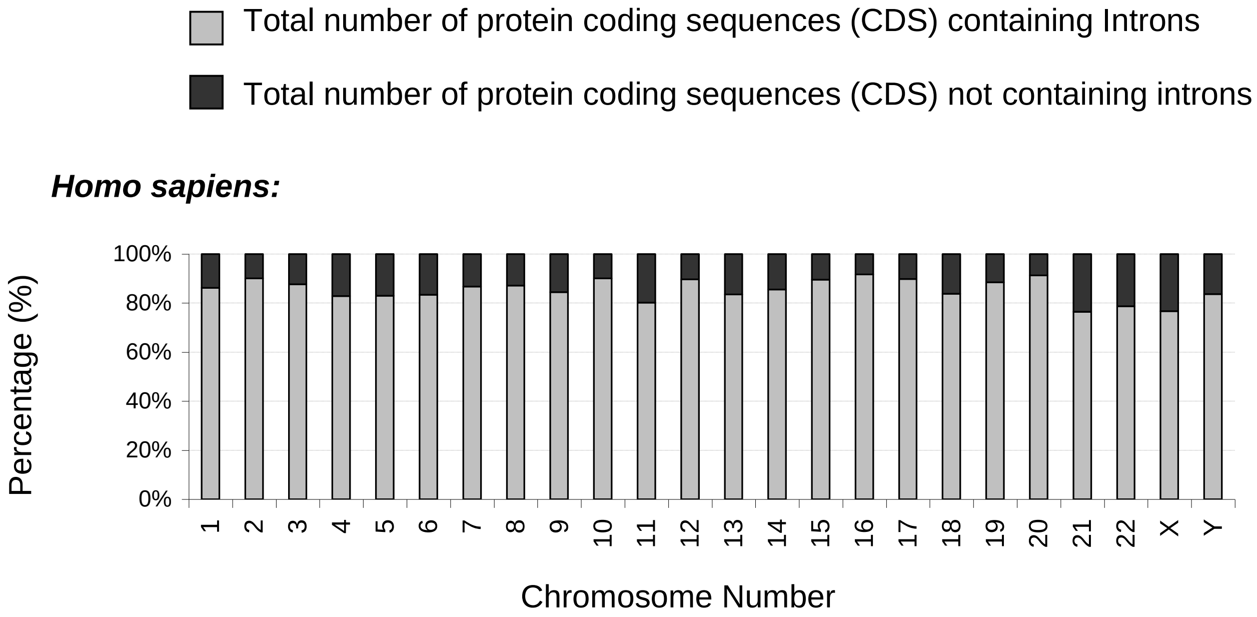


Figure 4.6 Total percentage of protein coding sequences (CDS) containing either introns or lacking introns with each chromosome of 15 different eukaryotes

The eukaryotes studied were *Plasmodium falciparum*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Anopheles gambiae*, *Tribolium castaneum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, *Pan troglodytes* and *Homo sapiens*. For the 15 different genomes of eukaryotes the number of proteins containing either introns or lacking introns for each chromosome was calculated from an MS Access database, where the CDS with introns table was compared to the CDS without introns table. The coding sequence with introns was calculated by subtracting the stop position by the start position plus 1 for each protein entry. The coding sequence without introns was calculated by multiplying the protein length by 3 and adding 3. Data was obtained from the RefSeq proteins tables found at http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi web site.

4.4 Discussion

In this chapter, 15 eukaryotic species were compared in relation to the coding sequence of proteins that included either introns or lacked introns, as well as the protein length. Observed differences between chromosome and species were established within the higher organisms, however the lower organisms such as the fungi species showed no observed differences. ANOVA and Kruskal-Wallis testing was performed on the data to support the observations made. Both tests confirmed significant differences between CDS with and without introns and protein length. There was only one discrepancy with the test for one organism, that being *D Hansenii*. Density of proteins per Mb and the total number of protein coding sequences for each chromosome that contained introns was also examined. This study has achieved its goals by verifying the proteins lengths of well documented eukaryotic species, and finding significant differences between chromosomes.

Within the eukaryote chromosomes there was differences only seen within the higher organisms. The biological significance from these results are still indeterminate, however

Zhang (2000) has proposed that larger proteins may be more complex in function than those that are smaller. Zhang (2000) also determined in the nematode and *Drosophila* that protein length and expression are positively correlated, supporting the concept that highly expressed genes are perhaps more important (Zhang, 2000).

Eukaryotic protein lengths have also been connected to 'functional regulators' sequence motifs that are interlaced within the protein coding sequences (Brocchieri and Karlin, 2005). Tan et al (2005) suggested that the longer proteins are essential in eukaryotes since these proteins are more connected, and this seems to be true especially for the higher eukaryote species, as seen in this chapter.

Bigger proteins may not be permissible in prokaryotes, due to the fact that the prokaryote genomes are highly compacted and must only retain those genes that are imperative (Zhang, 2000). However, in eukaryotes, multi-domain structures are formed, and may be linked with the evolution of the multi-exon proteins (Brocchieri and Karlin, 2005). The biological significance of the larger protein lengths within eukaryotes may be explained by the synthesis of single units seen in prokaryotes to create multi-domain units (Brocchieri and Karlin, 2005). The production of gene regulation networks in eukaryotes has been suggested by Zhang (2000) to be an evolutionary strategy, increasing protein lengths among higher organisms.

Environmental conditions and its' impact for all species of bacteria, archaeal and eukaryotes have also been related to the lengths of proteins. It has been suggested by Brocchieri & Karlin (2005) that in harsh, high temperature environments the evolution of shorter, and more stable proteins are chosen over more complex proteins. This was shown in small proteomes of parasitic organisms, which had longer median proteins due to the protected environment in which they live (Brocchieri and Karlin, 2005). The minimization of amino acid usages has also been connected to the length of proteins. Again, in environments that offer starving conditions, the selective pressure for the removal of more expensive proteins, has influenced the adaptive process for free-living species (Brocchieri and Karlin, 2005). A study that focused on *P falciparum* and *S cerevisiae* found that parasitism influences redundancy within each of these genomes. As *P falciparum* exhibits parasitism, it was found that there was a higher level of redundancy in the chromosomes compared to the similar sized *S cerevisiae* chromosomes (Achaz, et al., 2001).

Most eukaryote species studied exhibited the presence of introns for each protein coding sequence. There was selected species that only had a small portion of introns within each chromosome. Looking at a diverse range of eukaryotic genomes, it is still uncertain what mechanisms are responsible for the increase in introns gained in the last ½ billion years (Stoltzfus, 2004). Studies showing the intron gain and loss for mammalian genomes appear to be almost static, for example, between the human and mouse genomes there have only been a loss of 0.003 introns per gene with no clear gains (Jeffares, et al., 2006). With the separation of each chromosome for each species of eukaryotes, it may be possible to identify the areas in each species where the loss and gain of introns occurs (Roy and Gilbert, 2005). This may lead to a better understanding of the functions of these introns within an organism and between different species.

Furthermore, it has been found in other studies that a certain number of introns have important functions in multicellular eukaryotes, however, the proportion of these introns has not been identified up to now. If each chromosome is studied individually, it may make such an extensive task possible (Jeffares, et al., 2006; Vinogradov, 2002). This study has identified chromosomes, in the higher organisms that do have significant differences in each coding sequences with introns. The higher organisms also show a higher number of introns within each chromosome than fungi, which had a very small portion of protein coding sequences containing introns. Jeffares et al (2006) have suggested that the introns that have specific functions may have become essential in multicellular organisms and once they have been created, are not easily lost. This may explain why the higher organisms such as mouse (*M musculus*), human (*H sapiens*) and chimpanzee (*P troglodytes*) have a large portion of introns within each of their chromosomes. Why a particular chromosome has increased numbers of introns would need further investigation.

The fungi species studied had a very small portion of coding sequence that contained introns. Jeffares et al (2006) indicated that introns have been eliminated completely from highly reduced genomes. This would suggest that organisms such as prokaryotes that do not contain introns, may have come from ancestors that were intron rich (Jeffares, et al., 2006). This may explain why some of the fungi species did have a small percentage of introns within some of their chromosomes, and have been removed due to selection pressure. Roy (2006) related intron loss to evolution, suggesting it is driven by positive selection, meaning in larger population sizes the rates of intron loss would be greater than those seen in smaller population sizes. However, this study did highlight that these findings are not consistent with other intron losses, indicating another mechanism affecting the loss, which may include generation time (Roy, 2006).

Introns have been a good tool in previous studies to identify deletions of small scale mutations (Ogata, et al., 1996). The presence of introns with those organisms that did show differences within each chromosome, and indicated larger intron sizes, could suffer point mutations and deletions at a higher rate (Ogata, et al., 1996). The differences between chromosomes may be explained by different conditions that affect different parts of the genome in respect to these mutations (Ogata, et al., 1996). In *Drosophila*, the coding sequence with introns, particularly in chromosome 4, was very large. On an adaption basis, this could be explained by natural selection, where selection against very long introns is unproductive (Comeron, 2001).

It is still unclear why there is such a wide variation in the amount of noncoding DNA in genomes (Vinogradov, 2002). Lin and Zhang (2005) proposed that in *C. elegans* the total number of genes that did not contain intron was 2.7% for the whole genome compared to *P. troglodytes* which was 9.2%, *M. musculus* which was 16.1%, and *D. melanogaster* being 21.6% (Lin and Zhang, 2005). Our findings support these percentages, for example in *C. elegans* the study found the sequencing without introns only in chromosome 1. Vinogradov (2002) indicated that there is a general correlation between the genome size and the intron size in a wide variety of evolutionary diverse phyla. Vinogradov (2002) also indicated that the smaller the genome the more deletions are favoured over insertions (Vinogradov, 2001; Vinogradov, 2002). This could refine the results found in this study and explain why there were some differences in each chromosome. The balance of coding sequences that contained introns or lacked introns were very diverse, not only with each chromosome, but within the species themselves.

A study by Achaz et al (2001) investigated six species of eukaryotes (*S. cerevisiae, C. elegans, P. falciparum, A. thaliana, D. melanogaster and H. sapiens*). The results, based on the analysis of intrachromosomal repeats, indicated biological significance. The significance implied structures and mechanisms that are connected in the eukaryote kingdom, and are shared by all eukaryote chromosomes. Dujon (2006) looked at chromosome fragments and established that they aided in the identification of the whole-genome duplication process (Dujon, 2006).

The research in this chapter has extended the understanding and has opened more discussion on the role of length within and between organisms. In conclusion, this early study has highlighted differences between chromosomes for each organism, when examining the coding sequence (with and without introns) and protein length. Notably there were differences between the more complex organisms when the results were compared to the lower species. An understanding of intrachromosomal duplications have been studied by Achaz *et al* (2000) where it was implied that the coding repeats are conserved by functional pressures, and must be short due to the effect of length tolerance (Achaz, et al., 2000).

The caveats existing in influencing the results from this research topic, may be the vertebrates studied, generally having the same identical intron/exon structures with little gain and loss of these introns from the diversity of rodents to primates (Lin and Zhang, 2005). The fungi species could also have the same limitation imposed, warranting caution when analysing the results from this research. Other issues associated with comparative genomics include a particular portion of the genomic region may be conserved only because of the lower mutation rate in that area (Andofatto, 2005).

To focus on two heavily studied model organisms and to extend on the understanding of how the 5' UTR and coding sequence interacts, the research of this thesis altered direction and focus, to use more complex and innovative statistics to determine the relationship between the coding and noncoding length regions.

Chapter 5 – Coding and Noncoding Sequence Length Comparison with *Arabidopsis* and *Drosophila*

5 Coding and Noncoding Sequence Length Comparison with *Arabidopsis* and *Drosophila*

5.1 Introduction

This chapter is slightly modified from the following two papers:

Caldwell R., Lin, Y., and Zhang, R. (2008) Correlations of Length Distributions between noncoding and coding sequences of the *Arabidopsis thaliana*, Chapter: 2008 IEEE International Conference On Bioinformatics and Biomedicine BIBM 2008 (Philadelphia, Pennsylvania, USA) edited by Xue-wen Chen, Xiaohua Hu, and Sun Kim, IEEE Computer Society, 72-77.

Caldwell, R., Lin, Y., and Zhang, R. (2010) Assessment of length distributions between noncoding and coding sequences amongst two model organisms, *International Journal of Data Mining and Bioinformatics*, 4 (5), 535-552. doi:10.1504/IJDMB.2010.035899.

Data in this chapter was collected and the research published in 2008 / 2009. More recent data and techniques are included in the preceding chapters.

With large-scale methods for data generation, becoming more efficient and cost effective, biological research is seeing an expansion in the discipline of bioinformatics. All areas of biology will ultimately use bioinformatics to pursue a large range of questions and it will encourage collaborations between disciplines.

One direction of research using the sequences of a wide range of organisms has been with protein length, elucidating the development and biological differences amongst the three domains of life. Genome complexity in relation to protein length is also examined and it was established there is a positive correlation between average protein lengths and genome complexity (Tan, et al., 2005).

However, at gene level, little is known of the length distributions of noncoding regions. Higher organisms only use a small portion of the genome for encoding proteins, with the other segments not coding for anything, even though still transcribed. Interest in the function of these noncoding regions has intensified. One study evaluated the distances between neighbouring genes and the lengths of the 3’ un-translated regions (UTRs) and it has been found that length and distance between genes and their corresponding un-translated regions had important implications in gene expression and regulation (Chiaromonte, et al., 2003; Clark, 2001). Other research has found that the intron presence in the 3’ UTR was far less than those found in the 5’ un-translated regions (Hong, et al., 2006).

Previous research conducted has indicated that for each organism studied the distribution of distance from transcription start site to translation start site displayed its own specific character, and so these distances varied among different organisms (Dai, et al., 2006). This is consistent with aforementioned studies focusing on protein length, with similar results in the increase in distance from simple prokaryotes to more complicated eukaryotic organisms.

Research conducted using the *Arabidopsis* cDNA data discovered many features of gene structure and organization (Alexandrov, et al., 2006; Seki, et al., 2002). The 5' and 3' UTR data for the large dataset confirmed previous study results, suggesting the average length of the 5' UTR length ranges between 100 to 200 nucleotides, whereas the 3' UTRs are much more variable (Mignone, et al., 2002). Not only can this data help elucidate gene regulation mechanisms, but also allows extended research on comparisons between phyla (Rubin, et al., 2000).

This chapter investigates the relationship between the noncoding (both 5' and 3') and coding sequence regions, which as of yet has not been attempted. Countless analyses of intricate biological processes still exploit the use of linear models. However, a number of studies have determined differences between the coding and noncoding DNA regions based on nonlinear dynamical characteristic's (Mabrouk, et al., 2008). We propose a nonlinear function statistical approach to establish correlations between the length distributions of the coding and noncoding regions of an animal and plants species. The data analysis also comprises the presence or absence of introns, as a comparison (Vinogradov, 2002).

5.2 Statistical Analysis

Descriptive statistics were obtained from JMP 9 (SAS Institute Inc., North Carolina U.S.A) and SPSS version 19 (SPSS IBM, New York, U.S.A) statistical software. Refer to the region of interest in chapter 1 and abbreviation list for a description of the "D, d" values. Comparisons were conducted on the ratio of each region (length value over total). After initial statistical tests, it was found there was a significant nonlinear relationship between the coding region $d_1(D_1)$ and the ratio of $D_2^*(d_2^*)$. The value of $D_2^*(d_2^*)$ was calculated by:

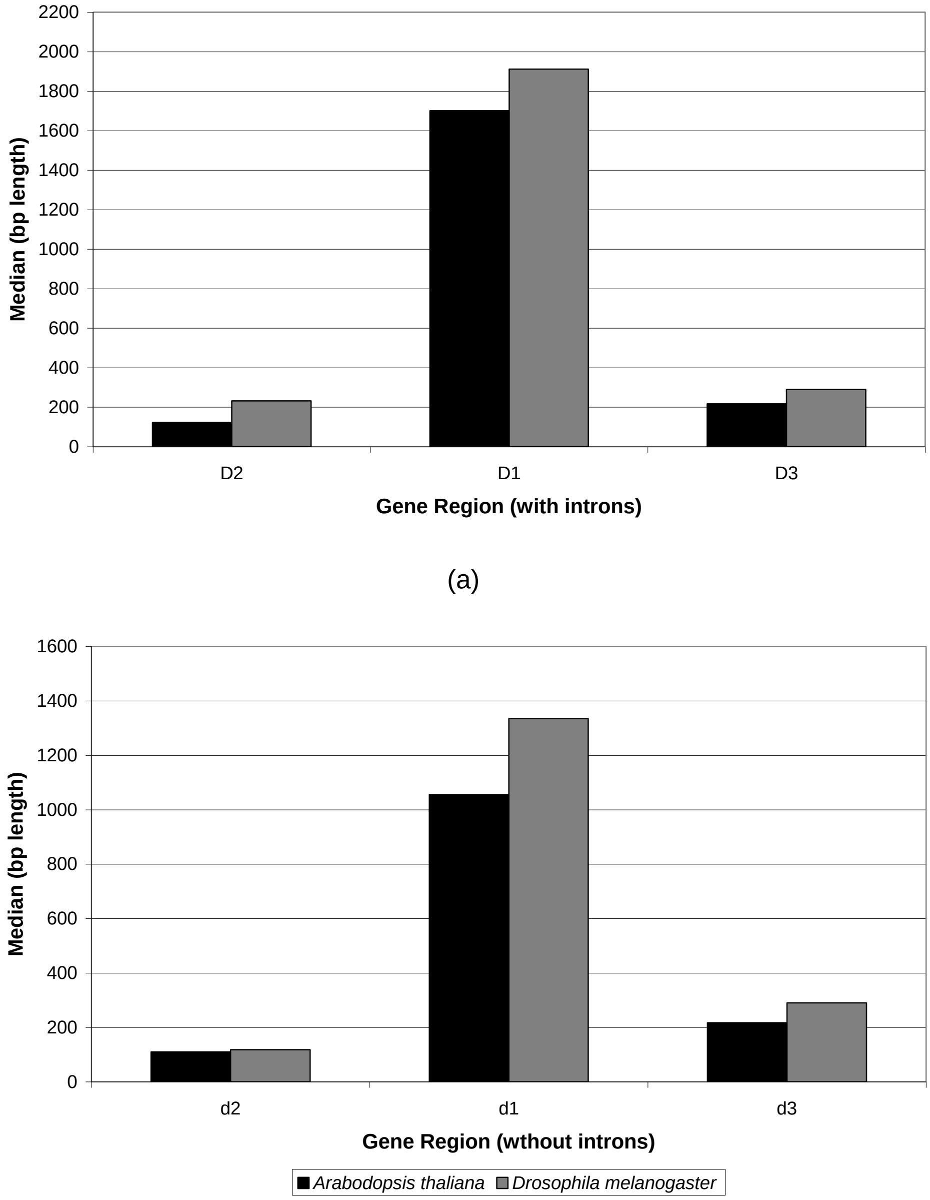
$$D_2^* = \frac{D_2}{D_1 + D_2 + D_3} \quad \text{and} \quad d_2^* = \frac{d_2}{d_1 + d_2 + d_3} \qquad (1)$$

The purpose is to predict d_1 through D_2^*, where D_2^* is the proportion of D_2 in the total length of protein coding gene (D_1+D_2+D_3). Analysis was conducted using JMP 9 (SAS Institute Inc., North Carolina U.S.A) and the data revealed significant nonlinear relationship between the coding region length $d_1(D_1)$ and the 5' noncoding region length ratio $d_2^*(D_2^*)$, which was conditional on the value of $\log(d_2) or \log(D_2)$. Each dataset for each organism was subset by the $\log(d_2) or \log(D_2)$ values and a nonlinear model was applied to each subset, to identify a nonlinear relationship between $d_1(D_1)$ and $d_2^*(D_2^*)$. In addition, ANOVA analysis was applied to each dataset to determine whether there are significant differences between the mean of each length region (coding and noncoding) and each protein category, which was grouped into four categories, information storage and processing, cellular processes and signalling, metabolism and poorly characterized based on the COG Functional categories. The datasets were imported into SPSS version 19 (SPSS IBM, New York, U.S.A) where mean testing analysis was performed.

5.3 Length Distributions among all Three Regions

Each gene region was examined and the median of the coding and noncoding region for each organism were obtained. Calculations were made on the data that included introns (D) and excluded introns (d). The median values for *Arabidopsis thaliana* ranged under 200 bps for the 5' un-translated region to over 1600 bps for the coding sequence. The 3' un-translated region values were just over 200 bps (Figure 5.1). Results from this research are comparable with previous analyses on this organism (Alexandrov, et al., 2006). Other studies have shown that the *Arabidopsis* 5' UTR average lengths range between 100 and 200 nucleotides. The 3' UTR for plants range from about 200 nucleotides (Mignone, et al., 2002). In comparison, the *Drosophila melanogaster* average length for the un-translated regions was diminutively larger than that of the plant species. Celniker & Rubin (2003) reported the size of the *Drosophila* un-translated regions as 265 nucleotides for the 5'UTR and 442 nt for the 3' UTR, which was confirmed by the data from this research.

The median value was considered in place of the mean values due to the skewness in the data. All data in each region, with and without introns when frequencies were plotted showed a long tail. The median length (bp) of the coding sequence without introns (d1) for *Arabidopsis* is ~62% of that of the coding sequence with introns (D1). In comparison, the median length for *Drosophila*, between the coding sequence with and without introns is ~70% (Figure 5.1). This indicates that in *Arabidopsis* the coding sequence region contains additional introns, than that of *Drosophila*.



(a)

(b)

Figure 5.1 Length distributions (bp) of noncoding and coding sequences of the Arabidopsis thaliana *and* Drosophila melanogaster.

Gene region consists of 5’ UTR median length with (a) and without introns (b) ($D_2(d_2)$); coding sequence median length with and without introns ($D_1(d_1)$); and 3’ UTR median length with and without introns ($D_3(d_3)$).

The noncoding region’s for *Arabidopsis* with and without introns shows higher median length percentages than the coding sequence. The median length for the 5’ UTR is ~89%, whereas the 3’ UTR is ~99.5%. The difference in the 5’ UTR and the 3’UTR in plants is consistent with other studies and may be attributed to nonsense mediated decay of mRNA (Alexandrov, et al., 2006; Hillman, et al., 2004). However, in *Drosophila* the 5’UTR median length was ~51%, suggesting a higher percentage of introns in this region, to the coding sequence. The 3' UTR was similar with the *Arabidopsis* showing a median length percentage of ~97% and again could be credited to the nonsense mediate decay of mRNA. Nonsense-mediated mRNA decay is a surveillance process to reduce errors in gene expression by eliminating mRNAs containing premature translation-termination codons (PTCs) (Brogna and Wen, 2009).

The datasets for each organism was further examined after being split into individual chromosomes. ANOVA analysis was applied to the data, and significant differences between each chromosome and each length region (noncoding and coding) were found, with and without introns (*p*-value < 0.001) (Figures 5.2 & 5.3). A large variation was observed in chromosome 4 for the *Drosophila* species between the length regions. This could be attributed to the small sample size for that particular chromosome once the data was compiled from the various data sources. Rearrangement changes, deletions, inversions and duplications in chromosomes are capable of accelerating species adaption as environmental conditions change (Coghlan, et al., 2005). This evolutionary influence can have an impact on the size, shape, and composition of eukaryotic chromosomes not only between organisms, but within a particular species (Schubert, 2007). This could substantiate why there were significant differences between the noncoding and coding lengths and each chromosome of these two species.

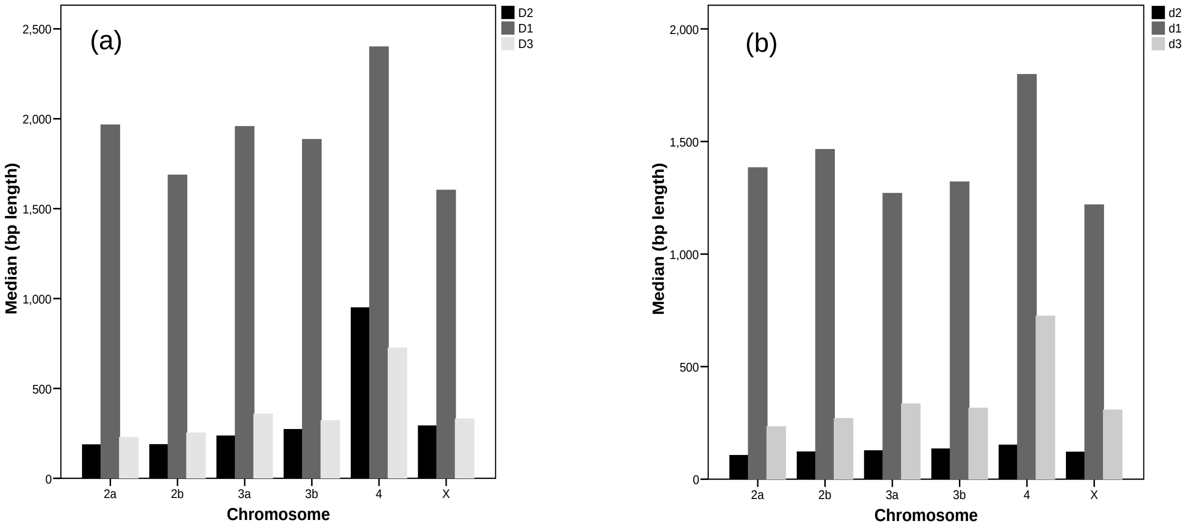


Figure 5.2 Median length values (bp) of each gene region (coding sequence ($D_1(d_1)$ *; 5' UTR (* $D_2(d_2)$ *) and 3' UTR (* $D_3(d_3)$ *) divided into chromosomes for* Drosophila melanogaster*. Figure (a) represents length regions with introns, and figure (b) represents length regions without introns.*

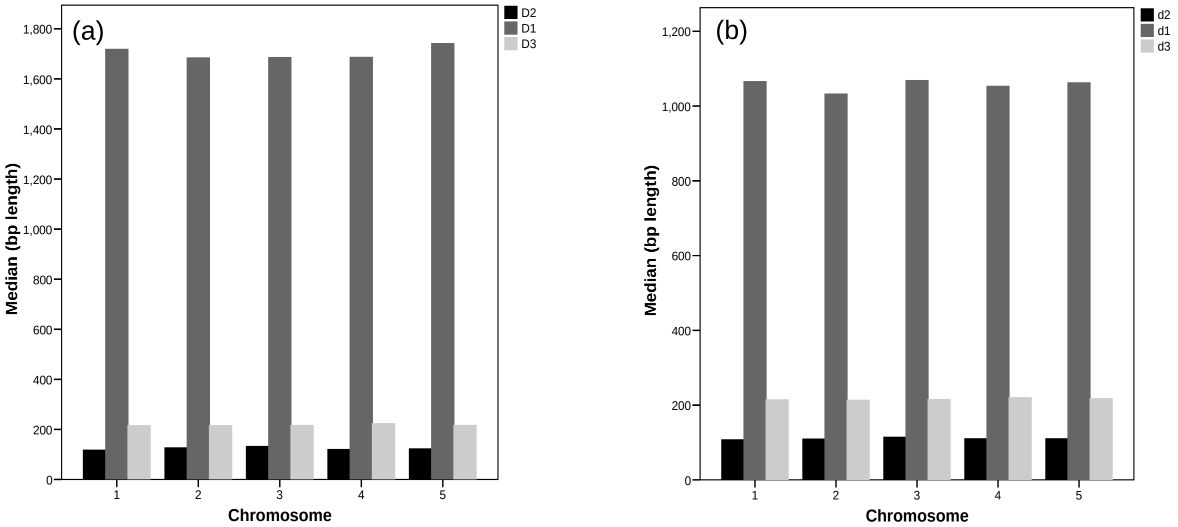


Figure 5.3 Median length values (bp) of each gene region (coding sequence ($D_1(d_1)$ *; 5' UTR (* $D_2(d_2)$ *) and 3' UTR (* $D_3(d_3)$ *) divided into chromosomes for* Arabidopsis thaliana*. Figure (a) represents length regions with introns, and figure (b) represents length regions without introns.*

5.4 Nonlinear Function relationship between $D_1(d_1)$ and $D_2^*(d_2^*)$ Values

Bivariate analysis was applied to the data to test for patterns and correlations between the coding and noncoding regions within each chromosome. Upon first inspection of the data, the relationship between the variables, showed more of a curved line, which prompted a nonlinear analysis approach. It was the intention to test the relationship between the coding and noncoding sequences using a nonlinear model, with the null hypothesis being that there is no relationship between the X and Y variables. By using log transformation, a nonlinear function relationship was established between d_1 and d_2^*

(D_1 and D_2^*) given the value of $\log(d_2) or \log(D_2)$. The log transformation was used on the $(d_2)/(D_2)$ data due to the highly skewed distributions and identified a clearer pattern in the data. Log transformation is often the first tool used when the data is faced with a curved relationship. The models used to fit the data are:

$$d_1 = \beta_0 + \beta_1 \frac{1}{d_2 *} + \beta_2(d_3 - d_2) + e \quad (e \text{ denotes random error}) \qquad (1)$$

and

$$D_1 = \beta_0 + \beta_1 \frac{1}{D_2 *} + \beta_2(D_3 - D_2) + e \quad (e \text{ denotes random error}) \qquad (2)$$

Where the β_0 β_1 and β_2 parameters in the model above are represented by the intercept and gradient estimates. The testing procedure is to run the data through a series of tests, firstly starting with linear regression, and add more terms to identify whether the R^2 is significantly greater than expected, and not due to chance. Once a best-fitting equation has been selected, it is tested for best fit against the linear model.

The data was subset, based on $\log(d_2) or \log(D_2)$ values. Splitting the data into these subsets proved the most accurate method in obtaining the best statistical outcome of the mathematical model used above, which was applied to each subset. The following results focuses on data without introns. From previous results data without introns performed better, and the data integrity from the external databases is proven. For the *Arabidopsis*, the data was subset into twelve subsets based on the value of $\log(d_2)$ (<1.0; 1.0-1.9; 2.0-2.4; 2.5-2.9; 3.0-3.4; 3.5-3.9; 4.0-4.4; 4.5-4.9; 5.0-5.4; 5.5-5.9; 6.0-6.4; > 6.4). Figure 5.4a shows nonlinear relationship between d_1 and d_2^* values based on the data from chromosome 4, within the subset of 3.5 ≤ $\log(d_2)$ < 4.0. Given the $\log(d_2)$ values >1.0, the R^2 values produced by the model are generally high (Figure 5.5). The R^2 values averaged around 0.9, making the correlation between these variables (d_1, d_2^* *and* $d_3 - d_2$) substantial. A similar trend was also seen in the *Drosophila* species, however the dataset for this organism was smaller therefore the data was subset into only four subsets (0 < 4.0; 4.0 – 4.9; 5.0 – 5.9; and > 6.0). The model was applied to each subset and strong correlation was also seen (Figure 5.4b - chromosome 2b subset 4.0 ≤ $log(d_2)$ <

4.9). Again, as the values of $\log(d_2)$ increased, so did the R^2 values to above 0.7 (Figure 5.6). For *Arabidopsis thaliana* the R^2 values were consistent, averaging around 0.9, however in the *Drosophila melanogaster*, there were variations seen across all subsets, with a sizeable drop in the R^2 value at subset 5.0 ≤ $\log(d_2)$ < 5.9. This could be attributed, again to the small sample size of the *Drosophila* or unexplained factors affecting the results, unseen by the model.

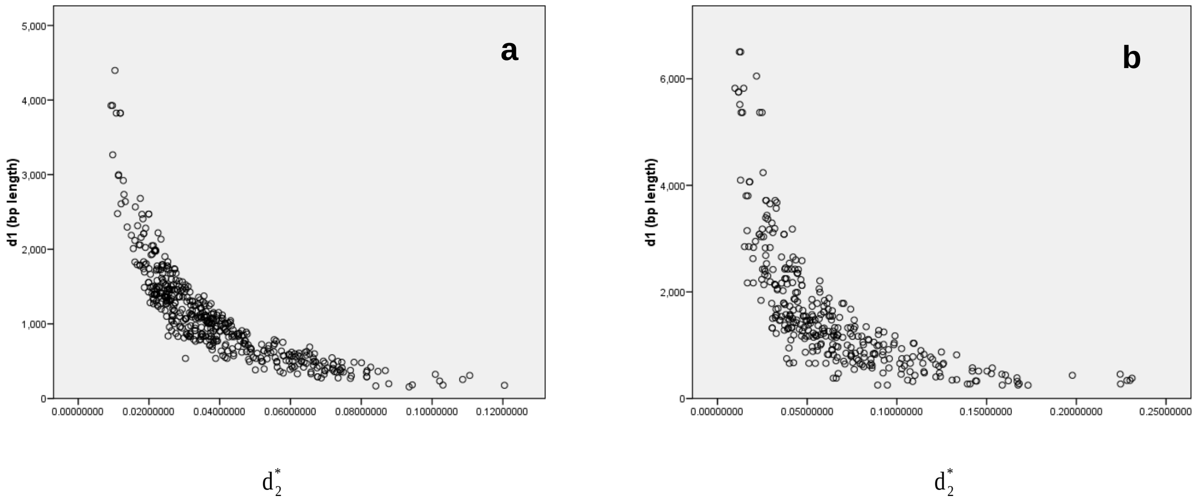


Figure 5.4 Nonlinear functional relationship between d_1 *and* d_2^*

Figure a represents chromosome 4 of the *Arabidopsis thaliana.* The data shown is from subset 3.5 ≤ *log*(d_2) < 4. Figure b represents chromosome 2b of the *Drosophila melanogaster.* The data shown is from subset 4.0 ≤ *log*(d_2) < 4.9.

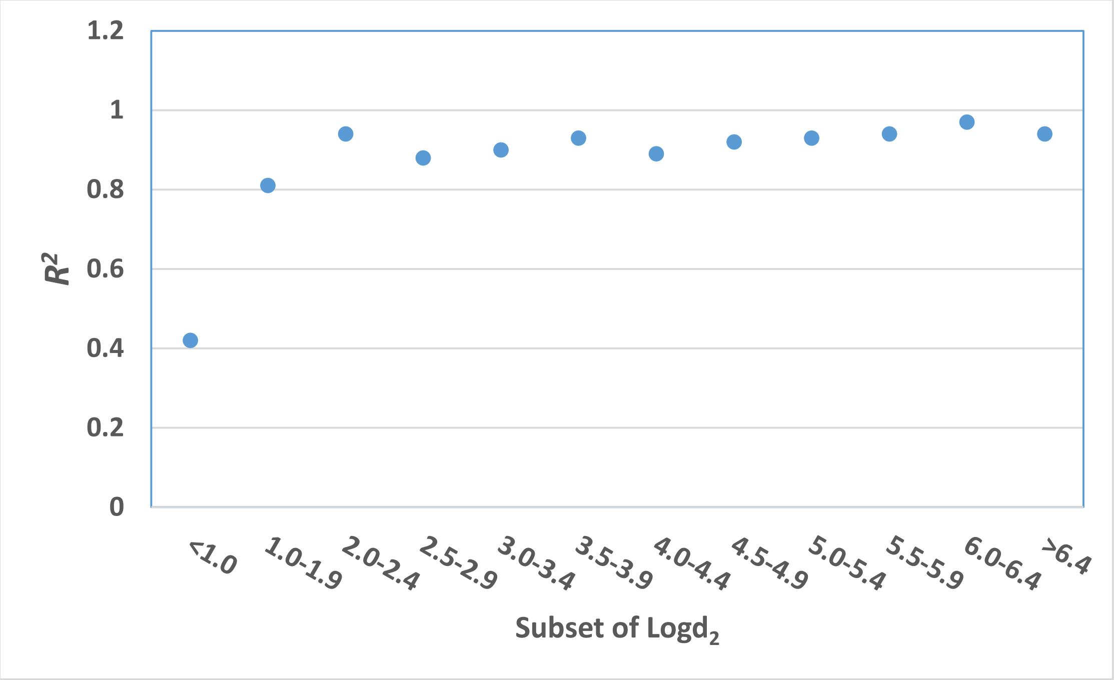


Figure 5.5 R-squared values based on nonlinear functional relationship between d_1 and d_2^* within chromosome 4 for Arabidopsis thaliana. Data was subset into 12 categories based on Logd₂ values.

Table 5.1 Summary of analysis based on the nonlinear model for Arabidopsis thaliana *on Chromosome 4. Data was subset into 12 categories based on Log(d_2) values.*

Summary of Fit	**Subset of *Log(d_2)***											
	<1.0	**1.0-1.9**	**2.0-2.4**	**2.5-2.9**	**3.0-3.4**	**3.5-3.9**	**4.0-4.4**	**4.5-4.9**	**5.0-5.4**	**5.5-5.9**	**6.0-6.4**	**>6.4**
RSquare	0.42	0.81	0.94	0.88	0.90	0.93	0.89	0.92	0.93	0.94	0.97	0.94
Root Mean Square Error	353	295	230	207	222	176	191	199	200	192	122	165
Observations	16	29	43	126	231	509	921	1110	844	429	171	44
β_0	471	390	92.5	54.2	95.4	-3.67	-31.38	-75.17	-155	-529	-911	-1900
β_1	0.67	2.78	8.57	15.1	25.4	41.58	64.70	103	163	282	445	793
β_2	0.39	-0.38	-0.78	-0.91	-1.54	-1.05	-0.95	-0.97	-1.14	-1.25	-1.31	-1.49

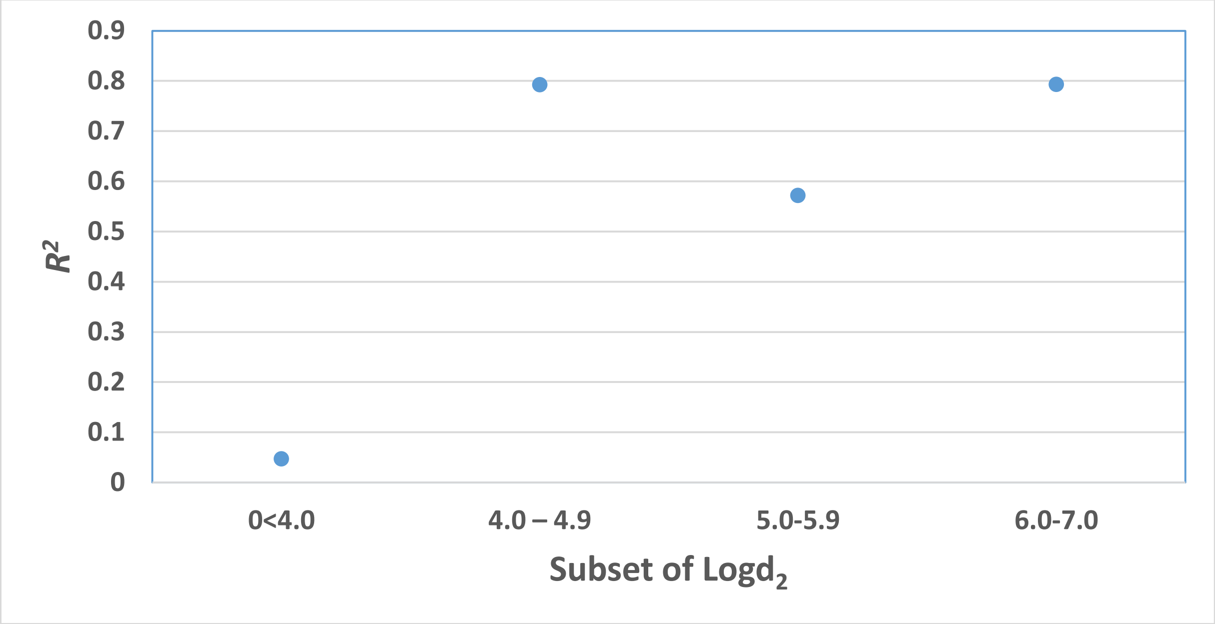


Figure 5.6 R-squared values based on nonlinear functional relationship between d_1 and d_2^* within chromosome 2b for Drosophila melanogaster. Data was subset into 4 categories on Logd$_2$ values.

Table 5.2 Summary of analysis based on the nonlinear model for Drosophila melanogaster *on Chromosome 2b. Data was subset into 4 categories based on Log(d$_2$) values.*

Summary of Fit	**Subset of *Log(d$_2$)***			
	0<4.0	**4.0 – 4.9**	**5.0-5.9**	**6.0-7.0**
RSquare	0.047541	0.792542	0.572273	0.793217
Root Mean Square Error	966.6695	452.2212	606.0774	406.4867
Observations	68	175	118	46
$\boldsymbol{\beta_0}$	1139.3685	227.55426	250.31857	-466.4913
$\boldsymbol{\beta_1}$	0.5072784	63.13861	148.22017	423.91478
$\boldsymbol{\beta_2}$	0.4972338	-0.57065	-0.54751	-0.856424

p values are all significant at α 0.05.

To identify if there were any correlations between these three regions, the study did not emphasis on how to classify and subset the values of $\log(d_2) or \log(D_2)$. The values of $\log(d_2) or \log(D_2)$ has a considerable impact on model fitting as well as the confidence on the prediction of the value of $d_2(D_2)$. Inappropriately grouping the values of $\log(d_2) or \log(D_2)$ might limit this application.

Emphasis on the differences between the coding and noncoding regions has been reported in various studies and has led to new perspectives in the understanding of DNA sequences. The *Trichomonas vaginalis* genome sequence study has found a higher G+C content and a lower frequency of repeated sequences in the coding regions when compared with the noncoding regions (Espinosa, et al., 2001). In a *Drosophila melanogaster* study, differences and similarities in composition of coding and noncoding sequences between the X chromosome and autosomes[7] were found (Singh, et al., 2005). The nonlinear model has revealed a significant relationship with the coding sequence and 5' UTR region and has complemented research that has already been investigated with these gene regions.

Future research is required to incorporate a wider range of organisms, along with other variables and biological functions to strengthen the understanding of this nonlinear trend, and to possibly associate it with evolutionary and biological phenomena. If the coding sequence and the 3' UTR sequence length are known, the 5' UTR length could be predicted, which could provide guidance in promoter studies (Bajic, et al., 2004; Burden, et al., 2005). The relationship between the length distributions of the coding and noncoding sequences is a thought-provoking question. Given the evidence of a nonlinear pattern with these regions, the next logical step would to incorporate other variables, such as protein function to determine the influence function has in relation to the coding and noncoding sequences.

5.5 Protein Function

The study of proteins and protein function is an important subject for biologists today. Proteins are the building blocks of all living organisms and play an important role in executing and regulating most biological processes. Sequence, structure and function are important components in the study of proteins, and the understanding of these components is now possible due to advanced techniques in sequencing.

Constraints on the evolution of proteins may be influenced by specific function, such as enzymes, regulators or signalling molecules (Lipman, et al., 2002). Examination of protein lengths in conjunction with functional classes, such as cellular processes and metabolism identified that the protein lengths of these functional groups were greater than those of some other groups (Brocchieri and Karlin, 2005).

Interest in genes that produce proteins of particular function has also been a growing and focused area. Coding and noncoding sequences are altered by the same mutational processes however, selection acts on these discriminately. Protein function adaptability can be contributed to many modifications in the sequence, including accumulation of sequence changes and gene duplications. Insertions and deletions (indels) within

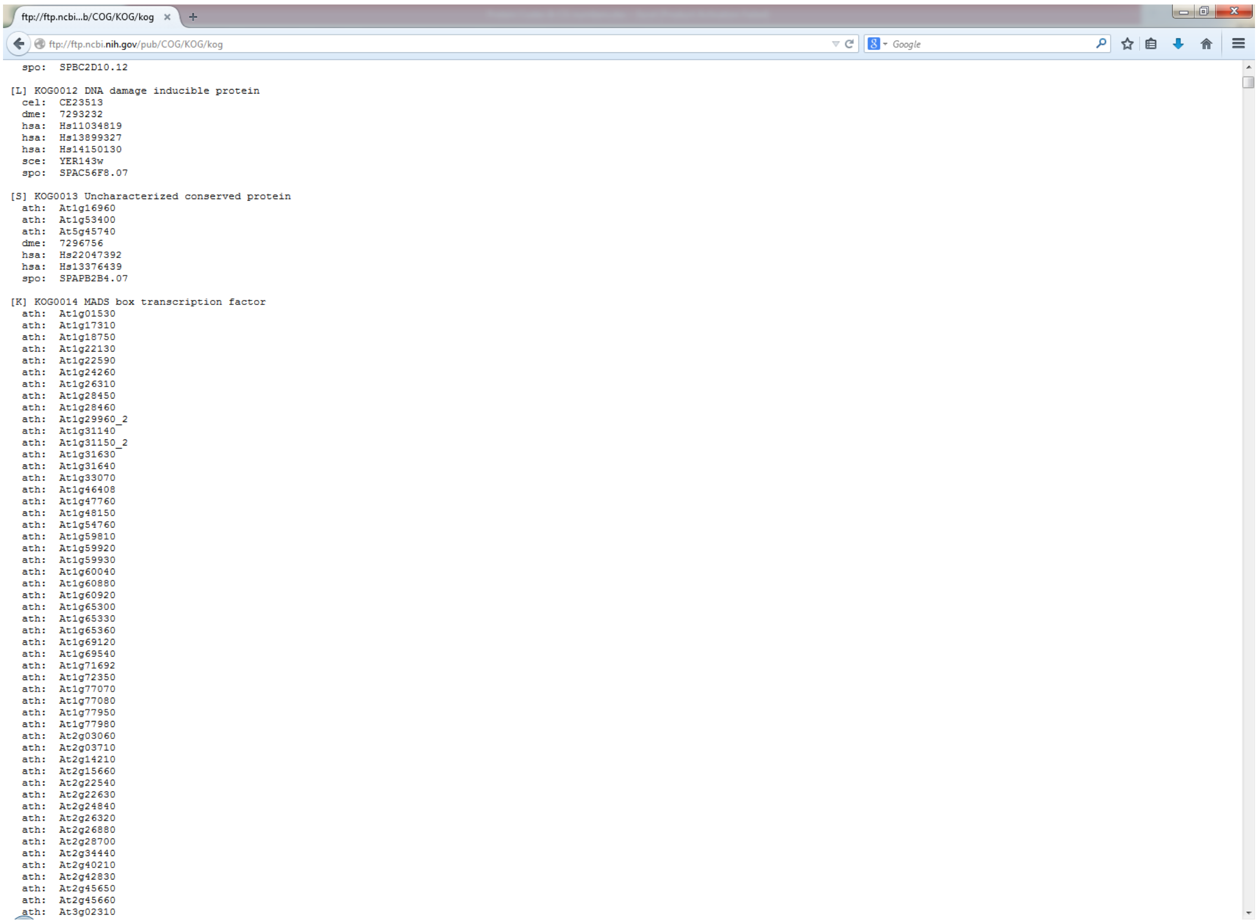
[7] Any chromosome that is not a sex chromosome

domains influence the length differences, with the presence of introns contributing to a larger expanse in protein length within eukaryotes than in prokaryotes. The size is also affected in eukaryotes by the accumulation of functional motifs that are involved in sophisticated regulatory networks (Zhang, 2000; Wang, 2005).

This section explores the noncoding and coding sequence length data of the *Arabidopsis thaliana* and *Drosophila melanogaster* to understand the relationship between the protein function and these lengths. The research extends previous investigation by examining not only the protein length data but the coding and un-translated region length data and will compliment what has already been found in previous chapters.

5.6 Functional Protein Classification

The length data for the coding and noncoding regions for each organism was merged with the Clusters of Orthologous Groups of proteins (COGs) database [http://www.ncbi.nlm.nih.gov/COG/] (Figure 5.7), using the IDs from each database. This database was generated by comparing predicted and known proteins in completed genomes of both microbial and eukaryotic organisms (Koonin, et al., 2004; Tatusov, et al., 1997). To investigate the sequence length in different protein functional groups the sequence length data was ranked into four main categories based on the COG functional classes (Table 5.3 / Appendix B).


ftp://ftp.ncbi...b/COG/KOG/kog
ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kog
Google

```
  spo:  SPBC2D10.12

[L] KOG0012 DNA damage inducible protein
  cel:  CE23513
  dme:  7293232
  hsa:  Hs11034819
  hsa:  Hs13899327
  hsa:  Hs14150130
  sce:  YER143w
  spo:  SPAC56F8.07

[S] KOG0013 Uncharacterized conserved protein
  ath:  At1g16960
  ath:  At1g53400
  ath:  At5g45740
  dme:  7296756
  hsa:  Hs22047392
  hsa:  Hs13376439
  spo:  SPAPB2B4.07

[K] KOG0014 MADS box transcription factor
  ath:  At1g01530
  ath:  At1g17310
  ath:  At1g18750
  ath:  At1g22130
  ath:  At1g22590
  ath:  At1g24260
  ath:  At1g26310
  ath:  At1g28450
  ath:  At1g28460
  ath:  At1g29960_2
  ath:  At1g31140
  ath:  At1g31150_2
  ath:  At1g31630
  ath:  At1g31640
  ath:  At1g33070
  ath:  At1g46408
  ath:  At1g47760
  ath:  At1g48150
  ath:  At1g54760
  ath:  At1g59810
  ath:  At1g59920
  ath:  At1g59930
  ath:  At1g60040
  ath:  At1g60880
  ath:  At1g60920
  ath:  At1g65300
  ath:  At1g65330
  ath:  At1g65360
  ath:  At1g69120
  ath:  At1g69540
  ath:  At1g71692
  ath:  At1g72350
  ath:  At1g77070
  ath:  At1g77080
  ath:  At1g77950
  ath:  At1g77980
  ath:  At2g03060
  ath:  At2g03710
  ath:  At2g14210
  ath:  At2g15660
  ath:  At2g22540
  ath:  At2g22630
  ath:  At2g24840
  ath:  At2g26320
  ath:  At2g26880
  ath:  At2g28700
  ath:  At2g34440
  ath:  At2g40210
  ath:  At2g42830
  ath:  At2g45650
  ath:  At2g45660
  ath:  At3g02310
```

Figure 5.7COG database FTP file format

Table 5.3 COG Functional Protein Classification [http://www.ncbi.nlm.nih.gov.COG/]. The classification was divided into 4 main categories 1) Information storage and processing; 2) Cellular processes and signalling; 3) Metabolism; and 4) Poorly Characterised.

(1) Information storage and processing	
J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
(2) Cellular processes and signaling	
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
Z	Cytoskeleton
W	Extracellular structures
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
(3) Metabolism	
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
(4) Poorly Characterized	
R	General function prediction only
S	Function unknown

5.7 Protein Function in relation to Coding and Noncoding Sequence Lengths

ANOVA analysis (analysis of variance) (Daniel, 1999) was performed on *Arabidopsis thaliana* (N = 13,245) and *Drosophila melanogaster* (N = 2,735) to compare differences between the coding and noncoding length sequences and protein function. The IBM SPSS19.0 software package (IBM, 2010) was used to conduct the analysis. ANOVA is a method of statistical hypothesis testing which reduces the rate of Type I errors (false positives) and is commonly used in analysis of experimental data (Ding, et al., 2014; Magwire, et al., 2010). ANOVA was used to test for differences between each length region (coding and noncoding sequences) within each protein functional group.

ANOVA analysis was conducted on each gene region, with and without introns, of both organisms, and significant differences were found in relation to the protein function categories. For the *Drosophila*, when the means were compared using ANOVA for each gene region (with and without introns) there were significant differences (*p*-value < 0.001) found between information storage and processing; cellular processes and signalling; metabolism; and poorly characterised protein categories (Figure 5.9A & 5.9B). In contrast, the *Arabidopsis* (Figure 5.8A & 5.8B) showed significant differences with each protein category with the exception of D2 and D3 mean difference. The only change between the two organisms was sample size with the *Arabidopsis* having a larger sample, which may have been more sensitive to the statistical testing.

Previous studies focusing on protein length found that the median values for the categories cellular process and metabolism are longest in all three phylogenetic domains (Eukarya, Bacteria and Archaea) (Brocchieri and Karlin, 2005). This is consistent with *Arabidopsis* data in this research, but *Drosophila* showed slight variation to this finding. When the noncoding regions were taken into consideration, similar length differences within the protein category groups were found at a smaller scale to the coding sequence length.

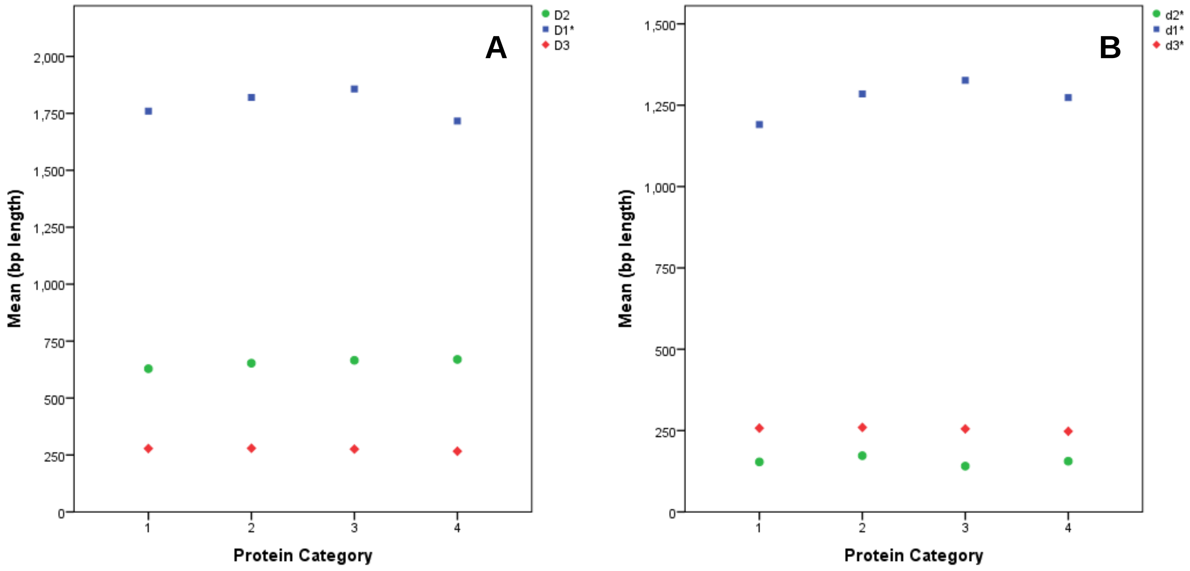


Figure 5.8 Protein category classifications in relation to noncoding and coding gene regions for Arabidopsis thaliana.

*A represents mean values with introns; B represents mean values without introns. Protein categories consist of 1: Information storage and processing; 2: Cellular processes and signalling; 3: Metabolism; & 4: Poorly characterized. ANOVA analysis conducted between each protein category and each region (with and without introns) found significant differences denoted by * (Alpha = 0.05). Sample size = N=13,245.*

Table 5.4 ANOVA analysis on Arabidopsis thaliana *between the length of the coding (with and without introns - D1 / d1) and noncoding (with and without introns - D2, D3 / d2, d3) gene regions in relation to protein function. Protein function was divided into four (4) main categories based on the COG classification.*

		Sum of Squares	**df**	**Mean Square**	**F Statistic**	**Sig.**
d2	Between Protein Category	1910306.459	3	636768.820	26.749	0.000*
	Within Groups	315202030.608	13241	23805.002		
	Total	317112337.067	13244			
D2	Between Protein Category	2941278.899	3	980426.300	0.936	0.422
	Within Groups	13863397428.676	13241	1047005.319		
	Total	13866338707.574	13244			
d1	Between Protein Category	28561720.276	3	9520573.425	18.119	0.000*
	Within Groups	6957615434.305	13241	525459.968		
	Total	6986177154.581	13244			
D1	Between Protein Category	38448351.415	3	12816117.138	6.151	0.000*
	Within Groups	27586772771.235	13241	2083435.750		
	Total	27625221122.650	13244			
d3	Between Protein Category	272403.177	3	90801.059	3.702	0.011*
	Within Groups	324775399.172	13241	24528.011		
	Total	325047802.349	13244			
D3	Between Protein Category	358683.203	3	119561.068	2.553	0.054
	Within Groups	619996565.352	13241	46823.999		
	Total	620355248.555	13244			

* Significant at α 0.05

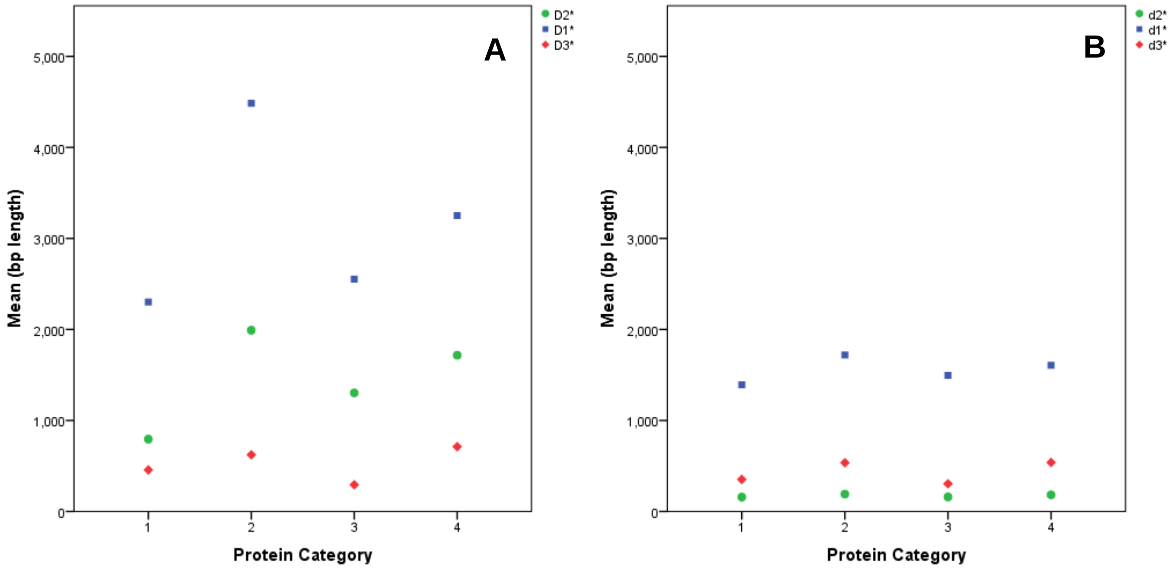


Figure 5.9 Protein category classifications in relation to noncoding and coding gene regions for Drosophila *melanogaster.*

*A represents mean values with introns; B represents mean values without introns. Protein categories consist of 1: Information storage and processing; 2: Cellular processes and signalling; 3: Metabolism; & 4: Poorly characterized. ANOVA analysis conducted between each protein category and each region (with and without introns) found significant differences denoted by * (Alpha = 0.05). Sample size N = 2,735.*

Table 5.5 ANOVA analysis on Drosophila melanogaster *between the length of the coding (D1 / d1) and noncoding (D2, D3 / d2, d3) gene regions in relation to protein function. Protein function was divided into four (4) main categories based on the COG classification.*

		Sum of Squares	df	Mean Square	F	Sig.
d2	Between Protein Category	574474.646	3	191491.549	5.723	.001*
	Within Groups	91376814.519	2731	33459.105		
	Total	91951289.165	2734			
D2	Between Protein Category	504995454.061	3	168331818.020	5.475	.001*
	Within Groups	71482359359.490	2325	30745100.800		
	Total	71987354813.551	2328			
d1	Between Protein Category	44623324.527	3	14874441.509	11.270	.000*
	Within Groups	3604348579.450	2731	1319790.765		
	Total	3648971903.976	2734			
D1	Between Protein Category	4281495213.351	3	1427165071.117	26.143	.000*
	Within Groups	149087313441.123	2731	54590740.916		
	Total	153368808654.474	2734			
d3	Between Protein Category	28885487.547	3	9628495.849	47.120	.000*
	Within Groups	558047821.403	2731	204338.272		
	Total	586933308.950	2734			
D3	Between Protein Category	52216955.197	3	17405651.732	20.971	.000*
	Within Groups	1929724322.131	2325	829988.956		
	Total	1981941277.327	2328			

* Significant at α 0.05

ANOVA analysis was used to compare the differences between the protein category groups in regard to the length of the coding and noncoding sequences. However, due to the nature of the data, Kruskal-Wallis testing was performed to support the ANOVA results. Interesting, the Kruskal-Wallis testing showed statistical differences between the categories of the protein groups (*P* = 0.000) and the length of the coding and noncoding sequences in both organisms.

An interesting picture emerges when comparing the two noncoding regions (5’ UTR and 3’ UTR) for both organisms (Figures 5.8 & 5.9). The mean length of 5’ UTR with introns (D2) is higher than the mean length of the 3’ UTR with introns (D3) whilst for the lengths without introns, the 3’ UTR (d3) is longer than the 5’ UTR (d2). These figures indicate the larger portion of introns in the 5’ UTR. Percentages of UTRs containing introns were estimated by Mignone et al (2002) and range from 15-35% for 5’ UTRs to 2-11% for 3’ UTRs. This has been confirmed in a recent study more accurately investigating the abundance, distribution and intron size within un-translated regions of genes in certain species (Hong, et al., 2006). The occupancy of introns in 5’ UTRs of *Arabidopsis thaliana* (2,012 numbers of introns) is lower than in the coding sequence (55,510), and with the 3’ UTR (382), it contained even smaller amounts of introns than that of the 5’ UTR. This is also true in the *Drosophila melanogaster* data which indicated the number of introns for the *Drosophila* was 1,490 for 5’ UTR; 10,507 for CDS; and 63 for 3’ UTR.

The experimental classification and function of genes, on a genome-wide scale is still in its early stages of development and determining which method of classification performs better have been yet to be achieved (Mi, et al., 2003). However, this research, even with small sample sizes has found significant differences between the available protein function classifications. The length distribution of genes and correlation between their regions in conjunction with protein function may reflect evolutionary trends among diverse organisms.

In delving into the patterns of statistical properties of different gene regions and their correlation it is intended to understand the spatial organization rules between various gene functional elements and the difference in such organizations among different living organisms and gene families. It is assumed that these rules and differences are the results of natural selection and reflect the complexity differences in the regulation of gene expression.

Again, the results from both organisms show very interesting results and guides the thesis project to start exploring the relevance length has on gene expression, as this is the most important process in all living organisms and was topical at the time in the literature.

Chapter 6 - Coding and Noncoding Sequences In Relation To Gene Expression

6 Coding and Noncoding Sequences In Relation To Gene Expression – *Arabidopsis thaliana Case Study*

This chapter is slightly modified from the paper:

Caldwell, R., Kongcharoen, J., Lin, Y., and Zhang, R. The Length Distributions of Noncoding and Coding Sequences in Relation to Gene Expression: A Study on Arabidopsis thaliana, Proceedings of IEEE International Conference on Bioinformatics and Computational Biology, 2010, Las Vegas, USA.

6.1 Introduction

Past attempts on understanding the influence of gene length on gene expression has yielded conflicting results. Most research conducted to date has focused on protein length of several model organisms. The relationship among gene expression and gene length for *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *A. thaliana* was found to be negatively correlated (Akashi, 2001; Raghava and Han, 2005). However, other research has found that there is positive correlation between these two factors. Ren et al (2006) studied the rice and *Arabidopsis* plant species to determine genes which are least compact in respect to gene expression. This study found that the length of the coding sequence per gene is larger in highly expressed genes. The conclusion drawn from this study is that highly expressed genes contain higher number of introns and exons (Ren, et al., 2006). In contrast, Raghava & Han (2005) and Subramanian (2004) found that a significant negative correlation was shown between the expression and length of a gene (Raghava and Han, 2005; Subramanian and Kumar, 2004). Li et al (2007) established that highly expressed genes are “miniaturized” when considering protein length, protein domain number, and intron number (Li, 2007).

Little research has explored gene expression in relation to protein function and length. Zhu (2008) investigated the correlation of tissue specific human genes in relation to genomic structure, phyletic age, evolutionary rates and promoter architecture. These included housekeeping genes (HK), which are genes that are expressed in all tissue and cell types as well as tissue-specific (TS) genes. It was found that in general the TS genes, were expressed at lower levels than the HK genes, and were shorter in length (Zhu, 2008).

The aim of this chapter was to use the noncoding and coding sequence length data of the *Arabidopsis thaliana* to determine whether there is a correlation between gene length and expression level for each protein coding gene. There was also an expectation that a relationship in the protein function of each gene in comparison to the gene length and gene expression levels will be established following the results found in chapter 5.

Conventional statistics were used first to identify correlations between the length distributions and gene expression parameters, and to understand the mechanics of the data itself. More complex statistics was used to expand on the initial findings.

6.2 Conventional Statistical Analysis

Pearson's correlation was applied to the datasets for testing the degree of linear relationship between the variables, gene expression and the length of each gene region. Pearson's correlation can be formulated from:

$$\frac{r=\sum_{i=1}^{N} X_iY_i - N\overline{X}\overline{Y}}{\sqrt{(\sum X^2 - N\overline{X}^2)(\sum Y^2 - N\overline{Y}^2)}} \qquad (1)$$

Median and mean values were compared due to the skewness nature of data. Skewness and kurtosis were calculated and were used to measure the observations that were clustered around a central point, or to measure the asymmetry of the distribution.

Skewness formula used through SPSS (v19.0):

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3} \qquad (2)$$

$\bar{Y}$ is the sample mean, *s* is the sample standard deviation and *N* is the sample size of the dataset. When the skewness value is zero, the data is symmetrically distributed. Negative values represent data skewed to the left and positive values are skewed to the right.

Kurtosis formula used through SPSS (v19.0):

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4} - 3 \qquad (3)$$

Minus 3 (-3) was used in the formula to generate a statistic of zero if a normal distribution is present.

The skewness and kurtosis analysis on the datasets identified left and right skewness in the data, no data was normally distributed. Therefore, to compare means, the Kruskal-Wallis test was performed to compare three or more independent groups of sampled data, which makes no assumptions about the distribution of the data.

Kruskal-Wallis formula used through SPSS (v19.0):

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \qquad (4)$$

Where n_i is the number of observations in group i and $r_{ij}$is the rank (among all observations) of observation j from group i. N is the total number of observations across all groups. $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ and $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

6.3 Gene expression and length distributions of coding and noncoding sequences in *Arabidopsis thaliana*

To start the investigation we looked at *Arabidopsis thaliana* where we obtained average intensity values from 1,000s of array experiments under varying environmental and tissue specific samples run by the Arabidopsis Functional Genomics Consortium (AFGC). The Average intensity values were obtained from the TAIR website. The average intensity value represents a large range of conditions and tissue types, therefore interpretation of the data can at this stage only be generalized. The data used was raw data, and is classified as “big data”. It can be difficult to identify patterns in the statistics in the underlying data from using descriptive statistics.

6.3.1 *Arabidopsis thaliana* Gene Expression and Coding and Noncoding Sequences

As a starting point, and based on previous research conducted on a similar study on gene expression and protein length, we used the same breakdown for the length data on our length regions (Brocchieri and Karlin, 2005). The purpose to split the data into five length categories is to get more insight into statistical information from the underlying data. If I did not consider breaking down the data into subsets, the signal of some statistical information of the data would become too weak to be identified. For the coding sequence, the length data was split into 5 length categories (Tables 6.1 & 6.2). In the lowest category ≤ 100 there were no values obtained, as the coding sequence started at length values above 100. In both datasets with and without introns the smaller the bp length of the coding sequence, the large the average intensity (gene expression). We used Pearson's correlation to determine the statistical significance between the length of the coding and noncoding sequence and the average gene expression intensity. The correlation is measured between expression levels and gene length. Pearson correlation confirmed the observations with the correlation being negatively significant (-0.108 without introns; -0.087 with introns). We used SPSS to calculate the correlation and the software also reports the level of significance and *T-test* results. This has been confirmed by previous studies that identified this trend in protein length (Li, 2007; Raghava and Han, 2005; Subramanian and Kumar, 2004). Raghava (2005) found significant negative correlation in the expression levels and gene length for *Saccharomyces cerevisiae*, with an r value of -0.18. This is consistent with the results obtained in this study for the *Arabidopsis*. Further testing from Raghava (2005) on two additional datasets also revealed the same results.

It is important to note that the data studied in this thesis are different from the data studied by other researchers, however the purpose of comparing results with other studies is not to check the accuracy, but further confirm the results given by other researchers.

Table 6.1 Coding sequence of Arabidopsis thaliana *without introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.*

Length (bp) of Coding Sequence (d1)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	NA	NA	NA	NA	NA	NA	NA	NA	NA
(100, 250)	70	210	210	11746	8803	-0.289	0.873	-0.210	-0.574
(250, 500)	1478	407	417	9406	7786	-0.409	1.155	-0.804	1.620
(500, 1000)	5057	778	789	8162	5658	-0.260	1.464	-1.037	1.897
≥ 1001	10800	1619	1416	7720	5051	3.706	1.739	26.534	3.381

Pearson Correlation: r = -0.108* (*Correlation is significant at the 0.1 level (2-tailed))

Table 6.2 Coding sequence of Arabidopsis thaliana *with introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.*

Length (bp) of Coding Sequence (D1)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	NA	NA	NA	NA	NA	NA	NA	NA	NA
(100, 250)	16	241	249	23359	27961	-1.934	-1.565	3.296	0.932
(250, 500)	222	415	441	8150	5901	-0.753	1.771	-0.686	2.768
(500, 1000)	1651	813	834	9593	7760	-0.608	1.124	-0.398	1.130
≥ 1001	15516	2430	2077	7821	5187	2.893	1.666	18.796	3.043

Pearson Correlation: r = -0.087* (*Correlation is significant at the 0.1 level (2-tailed))

The data for the 5’ UTR lengths, with and without introns were also divided into five length categories (Tables 6.3 & 6.4). For the 5’ UTR data there were gene lengths in the smallest length range of ≤ 100. The tables show a similar trend as seen in the coding sequence data, with the smaller the length of the 5’ UTR the higher the average intensity (gene expression) values. When Pearson’s correlation was applied to this dataset, significant negative correlations were also observed (r = -0.045 without introns; r = -0.036 with introns). The dataset without introns exhibited a larger variation in gene expression from each length category than the dataset with introns. It was also observed that the length values from > 100 to ≤ 1000 did not vary considerably in the average intensity values.

Table 6.3 5’ Un-translated region (UTR) of Arabidopsis thaliana *without introns. Average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.*

Length (bp) of 5’ UTR (d2)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	6833	65	70	8689	5804	-0.677	1.522	-0.355	2.412
(100, 250)	7612	157	146	7548	5308	0.562	1.643	-0.806	2.889
(250, 500)	2445	332	314	7658	4910	0.852	1.368	-0.124	1.196
(500, 1000)	485	645	589	7657	4808	0.909	1.647	-0.216	2.994
≥ 1001	30	2312	2424	3881	3119	-0.152	4.509	-0.178	22.586

Pearson Correlation: r = -0.045* (*Correlation is significant at the 0.1 level (2-tailed))

Table 6.4 5' Un-translated region (UTR) of Arabidopsis thaliana *with introns. Average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.*

Length (bp) of 5' UTR (D2)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	6161	64	69	8777	5832	-0.648	1.522	-0.435	2.366
(100, 250)	5803	158	147	7712	5393	0.535	1.588	-0.831	2.598
(250, 500)	2851	349	338	7608	5057	0.459	1.451	-0.856	1.574
(500, 1000)	2145	686	668	7216	4984	-.419	1.637	-0.797	3.083
≥ 1001	445	1596	1258	7606	4754	2.206	1.464	5.028	2.433

Pearson Correlation: r = -0.036* (*Correlation is significant at the 0.1 level (2-tailed))

Taken together, these observations indicate that the 5' UTR and the coding sequence of the *Arabidopsis thaliana* may be subject to evolutionary constraints in the management of gene expression. A theory many have considered is that to reduce the cost of energy in gene expression, natural selection supports shorter proteins and shorter introns (Castillo-Davis, et al., 2002). This could undoubtedly be the circumstance, with large protein lengths impacting on the energy cost of biosynthesis, with shorter protein lengths contributing to higher efficiency in synthesis (Wang, 2005). However, Wang (2005) found that newly evolved or derived proteins are on average, significantly longer than the older proteins, and these larger sizes may have some influence on protein stability and function (Claverie, 2003).

Because the 5' UTR and coding sequences are essential components of the production of proteins, in any living organism, a worthy question to ask would be does selection act on these sequences of genes to amplify transcription and translation effectiveness? Urrutia (2003) agree that due to the small size of the length sequences, in their case, protein size in relation to gene expression that selection is acting on these genes to maximise transcription and translation efficiency, since these sequences influence gene expression (Urrutia, 2003).

Interestingly, when comparisons are make between the 5'UTR sequence and the coding sequence, the density of introns in these two sequences are very similar, however the 3' UTR sequences contains less introns. Moreover, the introns in the 5' UTR are on average longer than those in the coding and 3' UTR sequence, which has been found in previous research (Chung, 2006).

Finally, the data for the 3' UTR lengths, with and without introns were split into five length categories (Tables 6.5 & 6.6). The number of genes were concentrated between > 100 ≤ 1000 length values, with small sample sizes in the smallest and largest length values. The datasets for the 3' UTR showed a very unique result compared to the coding and 5' UTR. Instead of the small length values having a high average gene expression intensity value it was lower. And in reverse the higher the length of the 3' UTR the greater the gene expression intensity. The Pearson correlation r values were also positively significant (r = 0.105 without introns; 0.063 with introns).

Table 6.5 3' Un-translated region (UTR) of Arabidopsis thaliana without introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.

Length (bp) of 3' UTR (d3)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	541	65	74	6878	3663	-0.625	1.426	-0.831	0.774
(100, 250)	7949	192	197	7044	4738	-0.464	1.715	-0.739	3.113
(250, 500)	7614	335	318	8852	6145	0.747	1.503	-0.485	2.465
(500, 1000)	1161	631	599	9382	6239	1.010	1.508	0.633	2.262
≥ 1001	140	1327	1212	9782	7302	0.481	0.829	-1.304	-0.467

Pearson Correlation: r = 0.105* (*Correlation is significant at the 0.1 level (2-tailed))

Table 6.6 3' Un-translated region (UTR) of Arabidopsis thaliana with introns. The average length (bp) in comparison to average intensity (gene expression). Sample size N=17,405 split into 5 length regions.

Length (bp) of 3' UTR (D3)	Gene Number	Mean Length	Median Length	Mean of Average Intensity	Median Average Intensity	Skewness		Kurtosis	
						Length	Average Intensity	Length	Average Intensity
≤ 100	531	65	74	6878	3623	-0.600	1.417	-0.866	0.738
(100, 250)	7803	192	197	7054	4737	-0.462	1.705	-0.740	3.080
(250, 500)	7106	332	314	8839	6192	0.763	1.445	-0.461	2.205
(500, 1000)	1655	649	603	9393	6208	0.940	1.583	0.064	2.430
≥ 1001	310	1437	1220	7504	5445	2.665	1.427	13.061	1.303

Pearson Correlation: r = 0.063* (*Correlation is significant at the 0.1 level (2-tailed))

The scatter plot shows the gene expression intensity and the 3' UTR length (≤ 100 bp) (Figure 6.1), with most of the genes around the average of 65 bp in length, being concentrated around the low end of the gene expression values.

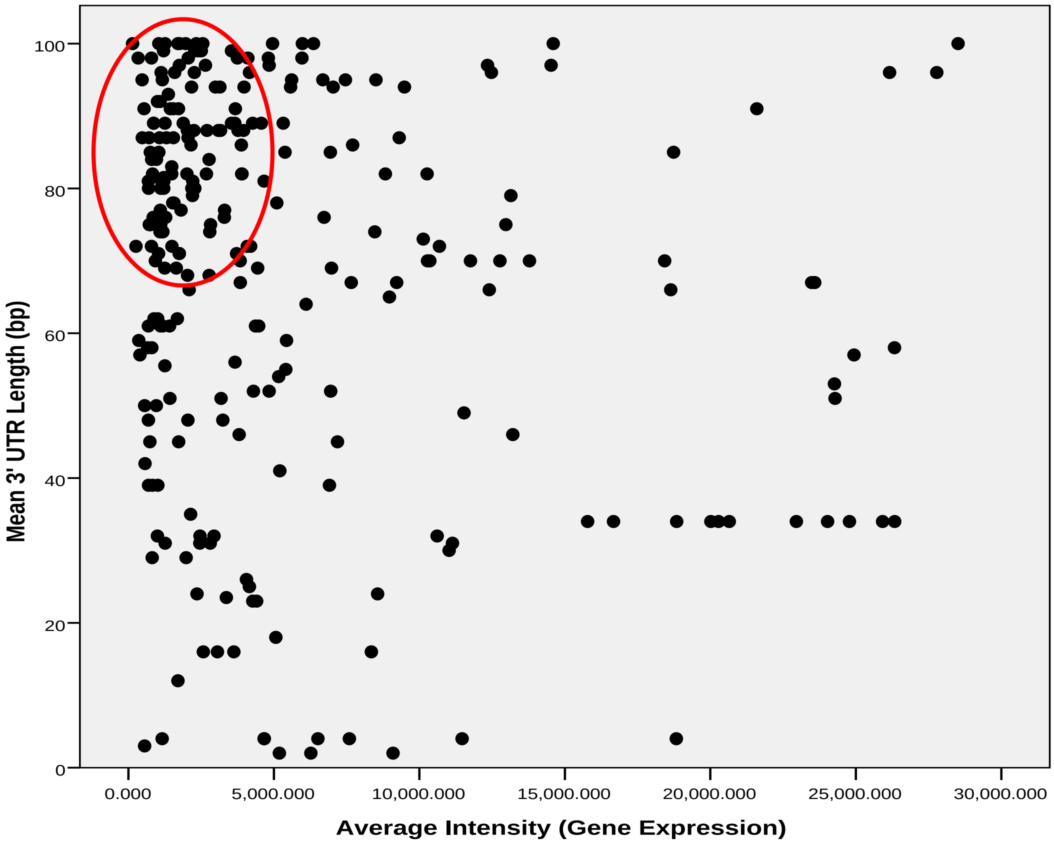


Figure 6.1 Scatter plot of the gene expression average intensity and the mean length of the 3'UTR region without introns. The 3' UTR length is categorised into ≤ 100 bp. Sample size of this category is N = 541 with an average length of 65 bp.

The 3' UTR gene regions lengths were opposite to that of the 5' UTR and coding sequence. This dataset showed a positive correlation between the 3' UTR length and gene expression intensity levels. A large amount of research has been accomplished on what function the poly(A) tail has in mRNA translation. From independent experiments performed over the last century it was established that the mRNA 3' poly(A) tail has a large influence on the initiation and stimulation of translation in eukaryotes (Preiss, 1998; Sachs, 1997). Therefore it would be reasonable to propose that the increase in the 3' UTR length may affect gene expression. Tanguay & Gallie (1996) concluded from experiments on the carrot protoplasts that there was an increase in stimulated expression by 24.5 fold when the 3' UTR was increased to 27 bases (Tanguay and Gallie, 1996). This would suggest that not only does the structural features and content of the 5' UTR sequence influence translational efficiency, but the 3' UTR length may also have some bearing on the stimulation of mRNA translation in eukaryotes, although in a differing capacity (Kuile, 2000).

Since the research had accessible data with and without introns, it would be erroneous to exclude a discussion on this aspect of the genes architecture, even though this is not the main focus of the research. It has been determined that introns play an important role in gene expression among many eukaryotic organisms. A great deal of energy is expensed in transcription with at least two ATP molecules used per nucleotide. Therefore, the presence of long introns for highly expressed genes can create a very high energy cost to the organism. A study conducted by Castillo-Davis *et al* (2002) found that, in general intron length varied among low gene expression levels however the average intron length in highly expressed genes were notably shorter (Castillo-Davis, et al., 2002). This was confirmed by this research in the *Arabidopsis* species (Figure 6.2), showing that intron length of < 100 bp had a higher proportion of gene expression than intron lengths > 100 bp. Comparisons of each length group found significance, the mean ranking of the gene expression are significantly different among the four intron length categories (H=64.8, 3 df, *p*=0.000). H represents the Kruskal-Wallis test, which is a non-parametric test and does not assume that the data comes from a distribution that can be completely described by two parameters. The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same.

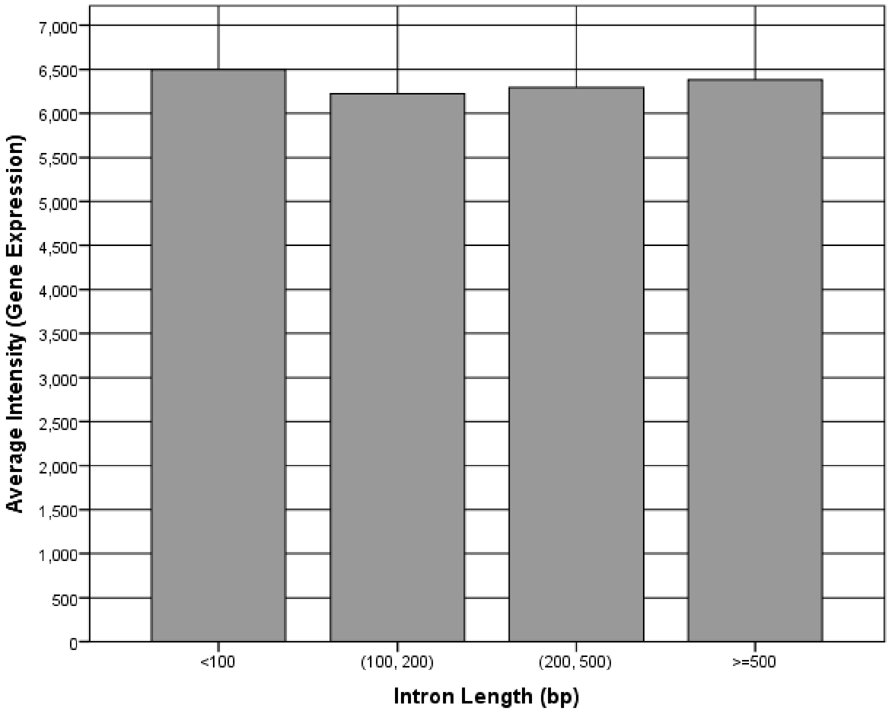


Figure 6.2 Total intron length of Arabidopsis thaliana*. The average length (bp) in comparison to average intensity (gene expression). Sample size N=123,854 split into 4 Intron length regions.*

6.4 Gene expression and protein function in relation to coding and noncoding sequences

The average intensity data tables were combined with the COG functional classification data tables for protein function and gene expression comparison analysis. The gene lengths were categorized based on the COG functional classification (refer to Chapter 5 for details). Poorly characterized gene categories were removed from the dataset for more concise analysis, represented by 21% of the data. The functions were grouped into 3 main categories, 1) information storage and processing; 2) cellular processes and signaling; and 3) Metabolism.

For each gene region, the length subsets were used for comparisons in each of the protein function classifications. The 5' UTR length region was subset into 5 length groupings (Figure 6.3). Each length subset for d_2 exhibited distinctive expression levels with the various protein function categories. The smaller d_2 lengths (≤ 100 bp) comprised functions for metabolism ([P] inorganic ion transport and metabolism) (Figure 6.3A), whereas in the subset length (100, 250) bp, the higher gene expression values spanned over metabolism and cellular processes and signaling functions ([G] carbohydrate transport and metabolism; [Z] cytoskeleton) (Figure 6.3B). As the median values lengths and gene expression levels increased, the functions were represented more in information storage and processing and cellular processes and signaling (Figure 6.3C-E). The trend seen in the 5' UTR data in all length regions is that metabolism was a frequent occurrence in the higher gene expression values.

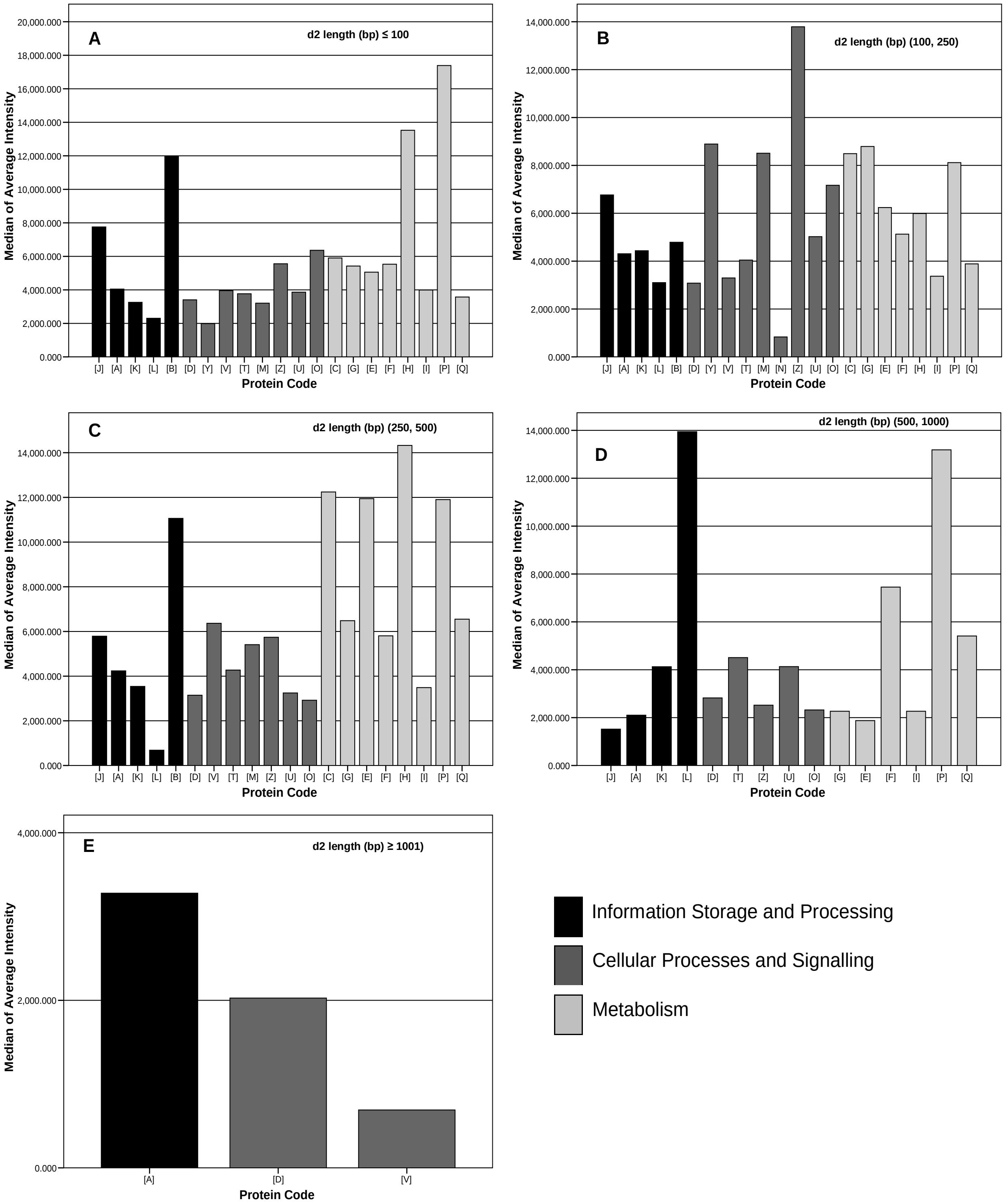


Figure 6.3 Arabidopsis thaliana *median length distributions for 5' UTR (d2) (without introns) and gene expression levels in comparison with protein function classifications. A) d2 length subset ≤ 100; B) d2 length subset (100,250); C) d2 length subset (250,500); D) d2 length subset (500, 1000); E) d2 length subset ≥ 1001. COG Protein classification groups [http://www.ncbi.nlm.nih.gov/COG/] were applied to the datasets.*

The smaller d_1 lengths ((100, 250) and (250, 500) bp) showed higher values in gene expression for the functions of information storage and processing and cellular processes and signaling (Figure 6.4A & 6.4B). As the length values increased the metabolism function was more prolific, in the high values of gene expression. For the coding sequence, d_1 lengths, there was no length category for ≤ 100 bp, the lengths were subset into four categories, as the length started at higher values.

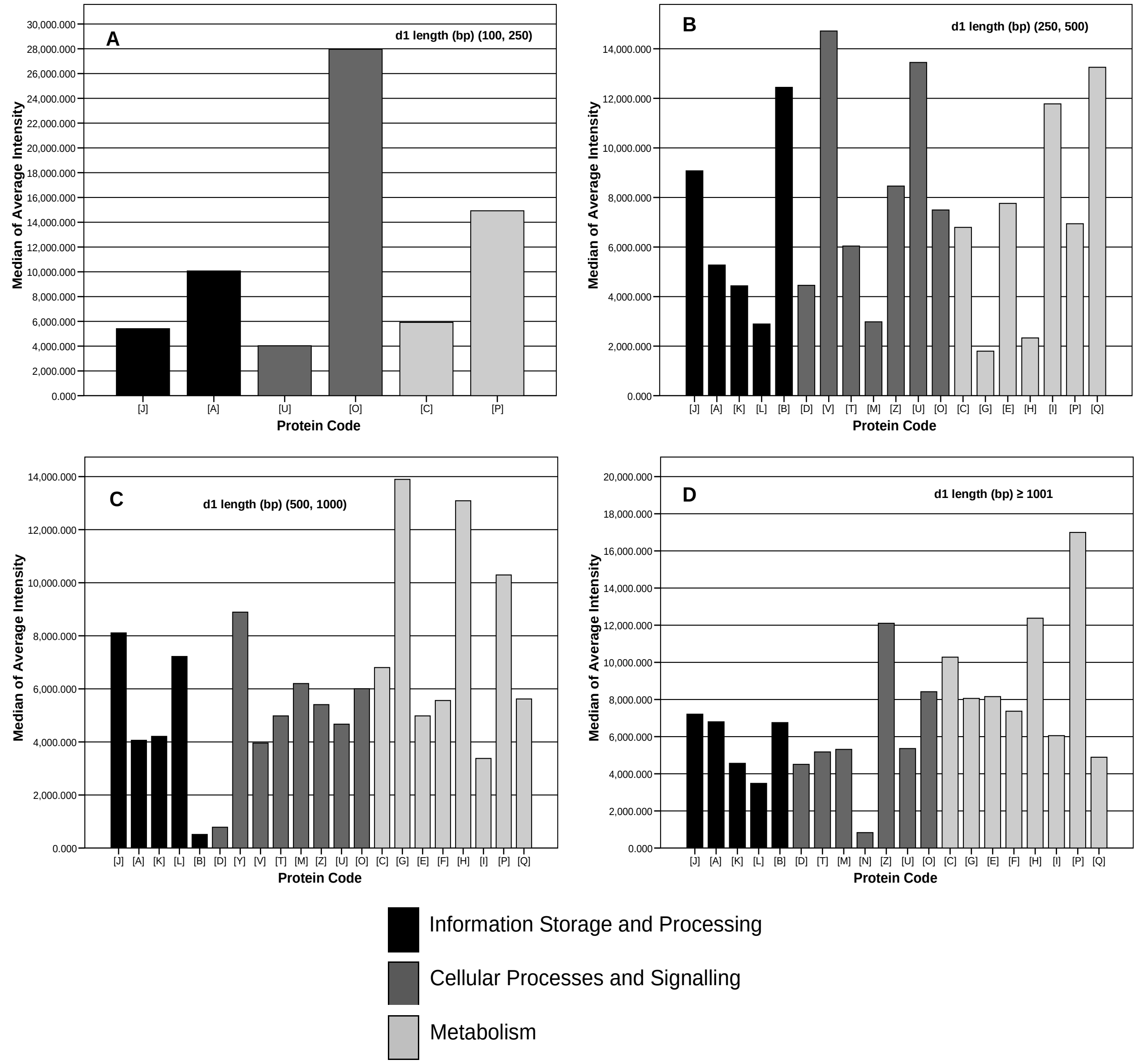


Figure 6.4 Arabidopsis thaliana *median length distributions for coding sequence (d1) (without introns) and gene expression levels in comparison with protein function classifications. A) d1 length subset (100, 250); B) d1 length subset (250, 500); C) d1 length subset (500, 1000); D) d1 length subset ≥ 1001. COG Protein classification groups [http://www.ncbi.nlm.nih.gov/COG/] were applied to the datasets.*

The d_3 lengths were also subset into five length groupings (Figure 6.5). Again the metabolism functional categories were observed over the range of d_3 length subsets. However, in the smallest length subset (≤ 100 bp) the functional categories that showed the highest gene expression levels were [A] RNA processing and modification, [B] Chromatin structure and dynamics (information storage and processing) (Figure 6.5A). This trend was also seen in the highest length subset (≥ 1001 bp) with RNA processing and modification showing the higher gene expression levels (Figure 6.5E).

For each gene region, coding and noncoding, in the upper length subsets, there was bias towards the metabolism functional categories. Overall, when looking at the whole dataset for the *Arabidopsis*, the highest gene expression obtained from the three functional categories was metabolism (Figure 6.6). Analysis to compare each functional category with the gene expression values found significant differences (H=408.9, 2 df, p=0.000).

The metabolism functional group presented higher gene expression levels than information storage and cellular processes, and fell above the average gene expression levels for all genes. In contrast, information storage and cellular processes fell below the average gene expression levels for all genes. These results are analogous to previous studies, where replication and transcription were below the average activity in all genes, however metabolism was found be around the average activity in all genes (Schmid, 2005). This would suggest that there are variations in functional classifications for gene expression in comparison to the average gene expression levels seen in all genes.

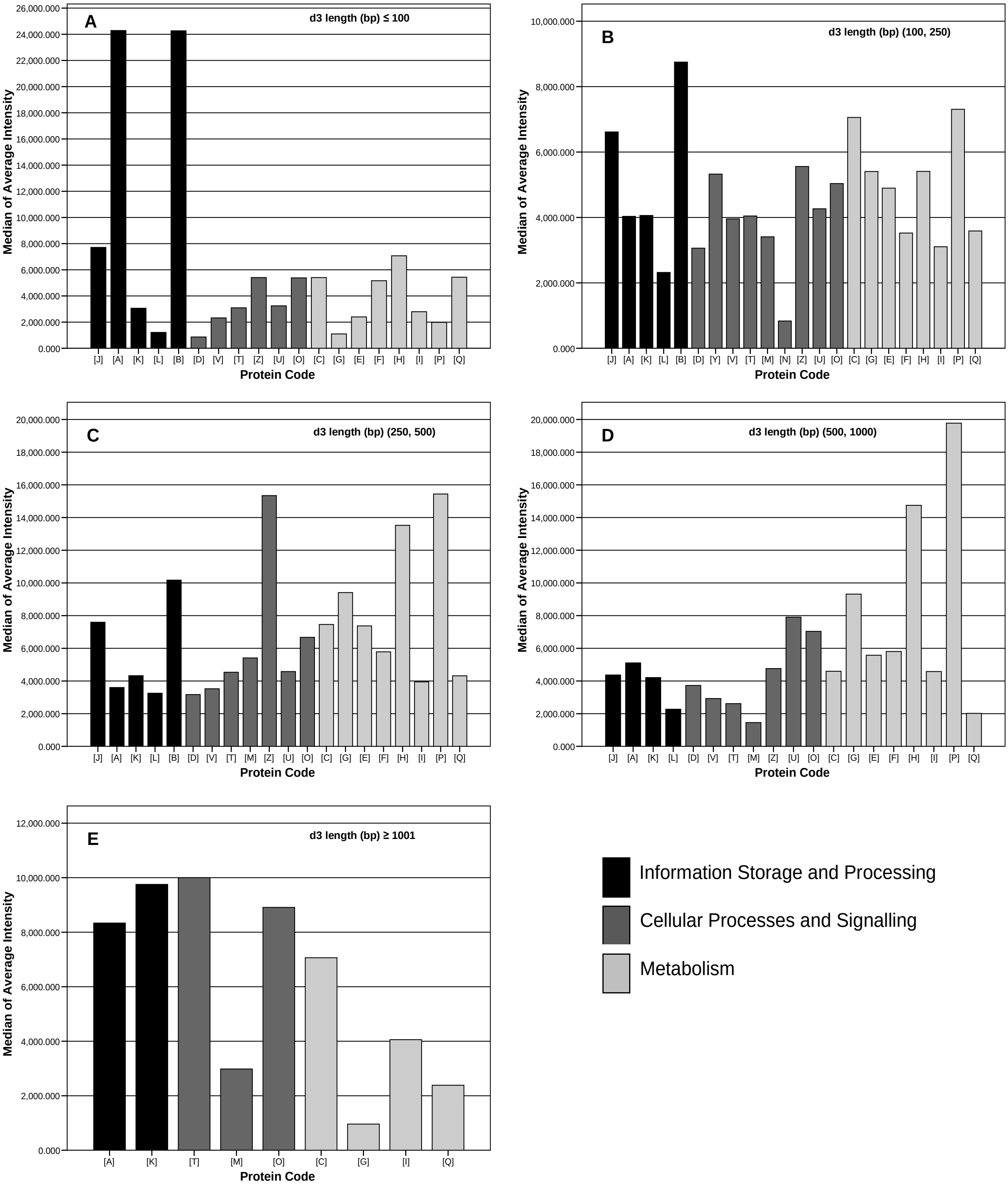


Figure 6.5 Arabidopsis thaliana *mean length distributions for 3' UTR (d3) (without introns) and gene expression levels in comparison with protein function classifications. A) d3 length subset ≤ 100; B) d3 length subset (100, 250); C) d3 length subset (250,500); D) d3 length subset (500, 1000); E) d3 length subset ≥ 1001. COG Protein classification*

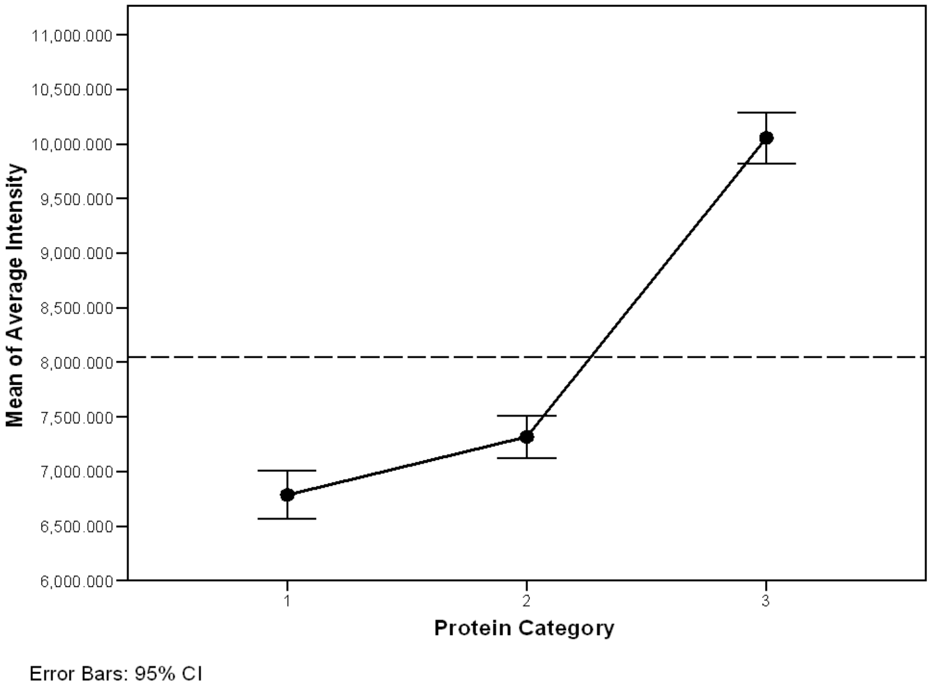


Figure 6.6 Arabidopsis thaliana *average intensity (gene expression) within 3 COG protein categories. Protein classifications as per COG functional classification is 1) Information storage and processing; 2) Cellular processes and signalling; 3) Metabolism. Sample size N=12,201, removal of poorly characterised proteins was applied to the dataset. - - - line indicates mean of all genes.*

Based on our results, there is concordance between what has been found in previous research and the data presented in this chapter. A summary of the statistics of the *Arabidopsis* genome completed by the Arabidopsis genome initiative (Initiative, 2000) found that 22.7% of the genes functional classes were cellular metabolism, followed by transcription (16.8%), based on a sample of 5,230. These results may explain why the sample of genes that were in these categories were high for our sample, metabolism being the most frequent function followed by transcription factors in all three regions. Furthermore, genes that are lowly expressed may only occur in a small range of tissue types, while highly expressed genes appear in the majority of tissues, making them easily distinguishable (Schmid, 2005).

Different regions of the *Arabidopsis* plant exhibit fluctuating gene expression levels. For example, the roots have higher relative expression levels than those in the apex and flower tissue samples (Schmid, 2005). In higher plants, these organisms have defense mechanisms in the form of pathogenesis-related (PR) proteins and genes to combat infections and damage (Kitajima and Sato, 1999). It has been established that mRNA of particular genes decreases when the plant has been exposed to wounds and pathogens (Liu and Mehdy, 2007).This could compromise the value of generalized interpretations of this gene expression data, which encompasses all tissues types and conditions.

The difficulty in interpreting the relationship between gene expression and protein function may be mired by the different conditions of each gene expression experiment and an inaccurate functional protein category database for global use, where it can be difficult to define a function across a wide variety of proteins and organisms (Gerstein and Jansen, 2000). It is difficult to surmise the patterns for gene function based on this study's gene expression dataset, due to the fact that the dataset covers a wide range of tissue types and conditions, however generalization of the data is possible, and a summary of the coding and noncoding length regions may offer some insight into particular patterns or relationships.

Other studies have focused on correlations between gene expression and protein interactions with varied outcomes. Weak correlation between gene expression and protein interactions may be rationalized by several hypotheses proposed by Bhardwaj & Lu (2005). Firstly, the correlation between gene expression and protein interactions are only weakly observed in yeast, therefore other species should be considered, secondly the expression data is too noisy to identify any relationship, and thirdly the correlation is weak in all species and the relationship is difficult to identify (Bhardwaj and Lu, 2005), which could also apply to gene expression and gene function. Classifying and analyzing the function of proteins is one of the most important activities biologists can achieve in the post-genomic era. Gene expression data in addition to the protein-protein interactions (PPI) data may be used to deduce functions of unknown genes, enhancing the gene ontology databases (Tu, et al., 2006).

Chapter 7 – Canonical Correlation Analysis (CCA) – *Drosophila melanogaster Case Study*

7 Canonical Correlation Analysis (CCA) - *Drosophila melanogaster Case Study*

This chapter is based on an unpublished paper.

7.1 Introduction

To try and improve on the standard statistics methods, to investigate the relationship between gene expression and the length of the coding and noncoding sequences, a variety of complex statistics were employed. Canonical Correlation Analysis (CCA) was chosen to extend previous research conducted in prior chapters. In this case study, *Drosophila* was chosen as there was a suitable sample of gene expression that incorporated environmental conditions, whereas for *Arabidopsis* gene expression analysis was conducted on a large sample of gene expression data containing all conditions and tissue types.

There are several circumstances in biological sciences where a researcher requires assessing a relationship between a set of dependent and a set of independent variables. Canonical Correlation Analysis (CCA) has been a useful statistical tool to examine patterns of interrelationships between sets of variables. Multivariate techniques in which CCA adopts, gives it a distinct advantage. Its benefits include reducing the need to run multiple comparisons, which not only saves time but can minimize Type I errors because it runs simultaneous comparisons in one test rather than over multiple statistical tests. It is more readily attainable due to the advent of statistical software, and can be more powerful under certain circumstances where other regression methods are lacking (Naylor, et al., 2010; Sherry and Henson, 2005). CCA has been effectively demonstrated in studies focusing on viral integration preferences (Gumus, et al., 2012), gene based tests in association with SNPs (Tang and Ferreira, 2012) and gene expression levels and genetic markers (Naylor, et al., 2010). Naylor et al found that CCA out-powered pairwise univariate regression models in their SNP Simulations.

Discovering genetic associations between the length distributions, of not only the coding sequence but the noncoding regions of a gene and gene expression levels under a number of environmental conditions has not been well illustrated. The research aims to employ the canonical correlation analysis method to attempt to establish correlations between gene expression and the length distributions of coding and noncoding sequences under various environmental conditions. Correlations between multiple datasets may expose some hidden biological phenomenon that may not be obvious with other statistical testing.

The sample size of the database was N=13,492. This database consisted of several replicates of genes that had various lengths within each of the gene regions. The sample size for individual genes was N=4,841. The length data was measured in base pairs (bp) and each of the sequence lengths included introns.

Please refer to chapter 3 for descriptions on length data collection for *Drosophila melanogaster*. Microarray Dataset: Gene expression data was collected via the GEO Datasets (NCBI) website http://www.ncbi.nlm.nih.gov/geo/ (GDS2830). The gene expression data consisted of *Drosophila melanogaster* females from three biological replicates from seven selection regimes and one control regime using whole genome gene expression arrays (SØRensen, et al., 2007). Replicated selection lines were selected for resistance to acute heat survival, high temperature knock down, constant 30°C during development, cold shock survival, desiccation, starvation, and longevity under non-stressful conditions. The raw CEL data from the microarray was transformed using R (v2.14.2) and the mas5 transform was applied to each of the replicates.

7.2 Canonical correlation analysis (CCA)

CCA is a useful technique that simultaneously tests the association between two sets of variables and can provide information concerning the nature of the links or patterns of interdependency that join the two sets, and also the number of (statistically significant) links between the sets. CCA can be considered as nothing more than a Pearson *r* test, however it is designed to maximize the correlation between the two canonical synthetic variables represented by the independent and dependent variables in each set.

The statistical problem entails identifying relationships between the length distributions of the coding and noncoding sequences and the gene expression intensity for each of the environmental conditions, with the goal of testing the strength of this relationship. The designation of the variables includes eight (8) metric-dependent and 3 metric-independent variables (Figure 7.1). Set 1 (X) composes the signal intensity of the gene expression across multiple environmental conditions and set 2 (Y) represents the length of each region of the gene including the coding and noncoding sequence length measured in base pairs.

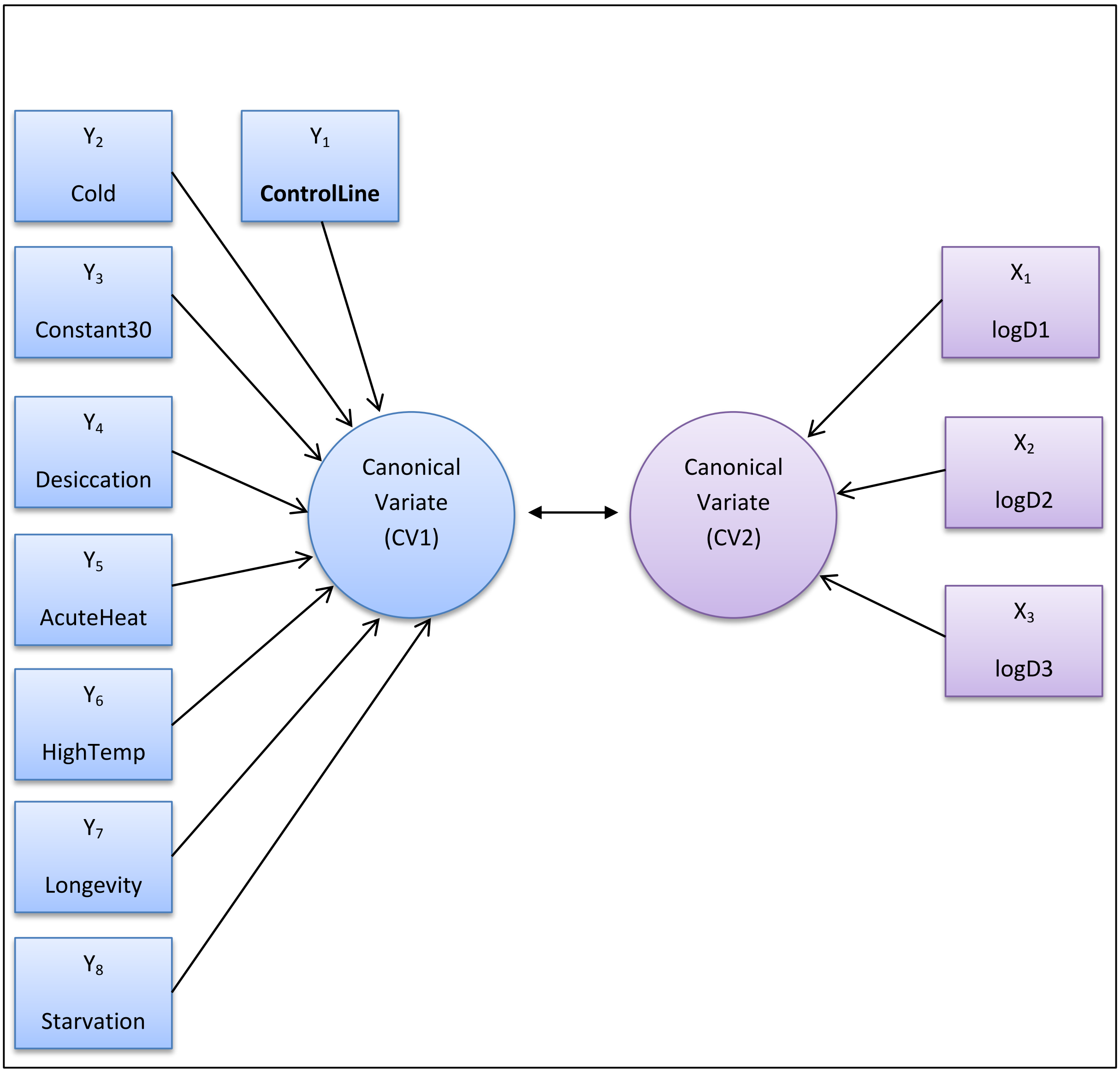


Figure 7.1 Representation of Canonical Variates – linear combinations of variables. Y values measured in gene expression signal intensity under a variety of environmental conditions, X values measured in base pairs.

The sample size of 13,492 was deemed too large for this test as it may affect the estimates of sampling error noticeably. Consequently replicates of each gene for varying lengths were averaged and only one value for each gene and each region length was recorded. The final sample size for testing was N=4,841, which is representative of the sample size for individual genes.

The independent variables were assessed for meeting the basic distributional assumptions and were found to be skewed. A log function was applied to the independent variables to pass this assumption. The dependent variables had a mas5 transform applied to set normality as is required with any microarray raw data.

The basic canonical correlation model is represented by two sets of variables X and Y (Figure 7.1). Each set is composed of variables, p variables in the X set and q variables in the Y set. Canonical correlation analysis was run through SPSS Statistics v19 using MANOVA and R CCA library (v2.14.2). To test the significance of the canonical correlations Wilks's lambda was used.

7.3 Pearson's Correlation

The most common correlation test used by biologists to measure correlation between two variables is Pearson's correlation. Pearson's correlation reflects the degree of linear relationship between two variables. The gene expression data was combined into one average value set over all experimental conditions. The data was used to analyse the correlation between the gene expression over all experimental conditions and the length of each gene region averaged over duplicate genes. Pearson's correlation (Table 7.1) showed significant positive correlations (*p* < 0.01) between gene expression under all experimental conditions in relation to the length distributions of the coding and noncoding sequences.

Table 7.1 Pearson's Correlation Analysis with Drosophila melanogaster *relating to the length distributions of coding and noncoding sequences to gene expression under all environmental conditions. Sample size of data N=4,841.*

bp	Mean	Pearson's Correlation
D1 – Length	10.076	0.038*
D2 – Length	6.802	0.160*
D3 – Length	7.444	0.140*

* Correlation is significant at the 0.01 level (2-tailed)

The gene expression data was then split into 3 groups using percentiles. Each group was labelled as Low, Medium and High gene expression (Figure 7.2).

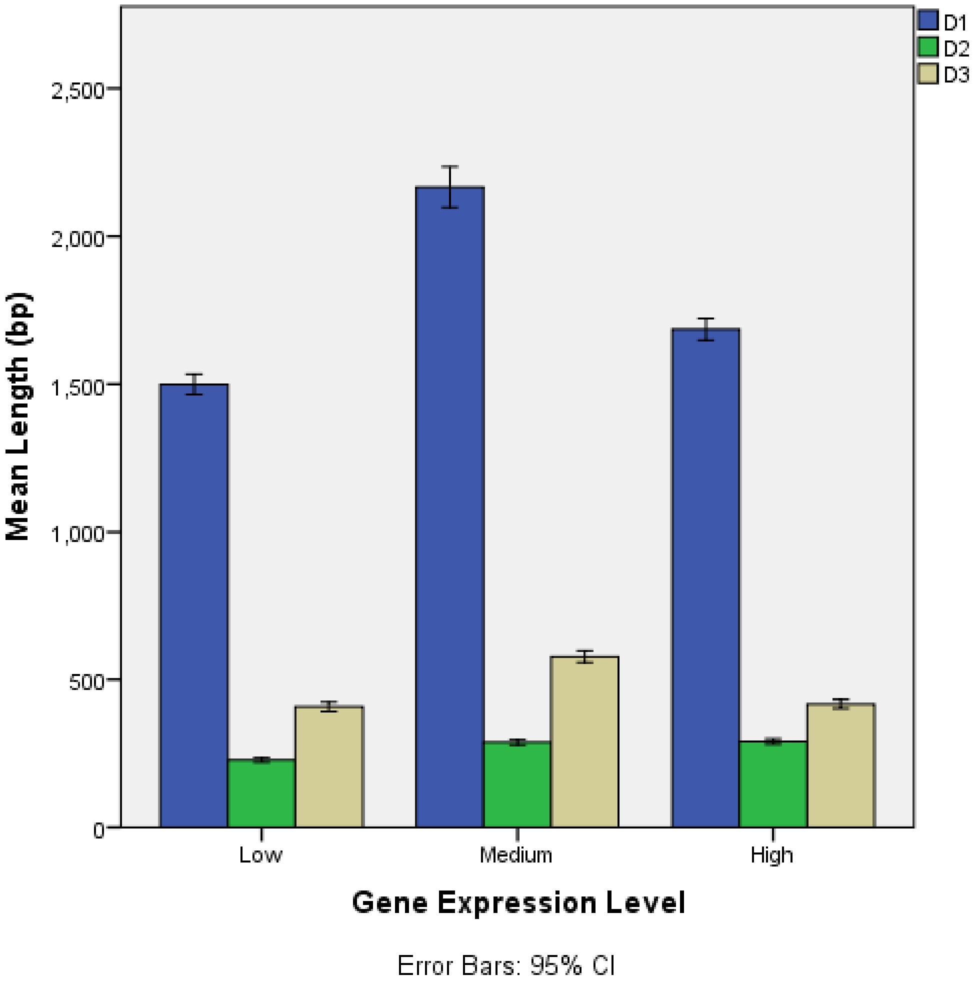


Figure 7.2 Mean length distributions (base pairs) of coding and noncoding sequences with Drosophila melanogaster over all experimental conditions split into low, medium, and high gene expression levels.

The length of each gene region shows different distribution over the three gene expression levels (Fig 7.2). D1 (coding sequence) shows a large drop in mean length beyond medium gene expression. D3 (3' un-translated sequence) displayed similar patterns to the D1 length. Notably D2 (5' un-translated sequence) shows a slight increase in length from low to high gene expression levels.

We utilized CCA to determine a relationship between the length of the coding and noncoding regions of a protein coding gene, and the gene expression levels of *Drosophila melanogaster* females subjected to various environmental conditions. The canonical correlation analysis was restricted to deriving three canonical functions because the independent variable set contained only three variables. To determine the number of canonical functions to include in the interpretation phase, the analysis focused on the level of statistical significance.

On examination of the canonical correlation values, the first two canonical correlation functions were considered noteworthy in the contexts of this research (Table 7.2). The Wilks' lambda statistic was employed which is a commonly used statistic to test for canonical correlation significance. The canonical functions 1 and 2 showed significance at α = 0.05. After scrutinizing the balance of variance over the two data sets, the first canonical correlation function was only reported. In the first canonical function, the canonical correlation (R_c) = 0.18421, indicating that approximately 3.4% of the variance is shared between the two variable sets and is represented by the gene expression category of longevity and the length region D2. The use of R^2 and R_c significance tests determine the canonical functions to interpret. The Wilks' Lambda significance determines the number of canonical functions used in the analysis.

Table 7.2 Canonical Correlation Analysis with Drosophila melanogaster *relating to the length distributions of coding and noncoding sequences to gene expression under various environmental conditions. * Represents significance at α 0.05*

Measures of Overall Model Fit for Canonical Correlation Analysis		
Canonical Function	Canonical Correlation	Canonical R^2
1	**0.18421**	**0.03393**
2	**0.07390**	**0.00546**
3	0.04273	0.00183

Multivariate Test of Significance			
Statistic	Value	F Statistic	Sig. of F
Wilks' Lambda			
1	0.95904	8.47913	0.000*
2	0.99272	2.52521	0.001*
3	0.99817	1.47344	0.183

For multiple x and y the canonical correlation analysis constructs two variates CVX1 and CVY1. The canonical weights are chosen so that they maximize the correlation between the canonical variates CVX1 and CVY1.

For the first canonical function, the four highest canonical loadings, which are correlations between variables and the canonical variates, from the eight variables in Set 1, were longevity (-0.991), Heat shock (-0.975), starvation (-0.972 and heat knockdown (-0.972) (Table 7.3) in the canonical variate 1/ set 1 (CV1-1). However all variables in Set 1 showed similar values using a cut off of 0.30 which has been the standard measure in other related studies. CV1-1 accounts for 94.4% of the variances in Set 1, while the other variate, CV2-1 shares 32% of its variance with Set 1.

Table 7.3 Canonical loadings for the First Canonical Function

Set 1	**Loading**		R_c = 0.184		**Set 2**	**Loading**
Cold	-0.971		→ **34%** ←		D1	-0.204
Constant 30	-0.970	CV1-1 (32%)		CV2-1 (16%)	D2	**-0.889**
ControlLine	-0.961	94.4%		47.5%	D3	-0.771
Desiccation	-09.61					
Heat	-0.975					
KnockDown	-0.972					
Longevity	**-0.991**					
Starvation	-0.972					

Of the three variables in Set 2, the two highest canonical loadings with a cut-off of 0.30 were D2 (-0.889) and D3 (-0.771) (Table 7.3) in CV2-1. CV2-1 accounts for 47.5% of the variances in Set 2, while the other variate, CV1-1 shares 16% of the variances with Set 2. All of the loadings for the canonical variate in Set 1 and Set 2 are negative, indicating the large values of the variables in Set 1 are associated with the large values of the variables in Set 2.

Regression analysis was performed and significance was found for each dependent variable only with the D2 and D3 length covariates ($p < 0.05$) which was also the relationship seen in the canonical loadings for function 1.

7.4 Discussion

The standard Pearson's correlation analysis on all experimental conditions showed positive correlations, indicating that the gene expression as a whole for the *Drosophila melanogaster* increases as the length distribution for each region increases. Furthermore when the gene expression data is segmented into low, medium and high expression levels the mean length changes over these expression subsets. There is an obvious decline in the length for D1 as the expression levels increase, indicating a negative relationship beyond the medium to high expression levels. D2 and D3 displayed similar trends. Other studies have found negative correlations associated with protein length and suggest the protein sequences and gene expression are subject to similar evolutionary dynamics (Duret and Mouchiroud, 1999; Lemos, et al., 2005). Previous research conducted by the author found negative correlations for the D1 and D2 lengths and positive correlations for D3 length in a model plant species (Caldwell, et al., 2010). The studies of gene expression changes associated with protein length, coding and noncoding sequences has helped to increase understanding of the fundamental connection between these biological processes and structures, however it creates many more questions.

Pearson's correlation offer a generalized view and understanding of the relationship between the length distributions and gene expression over combined environmental conditions and can be extremely time consuming if there are multiple variables to test. CCA was easily applied to the two data sets using statistical software to further analysis the intricate relationship between length of the coding and noncoding regions and gene expression under varying environmental conditions. Lemos et al emphasize the relevance of incorporating a number of biologically important variables to genome-wide relationships to understand the influence of protein and gene expression evolution.

The breakdown of the analysis showed two canonical correlation functions as being significant, and that for each dependent variable there was a weak relationship with D1. Importantly, the results show the maximized correlation for each data set for each variable, was between longevity (extended life span under non-stressful conditions) and D2. Both of these values were negative, indicating that the higher the expression levels of longevity, the longer the length of D2, 5' un-translated region. The notion that aging is somehow a result of a lifetime of stresses, may show age dependent expression changes among those genes that are regulated by stress (Golden and Melov, 2007). Sorensen et al methodology in the gene expression test for stress response was to apply cold shock, heat shock, heat knockdown, desiccation and starvation to flies, following this protocol longevity selection was measured. This could indicate the stress response prior to measuring longevity impacted on gene expression. However, there were very small differences between all experimental regimes and requires further investigation. Whole-genome analysis research on the *C. elegans* has supported the assessment that some changes in gene expression may play a role in specifying life span (Lund, et al.). Other relevant research to longevity and genome size has been shown in birds, where a highly significant relationship was seen (Monaghan and Metcalfe, 2001). What these findings represent and the mechanisms that influence it remain to be investigated.

Many studies have shown that 5' and 3' un-translated regions influence post-transcriptional regulation (Doran, 2008; Pesole, et al., 2000). Structural characteristics of 5' UTRs such as length have a high impact on the efficiency of the translational process. Intron presence and length has also been found to be a contributing factor in the enhanced expression levels among *Arabidopsis* (Chung, et al., 2006). There is also a higher occurrence of introns in the gene region corresponding to 5'UTR region, indicating shorter exons (Pesole, et al., 2000). This may explain our results as the length data of all the regions contained introns, and these characteristics for the D2 data may have been identified from the canonical correlation analysis. The length region association with longevity is an interesting outcome of this analysis and requires further exploration.

However, there are caveats in the application and interpretation of the results using CCA. Firstly this method has several assumptions as with all analyses. Adequate sample size is important to reduce the chances of Type II errors. In the preliminary testing CCA was applied to the N=13,492 dataset and the multivariate testing found significance with all canonical correlations, once the sample size was dropped to N=4,841 the R_c values improved and only the first two canonical correlations were significant. And secondly, CCA is used to test the linearity of relationships between variables, and may not be sensitive to nonlinear relationships as found in our previous research.

Can we say that this method outweighs standard statistical tests? This method is not commonly used by researchers in published papers, and the main reason for this is due to the complexity of interpreting the results (Thompson, 1980). However, if the technique is implemented correctly, and a good understanding of the results is produced the results offer the researcher a more complete view of the biological question. The method only saved a fraction of time in running the analysis, however, I did find it difficult to interpret and spent much more time deciphering the results into something meaningful. This led me to find another analysis tool to study the relationship between gene expression and the length of the coding and noncoding regions.

Chapter 8 - Genome Comparisons using Quantile regression Analysis Between Gene Expression and Length

8 Genome Comparisons using Quantile Regression Analysis between Gene Expression and Length

8.1 Introduction

This chapter is slightly modified from the following two papers:

Caldwell, R., Kongcharoen, J., Lin, Y., and Zhang, R. The Length Distributions of Noncoding and Coding Sequences in Relation to Gene Expression: A Study on Arabidopsis thaliana, Proceedings of IEEE International Conference on Bioinformatics and Computational Biology, 2010, Las Vegas, USA.

Caldwell, R., Lin, Y., and Zhang, R. (2015) Comparisons between Arabidopsis thaliana and Drosophila melanogaster in relation to Coding and Noncoding Sequence Length and Gene Expression, *International Journal of Genomics*, vol. 2015, Article ID 269127, 13 pages, 2015. doi:10.1155/2015/269127.

Regression analysis is a special case of Canonical Correlation Analysis, and therefore used in this chapter to investigate the relationship between gene expression and the coding and noncoding sequences. Statistical approaches, such as quantile regression, is a practical statistical method utilized by many biologists in a range of ecological (Cade and Noon 2003) and bioinformatics (Huang, Zhu et al. 2008; Wang and He 2008) studies to investigate relationships between variables. The advantage of using such a model includes the robustness against outliners, and helps obtain a more comprehensive analysis of the relationship between variables by using different measures of central tendency and statistical dispersion. When dealing with sequence length and gene expression data, modelling techniques often have difficulty with this data, due to the data values ranging over several orders of magnitude. It is general practice to log transform the data, particularly when parametric statistical tests, such as t-test, ANOVA or linear regression are used. The log function tends to squeeze together the larger values and stretches out the smaller values allowing a better view of the data.

8.2 Quantile regression Analysis on gene expression and length distributions in *Arabidopsis thaliana*

Another extension to the standard statistics methods was to apply quantile regression analysis on our gene expression and length data. After preliminary analysis of the length distributions and gene expression using standard Pearson’s correlation, quantile regression was used to extend the effect of gene length distribution on the average gene expression intensity. This type of analysis exposes the influence of independent variable(s) on a dependent variable in terms of variation range and conditional distribution status in greater depth (Chen and Ding, 2008).

Please refer to chapter 3 regarding the data collection of *Arabidopsis thaliana*.

Quantile regression models were used in this research to model average gene expression on the length of noncoding regions (3' UTR and 5' UTR's) and coding regions for *Arabidopsis thaliana* using the dataset without introns, which was more reliable and has been validated by the data community and many published research studies, to test the viability of the statistical method. To build up an appropriate quantile regression model for the average gene expression intensity and the length of coding region dataset, we started with the linear quantile regression model. Then we tested the quadratic, the cubic and higher order quantile regressions until an appropriate model was found.

Comparisons between different linear and nonlinear quantile regression models are based on model fit criteria Akaike Information Criterion (AIC) (Akaike, 1974) values of our final models to a number of alternative models of varying complexity levels at the same quantile. To assess whether the selection method resulted in an appropriate model, the AIC for quantile regression models are calculated as

$$\text{AIC} = n \times \ln(\text{SAF}(\tau)/n) + 2p$$

where n is the number of observations

SAF(τ) is the weighted sum of absolute deviations minimized when estimating the τ th regression quantile with p parameters

p is the number of parameters

The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. The smaller the AIC is the better the model. The following models were used to fit *Arabidopsis thaliana* dataset:

$$Q_{\text{int}}(\tau|d_1) = \beta_0(\tau) + \beta_1(\tau)d_1 + \beta_2(\tau)d_1^2 + \varepsilon(\tau) \quad (1)$$

$$Q_{\text{int}}(\tau|d_2) = \beta_0(\tau) + \beta_1(\tau)d_2 + \varepsilon(\tau) \quad (2)$$

$$Q_{\text{int}}(\tau|d_3) = \beta_0(\tau) + \beta_1(\tau)d_3 + \beta_2(\tau)d_3^2 + \varepsilon(\tau) \quad (3)$$

where $Q_{\text{int}}(\tau|d_1), Q_{\text{int}}(\tau|d_2)$ and $Q_{\text{int}}(\tau|d_3)$ are the τ^{th} quantile of the average gene expression intensity on the length of coding region, the length of 5' UTR region and the length of 3' UTR region covariates respectively. $\beta_i(\tau)_{;i=0,1,2}$ are unknown parameters in the model and need to be estimated: $\varepsilon(\tau)$ is the error term in the model ; $0 < \tau < 1$

Equations (1) and (3) are quadratic quantile regression models of the average gene expression intensity on the length of coding region and the length of 3'UTR region respectively. Equation (2) is a linear quantile regression model of the average gene expression intensity on the length of 5'UTR region.

Quantile regression was conducted on the length data for the coding region, 5'UTR region and 3'UTR region in relation to gene expression (8.1 a, b, c). The coefficients for d_1 and d_2 in models (1) and (2) are negative for all quantile cases, however the coefficients for d_3 (in model (3)) are positive. This indicates that the length of the coding region and the length of 5'UTR region (without introns) are negatively related to the quantiles of the average gene expression intensity while the length of 3'UTR region (without introns) are positively related. The patterns observed (Figure 8.1) shows the values of the quantile of the average gene expression intensity decreases as the value of d_1 (a) or d_2 (b) increases. However, as the value of d_3 (c) increases so does the value of the quantile of the average gene expression intensity increase only in the length range of 0 to 1000 bp. As d_3 increases after 1000 bp the quantile of the average gene expression intensity decreases. Therefore, the larger the quantile, the faster the quantile curve proceeds down, d_1 increases, while the quantile lines are steadier for d_2. After initial increases, the average gene expression intensity decreases as d_3 increases.

Our study using the average gene expression intensity data of *Arabidopsis thaliana* has verified previous research (Li, 2007; Raghava and Han, 2005; Subramanian and Kumar, 2004) that there is negative correlation between the length of the coding sequence (d_1) as well as the 5' un-translated region (d_2) and gene expression levels. Further analysis has also found that the 3' UTR showed a positive correlation. Previous research conducted by us found that there is a non-linear function relationship between the coding sequence length and the 5' UTR region (Caldwell, et al., 2008), and supports the fact that there is a nonlinear relationship in the *Arabidopsis* data in relation to gene expression. Using quantile regression modelling, to further test this correlation, it has confirmed the results, and is capable of aiding in the investigation of coding and noncoding length distributions on gene expression.

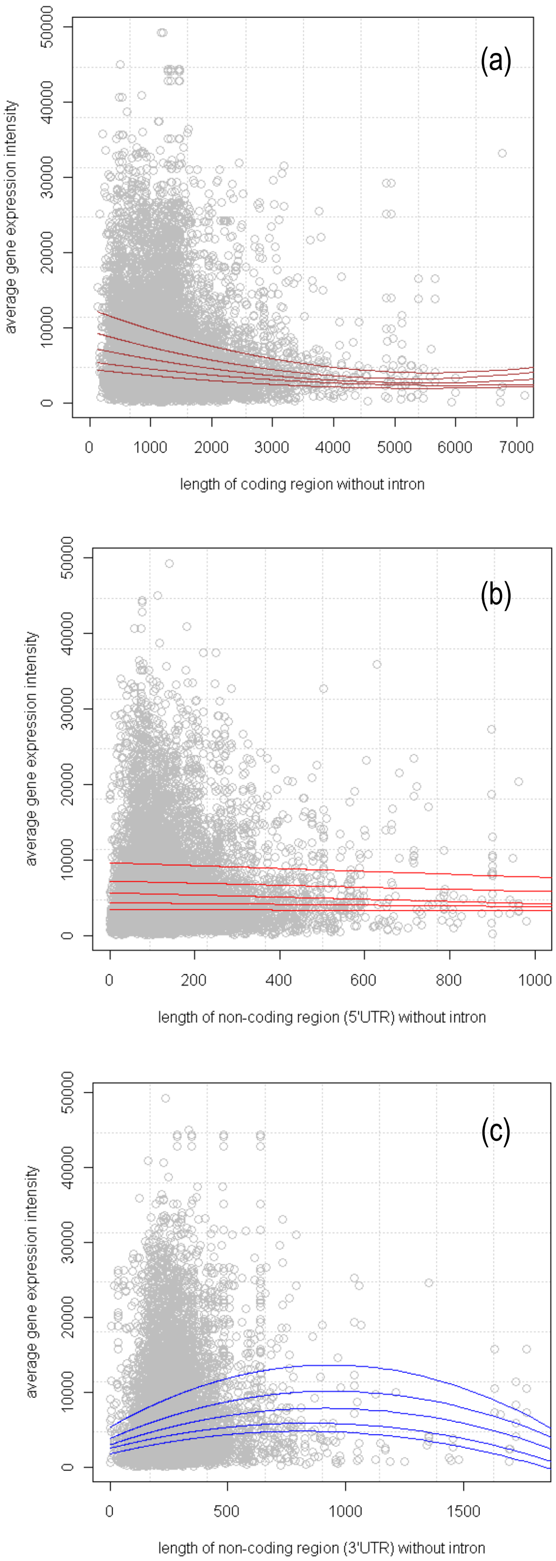


Figure 8.1 The quantile curves of the average gene expression intensity on the length of coding region (a); the length of 5'UTR region (b) and the length of 3'UTR region (c). The conditional quantiles include the range of 0.3 to 0.7 in quantile increments of 0.1 with Arabidopsis thaliana.

Negative correlations were found between the length of the 5' UTR and coding sequence and gene expression. The observations of the 5' UTR and coding sequence indicate that for *Arabidopsis thaliana* may be subject to evolutionary constraints in the management of gene expression. Longer 5' UTR regions in eukaryotes can produce defective proteins due to a higher instance of mutation to the translation-initiation codons (Lynch, et al., 2005). As discussed earlier, since the 5' UTR and coding sequences are essential components of the production of proteins in any living organism, it is reasonable to assume that selection act on these sequences of genes to amplify transcription and translation effectiveness. Urrutia & Hurst (Urrutia, 2003) postulate that due to the small length size of the sequences, in their case, protein size in relation to gene expression that selection is acting on these genes to maximise transcription and translation efficiency. However, a model proposed by Lynch (Lynch, et al., 2005), for the evolution of 5' UTR length suggests that the evolution of the length of this region is influenced by stochastic processes, rendering it selectively neutral. Reuter (Reuter, et al., 2008) disputed this model suggesting that UTR length evolution is affected by the gene's function and secondary mRNA structures. The length of the 5' UTR showed some influence in gene expression, to extend on this research further, gene function may indicate the evolutionary weight to changes in these lengths.

Our results on the 3' UTR gene regions lengths were reverse to that of the 5' UTR and coding sequence. They showed a positive correlation between the 3' UTR length and gene expression intensity levels. 3' UTRs have been related to the stability of mRNA processing, but it can be difficult to interpret due to the involvement of the mRNA in all processes. The importance of this un-translated region is evident in many studies examining the presence of 3' UTR in tumour growth (Briestanska and Plachy, 1996), 3' –processing end sequences on gene expression in plant cells (Ingelbrecht, et al., 1989), regulation of mouse κ Opioid receptor gene expression by different 3' Un-translated regions (Hu, et al., 2002).

Extension of the 3' UTR has also been allied with a pathway known as nonsense-mediated mRNA decay (NMD), where it was seen in *Saccharomyces cerevisiae* that 91% of the longer 3' UTR mRNAs tested were affected by NMD (Kebaara and Atkin, 2009). Mutually, the 5' and 3' UTR's involvement in gene expression is broadened to include further quality control mechanisms to strengthen the dependability of accurate protein formation (Chang, et al., 2007), and length is a contributing factor to these control mechanisms. The lengths of the 3' UTR's varies substantially within eukaryote genomes. Humans present longer 3' UTRs, compared to plants with a difference of 33% in length (Pesole, et al., 2000). The evolution of longer 3' UTR's, as seen in humans, may be contributed to the regulation of gene expression which use this increase in length for post-transcriptional control mechanisms (Hesketh, 2004). The results for the 3' UTR for this particular plant species, showing higher levels of gene expression indicates that there are evolutionary forces at work and the increased length plays a role in the regulation of gene expression. Tanguay & Gallie (Tanguay and Gallie, 1996) concluded from experiments on carrot protoplasts that there was an increase in stimulated expression by 24.5 fold when the 3' UTR was increased to 27 bases. The un-translated region influence gene expression by way of RNA stability and translational efficiency (Hesketh, 2004; Tanguay and Gallie, 1996) (3' UTR) and facilitating translation (5' UTR). Our results support the role and importance of these regions in the regulation of gene expression.

The results were interesting and not previously published regarding the positive correlation with the 3' UTR and gene expression. We are now interested in comparing an animal and plant species to determine if the same patterns are ascertained or dissimilar, in both the coding and noncoding regions, with an emphasis on the 3' UTR.

8.3 Genome Comparisons using Quantile regression Analysis on gene expression and length distributions

Advances in high-quality sequencing technologies (Franca, et al., 2002; Shapiro, et al., 2013), and large-scale resource data sets (Marygold, et al., 2013; SY, et al., 2003) have enhanced genomics research. Conducting large-scale sequence comparisons has the advantage of identifying the genetic variation and speciation among organisms (Ball and Cherry, 2001). Whole-genome expression experiments have also expanded a new era in bioinformatics analyses (Kilian, et al., 2007; Richards, et al., 2012; Robinson, et al., 2012; Sorensen, et al., 2005). Understanding relationships and cross-referencing of expression data to large genome data can now be attained and facilitates a greater insight of organismal complexity and the tightly regulated process of gene expression.

There is a continuing interest in the analysis of gene architecture and gene expression to determine the relationship that may exist (Murat, et al., 2012). Current investigations on the similarities and differences between plant and animal genome structure have led to a greater understanding in biochemical pathways, genetic mechanisms, sequence structures and functions (Kejnovsky, et al., 2009), and comparative studies are more powerful than studying the sequence of a single genome (Ball and Cherry, 2001). Furthermore, control of gene expression has been used as a measurement of variation and is often well conserved between species in the coding sequences. In unicellular organisms such as the yeast *Saccharomyces cerevisiae*, research has found that highly expressed genes tend to have smaller compact protein sizes (Warringer and Blomberg, 2006). Other animal genome studies have found that highly expressed genes have fewer and shorter introns, shorter coding sequences and protein lengths and favour more compactness in highly expressed genes (Rao, et al., 2010; Subramanian and Kumar, 2004). Previous research, however, is divided in opinion, with highly expressed genes not always being compact in plants. There is evidence that suggests in higher plant genomes, highly expressed genes comprise longer introns and primary transcripts (Ren, et al., 2006) in contrast, with other research on *Arabidopsis* and rice, finding that highly expressed genes are more compact (Yang, 2009), specifically the lengths of the coding sequence (CDS) (Caldwell, et al., 2010). Negative correlation between protein length and gene expression breadth in the plant species *Populus tremula* was also observed (Ingvarsson, 2007). Taken together, these observations suggest that the differences in length in relation to gene expression is not merely due to adaptive evolution, but rather has specific biological significance (Smith and Eyre-Walker, 2002).

Significance of noncoding regions is less understood across species compared to the coding regions. A range of genomic studies over the last decade has supported the opinion that there are tightly regulated processes and levels of control in the regulation of gene expression. This has included the untranslated gene regions, notably the 5' and 3' untranslated regions (UTRs) which may play the most important role in the regulation of gene expression (Andofatto, 2005). A study by Lin & Li (2012) revealed a strong negative correlation between the 5' UTR length and expression correlation with cytosolic ribosomal protein patterns in *S. cerevisiae* and *C. albicans* (Lin and Li, 2012), with highly expressed eukaryotic genes tending to have more compact 5' UTR regions (Grisdale and Fast, 2011). A plant study on both *Arabidopsis* and rice also reported negative correlation between expression levels and noncoding sequences (both 5' and 3' UTRs) (Yang, 2009).

The aim of this study was to apply a quantile regression model to re-examine the correlation of gene region lengths and expression levels of *Arabidopsis* using a different and larger set of gene expression data. The research also extended to another species, *Drosophila melanogaster*, so this study not only expanded objects but also conducted a comparison between a plant and animal species.

8.4 Methods

8.4.1 Datasets

Sequence and gene expression data were collected from a selection of publicly accessible databases and websites for each of the plant and animal species.

The *Arabidopsis thaliana* sequence data were downloaded from the TAIR website ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/. The sequence data used were generated from the TAIR10 (December 2010) release. Gene expression data were downloaded from the NCBI GEO Datasets database (series GSE34188) (Hanada, et al., 2013) including the annotation file which contained only one gene model for each gene. The downloaded expression data were already normalized by Bioconductor (www.bioconductor.org) R software. The final sample size for analysis was 18,445 genes, excluding two (2) genes from the coding sequence that only had 1 bp which was classified as an intron. The accession string and ID reference from the arrays were used to link the data together to create a master database of length and gene expression data for analysis.

The *Drosophila melanogaster* sequence data were downloaded from the Flybase website: http://www.flybase.com.au/. The raw CEL gene expression data files were downloaded from the NCBI GEO Datasets database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42255) under series GSE42255 (Landis, et al., 2012). Affymetrix microarrays were used to analyse the adult *Drosophila* and the raw CEL files were normalized using the Bioconductor (www.bioconductor.org) affy package in the R software environment. The annotation file was included and the Entrez UniGene name (GC numbers) and the ID from the platform data table was used to link the data together to create a master database of length and gene expression. The final sample size of unique genes was 3,290 for analysis.

The downloaded text files for each organism were cleaned using visual basic scripts and imported into MS Excel, all length data for both coding and noncoding sequences excludes introns. For each organism the gene expression experiments included multiple replicates of the control as well as abiotic stress conditions. For this study we have only reported on the control condition expression from the GEO datasets for both organisms, to simplify the analysis reporting. Abiotic stress conditions will be investigated at a later stage.

8.4.2 Statistical Analysis

Pearson's correlation was used to test the gene expression data to determine the reliability of the control replicates. The R^2 value was found never below 0.95, demonstrating the accuracy and reproducibility of the raw data. Therefore, the mean of the results of the control biological replicas were used in the statistical analysis reporting. The gene expression measurements are represented by gene expression signal intensity.

In this study we are interested in whether the length of the coding and noncoding sequence has a significant impact on the probability distribution of the gene expression under control conditions. Quantiles are statistics that describe the subdivisions of a ranked set of data values into equal proportions. Divisions can be made in four parts corresponding to 25%, 50% and 75% of the data. Firstly, to examine how the data behaves between the sequence length of each region, and gene expression, the length data for each region (5' UTR, CDS, and 3' UTR) were split into 4 quartiles (group 1, 2, 3, & 4).

Strong skewness was identified in all the length datasets for each gene region. For example, the distribution of the 5' UTR length without introns in *Arabidopsis thaliana* was positively skewed (skewness = 2.511) (Figure 8.2). Consequently, the Kruskal-Wallis nonparametric analysis method using SPSS version 19 (SPSS IBM, New York, U.S.A) was applied to the data to determine whether there are differences between the quartile groups, in relation to gene expression and the length of the coding and noncoding regions. This test makes no assumptions about the distribution of the data.

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{4}$$

Where n_i is the number of observations in group i and $r_{ij}$is the rank (among all observations) of observation j from group i. N is the total number of observations across all groups. $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ and $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

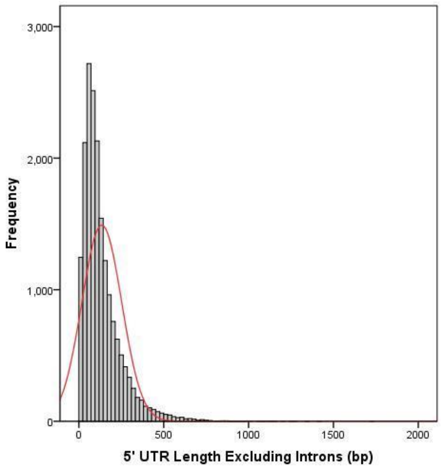


Figure 8.2 18,445 genes in Arabidopsis thaliana *for the 5' untranslated (UTR) region length, excluding introns. The distribution of this data is positively skewed (skewness = 2.511)*

8.4.3 Quantile Regression Analysis

The purpose of regression analysis is to expose the relationship between the independent variable (*x*) and dependent variables (*y*). Conditional quantile regression is useful in modelling the quantile value of the dependent variable on the independent variable. In this study, the dependent variable is represented by the log of gene expression, under control conditions, and the independent variable is represented by the log of the sequence length. The lengths considered include the coding and noncoding sequence (5' UTR, CDS, 3' UTR). The model considered was linear and is represented by:

$$\log Control = \beta_0 + \beta_1 \log x + \varepsilon \tag{5}$$

x represents the following attributes: Log 5' UTR, Log CDS, and Log 3' UTR.

The quantile subsets used ranged from 0.1 to 0.9 in 0.1 increments. The Log of the data was used to expand the data points for an enhanced view of the quantile regions. Regression analysis was performed in R. We used a linear model to be consistent with the analysis between organisms and to alleviate discrepancies in the analysis.

8.5 Results – Length Subset Analysis

To understand the relationship of the length of the coding and noncoding sequences and gene expression, the data of the lengths for each type of coding and noncoding region were grouped into four quartile subsets, respectively. For each quartile subset (1, 2, 3, 4), the gene expression data in each of these quartiles were averaged. Through the nonparametric analysis method, the mean of the gene expression conditional on the four quartile groups for each length region, respectively, were significantly different (*p*-value < 0.000) (Figure 8.3A & 8.4A).

For *Arabidopsis* the coding sequences shows a linear negative relationship between the four quartiles (groups 1 – 4) and their average gene expression intensity, indicating as the length increases, the gene expression intensity decreases. This pattern is also seen in *Drosophila* (Figure 8.4A). The same pattern is also seen in the full transcript length, which follows the same negative relationship, in both the animal and plant species.

However, the noncoding sequences, show dissimilar trends from the coding sequence. The relationship between the length of the 5’ UTR and gene expression intensity for *Arabidopsis* indicates a quadratic form, with an increase in length until the average gene expression intensity peaks for those genes in the 3rd group determined by the 3rd quartile, and then starts to decrease (Figure 8.3A).

The pattern seen in the 3’ UTR length data was more positively correlated in relation to the average gene expression intensity, in contrast to the CDS and 5’ UTR sequence length. This pattern implies that as the length of the 3’ UTR increases (from 1 to 3318 base pairs) the gene expression intensity increases (Figure 8.3A).

Furthermore, in *Drosophila*, the noncoding sequences in relation to the average gene expression intensity varied considerably from *Arabidopsis*. The patterns showed a reversal in the 3’ and 5’ UTR sequence length in relation to the average gene expression. The 3’ UTR gene expression intensity increased until the 2nd quartile and then decreased at the 4th quartile, again showing signs of a non-linear relationship. The pattern in the 5’ UTR for *Drosophila* was very distinctive, displaying a cubic polynomial pattern with one turning point, (Figure 8.4A).

In summary, the findings based on the 4 quartile subsets shows some variability between the coding and noncoding sequences as well as between animal and plant species. The quartile analysis indicates that the coding sequence is negatively correlated to the average gene expression intensity for both the animal and plant species. The full transcript sequence, which includes the flanking 5’ and 3’ UTRs also shows negative correlation to the average gene expression intensity again in both species. However, when the gene is divided into coding and noncoding regions, differing patterns emerge from each of these gene regions in the plant and animal species. It is important to note that these gene region lengths do not include introns, the gene expression values are measured under control conditions, and the gene length and gene expression data for this analysis has not been transformed.

To determine the validity of the findings in the first set of gene expression experiments, a second set of gene expression data was downloaded from the GEO Dataset website. The raw CEL gene expression files were downloaded GDS3933 (González-Pérez, et al., 2011) – *Arabidopsis* and GSE36507 – *Drosophila* and normalised by MAS5 using R. The label and hybridization protocols for *Arabidopsis* varied between each experimental sample, the first sample using Agilent Low RNA Input Linear Amplification Kit and the second sample using GeneChip® 3' IVT Express Kit. In both samples, total RNA was extracted.

For *Drosophila* both the gene expression samples used 7-9 day old adults, with total RNA extraction. The labels used were biotin however the protocols for labelling varied between the gene expression samples. Hybridization protocols followed similar methods. Length data and the master databases containing the length and gene expression data were generated with the same method as outlined in the methods section above.

The quartile results show similar results to the first set of gene expression analysis. The noncoding sequences (5' and 3' UTRs) in both the animal and plant species displayed an increase in the first two quartiles, then decreased. However, for the coding sequence there was not such a dramatic decline in gene expression from each quartile (Figure 8.3B & 8.4B).

To test the distribution of gene expression across the four quartile groups, nonparametric analysis was applied to the new gene expression samples. As seen in the previous example, the mean of the gene expression conditional on the four quartile groups for each length region, respectively, were significantly different (*p*-value <0.0000 at significance level 0.05) (Figure 8.3B & 8.4B).

For the experimental analyses with the quartile length subsets, it is difficult to achieve a general opinion on patterns observed in the coding and noncoding sequences in relation to gene expression. The data in the four subsets do not have sufficient resolution to determine accurately, identifiable patterns in both the animal and plant species. However, based on the nonparametric analysis both samples' results were unanimous in showing significant differences between the gene expression and the four quartile length groups. The results reported in the length subset analysis of this paper and the results on the relationship between gene expression intensity, and length in general, published in the literature, have directed us to employ a different analytical method to examine more precisely this relationship.

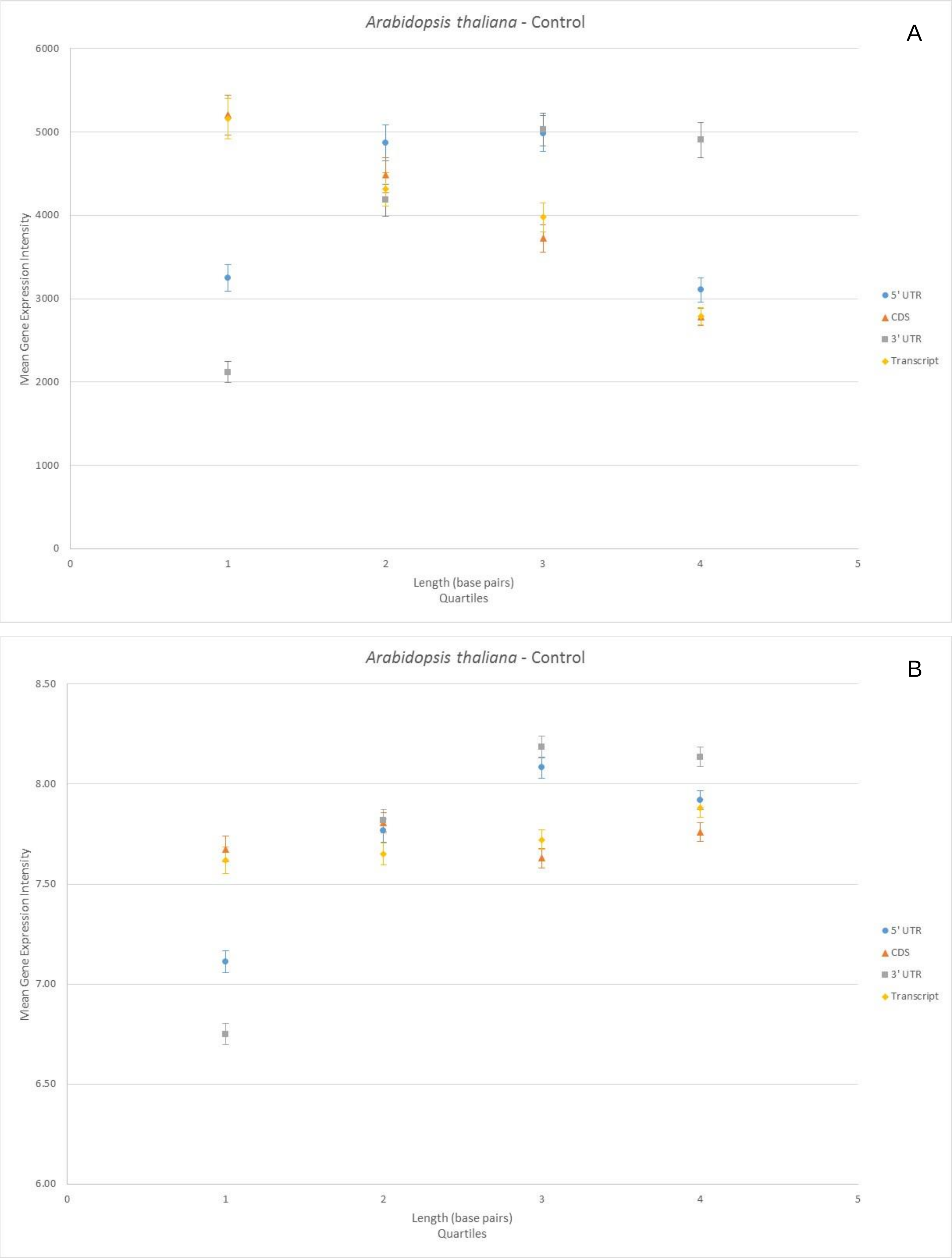


Figure 8.3 Relationship of gene expression in Arabidopsis thaliana within the coding and noncoding sequence regions. The gene expression intensity from GEO Datasets - GSE31488 (A) and GDS3933 (B) are plotted versus the quartile score for coding sequence, transcript, 5' UTR and 3' UTR regions. Each data point represents the mean for the samples in each quartile. Error bars represent standard error.

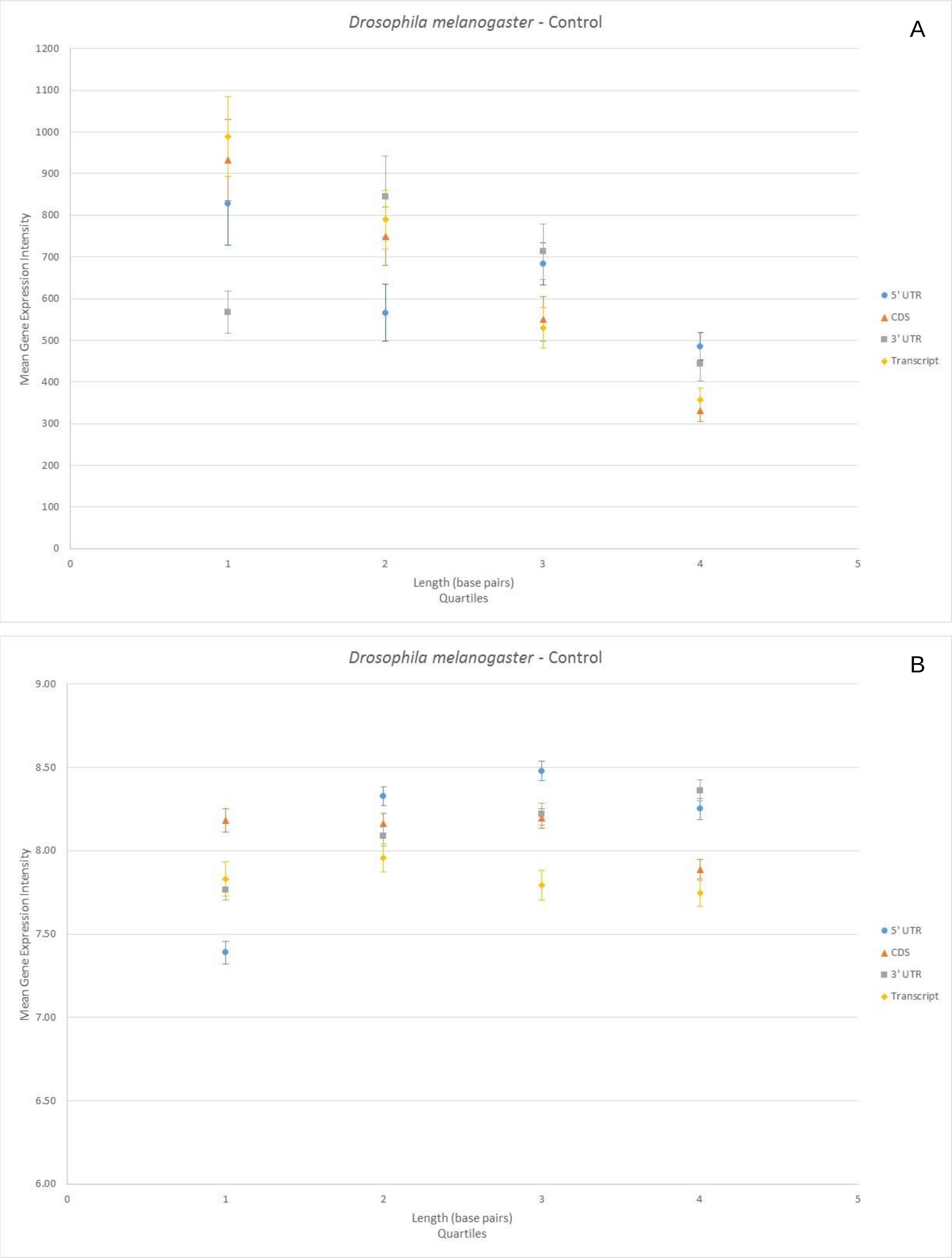


Figure 8.4 Relationship of gene expression in Drosophila melanogaster *within coding and noncoding regions. The gene expression intensity from GEO Datasets – GSE42255 (A) and GSE36507 (B) are plotted versus the quartile score for coding sequence, transcript, 5' UTR and 3' UTR regions. Each data point represents the mean for the samples in each quartile. Error bars represent standard error.*

8.6 Quantile Regression Analysis

The Log function was used to transform the data for an improved view of the quantile regions, a method not applied in the analysis above. Distinct patterns in the quantile regression for both the animal and plant species are evident in the analysis. Firstly, the length of the 5' UTR and the gene expression in both *Arabidopsis* (Table 8.1 / Figure 8.5) and *Drosophila* (Table 8.4 / Figure 8.8) show a positive correlation in the majority of quantiles, indicating as the length of the 5' UTR increases gene expression increases. However, in the *Drosophila* at the 9th quartile, the pattern changes, and shows a negative correlation, indicating that in this quartile for the *Drosophila*, the 5' UTR length increases as the gene expression decreases.

For the CDS length, each species shows a different pattern among the quantiles. For *Arabidopsis* (Table 8.2 / Figure 8.6), the pattern shows a positive correlation for the first six (6) quantiles, and then from 7th quantile there appears to be negative correlation. This would indicate that within the first six quantiles as the CDS length increases, the gene expression increases, and this is reversed past the 7th quantile. The *Drosophila* result (Table 8.5 / Figure 8.9) in all quantiles shows negative correlation, indicating as the CDS length increases, gene expression decreases. This shows two very distinctive patterns between the animal and plant species when the CDS is examined.

Finally for the 3' UTR length, the interesting result for both *Arabidopsis* (Table 8.3 / Figure 8.7) and *Drosophila* (Table 8.6 / Figure 8.10) was that all quantiles showed positive correlation between the 3' UTR length and gene expression. This suggests that as the 3' UTR length increases, gene expression increases.

Overall, the CDS length and gene expression appeared dissimilar between the animal and plant species, with different patterns observed. However, when comparing the 5' UTR and 3' UTR lengths (noncoding regions of the gene) with gene expression data, similarities emerged.

Table 8.1 Quantile regression analysis results on Arabidopsis thaliana between the log of 5' UTR sequence length and the log of gene expression (GSE31488 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	-0.49412	0.26154	-1.88925	0.05887
	Log 5' UTR	0.76284	0.05660	13.47864	0.00000
0.2	Intercept	1.59094	0.19615	8.11066	0.00000
	Log 5' UTR	0.66533	0.04014	16.57554	0.00000
0.3	Intercept	3.42035	0.14372	23.79799	0.00000
	Log 5' UTR	0.47395	0.02935	16.14634	0.00000
0.4	Intercept	4.54025	0.11701	38.80151	0.00000
	Log 5' UTR	0.36948	0.02413	15.30925	0.00000
0.5	Intercept	5.41919	0.10368	52.26962	0.00000
	Log 5' UTR	0.29044	0.02129	13.64508	0.00000
0.6	Intercept	6.21008	0.09453	65.69373	0.00000
	Log 5' UTR	0.22153	0.01970	11.24759	0.00000
0.7	Intercept	6.86249	0.09495	72.27477	0.00000
	Log 5' UTR	0.18379	0.01959	9.38290	0.00000
0.8	Intercept	7.78214	0.09417	82.63627	0.00000
	Log 5' UTR	0.10587	0.02002	5.28773	0.00000
0.9	Intercept	8.71224	0.13789	63.18310	0.00000
	Log 5' UTR	0.06857	0.02834	2.41939	0.01556

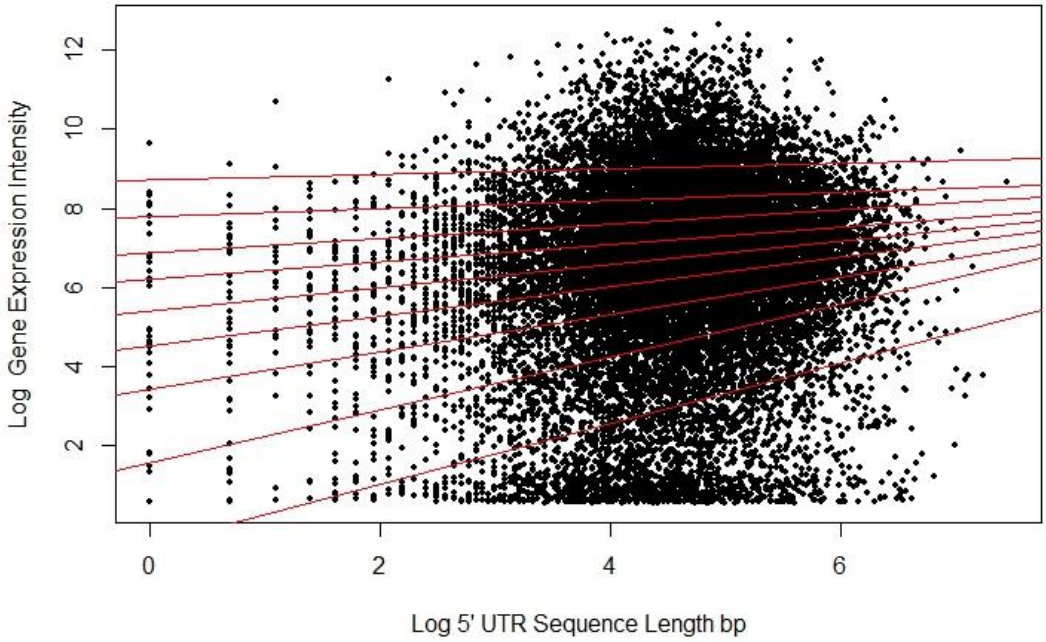


Figure 8.5 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.2 Quantile regression analysis results on Arabidopsis thaliana between the log of CDS sequence length and the log of gene expression (GSE31488 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	-7.25569	0.44465	-16.31763	0.00000
	Log CDS	1.49091	0.06397	23.30774	0.00000
0.2	Intercept	-3.42118	0.35382	-9.66938	0.00000
	Log CDS	1.15369	0.04837	23.85127	0.00000
0.3	Intercept	0.34012	0.25491	1.33425	0.18214
	Log CDS	0.74619	0.03464	21.54367	0.00000
0.4	Intercept	2.92526	0.21057	13.89182	0.00000
	Log CDS	0.47024	0.02846	16.52017	0.00000
0.5	Intercept	5.05024	0.20765	24.32114	0.00000
	Log CDS	0.24374	0.02887	8.44371	0.00000
0.6	Intercept	6.58158	0.19429	33.87460	0.00000
	Log CDS	0.09393	0.02720	3.45279	0.00056
0.7	Intercept	7.89733	0.19186	41.16143	0.00000
	Log CDS	-0.02494	0.02698	-0.92442	0.35528
0.8	Intercept	9.37143	0.20560	45.58077	0.00000
	Log CDS	-0.15842	0.02889	-5.48414	0.00000
0.9	Intercept	11.57666	0.23543	49.17176	0.00000
	Log CDS	-0.36614	0.03348	-10.93737	0.01556

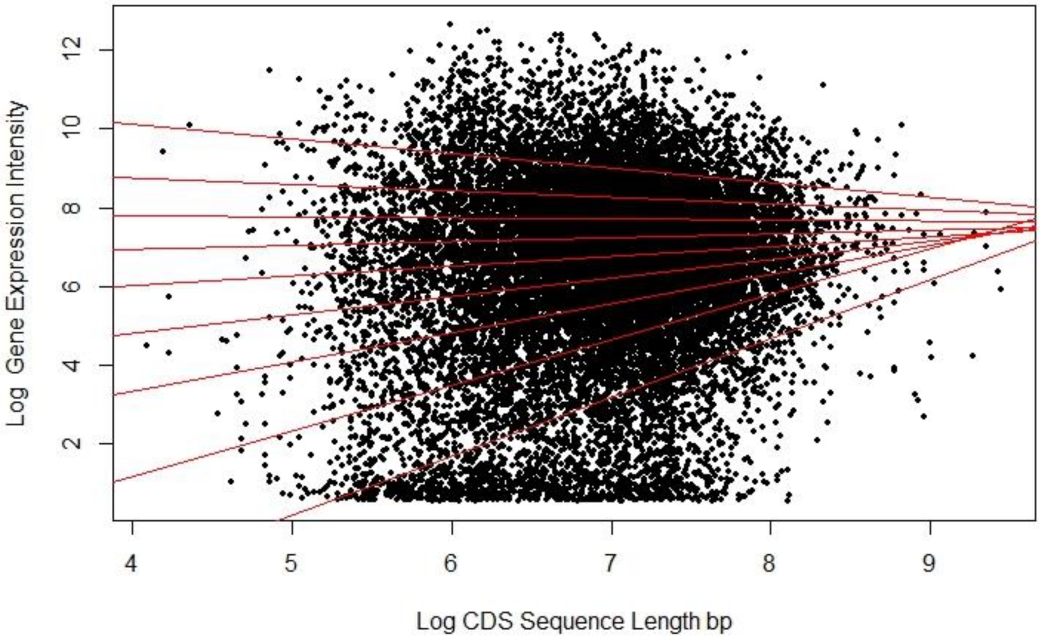


Figure 8.6 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.3 Quantile regression analysis results on Arabidopsis thaliana between the log of 3' UTR sequence length and the log of gene expression (GSE31488 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	-1.28003	0.43043	-2.97380	0.00295
	Log 3' UTR	0.80649	0.08240	9.78727	0.00000
0.2	Intercept	-0.05783	0.36157	-0.15994	0.87293
	Log 3' UTR	0.90699	0.06794	13.34944	0.00000
0.3	Intercept	0.76806	0.22584	3.40086	0.00067
	Log 3' UTR	0.92050	0.04268	21.56976	0.00000
0.4	Intercept	1.59305	0.20227	7.87587	0.00000
	Log 3' UTR	0.88246	0.03802	23.21245	0.00000
0.5	Intercept	2.10509	0.16906	12.45189	0.00000
	Log 3' UTR	0.88162	0.03187	27.66366	0.00000
0.6	Intercept	2.77838	0.16847	16.49139	0.00000
	Log 3' UTR	0.84038	0.03151	26.66661	0.00000
0.7	Intercept	3.31070	0.15947	20.76044	0.00000
	Log 3' UTR	0.82708	0.03010	27.48028	0.00000
0.8	Intercept	4.04752	0.19399	20.86466	0.00000
	Log 3' UTR	0.79168	0.03630	21.81226	0.00000
0.9	Intercept	4.96045	0.15761	31.47265	0.00000
	Log 3' UTR	0.76333	0.03044	25.07947	0.00000

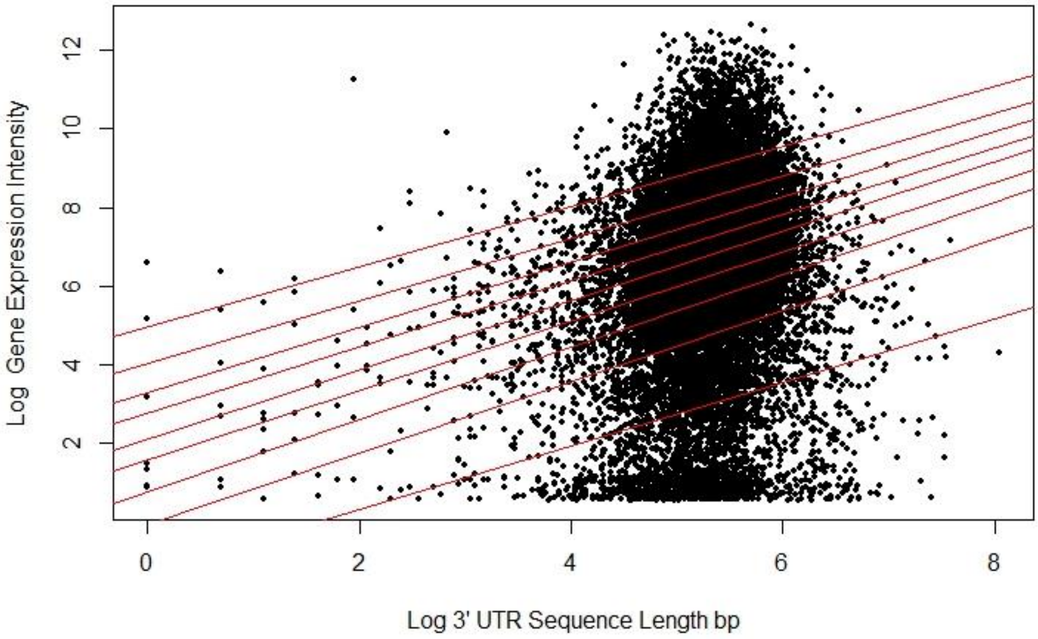


Figure 8.7 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.4 Quantile regression analysis results on Drosophila melanogaster between the log of 5' UTR sequence length and the log of gene expression (GSE42255 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	2.99298	0.24447	12.24295	0.00000
	Log 5' UTR	0.09999	0.05405	1.85007	0.06439
0.2	Intercept	3.70770	0.18407	20.14305	0.00000
	Log 5' UTR	0.10567	0.04101	2.57692	0.01001
0.3	Intercept	4.23852	0.16040	26.42456	0.00000
	Log 5' UTR	0.10295	0.03436	2.99667	0.00275
0.4	Intercept	4.60339	0.13736	33.51435	0.00000
	Log 5' UTR	0.10315	0.03002	3.43561	0.00060
0.5	Intercept	4.95174	0.12961	38.20391	0.00000
	Log 5' UTR	0.09782	0.02782	3.51598	0.00044
0.6	Intercept	5.31990	0.11417	46.59480	0.00000
	Log 5' UTR	0.08538	0.02580	3.30864	0.00095
0.7	Intercept	5.57721	0.14373	38.80300	0.00000
	Log 5' UTR	0.10068	0.02962	3.39947	0.00068
0.8	Intercept	6.44681	0.17437	36.97142	0.00000
	Log 5' UTR	0.00413	0.03571	0.11577	0.90784
0.9	Intercept	7.68730	0.22150	34.70589	0.00000
	Log 5' UTR	-0.12948	0.04432	-2.92142	0.00351

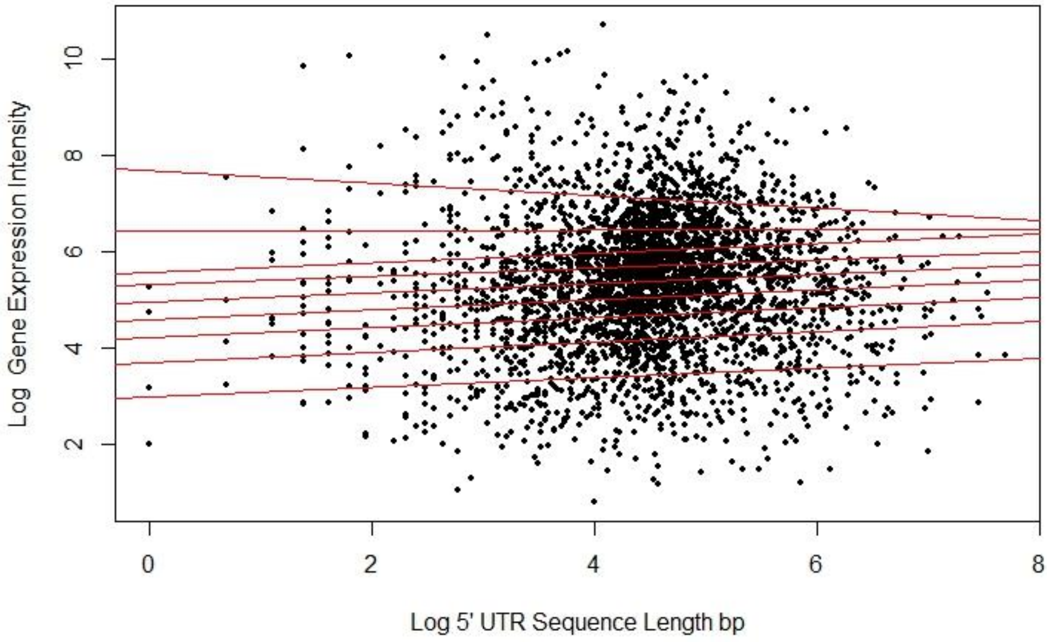


Figure 8.8 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.5 Quantile regression analysis results on Drosophila melanogaster between the log of CDS sequence length and the log of gene expression (GSE42255 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	4.84616	0.57613	8.41160	0.00000
	Log CDS	-0.19783	0.08014	-2.46860	0.01361
0.2	Intercept	6.07890	0.44560	13.64200	0.00000
	Log CDS	-0.27198	0.06264	-4.34194	0.00001
0.3	Intercept	7.03180	0.34350	20.47108	0.00000
	Log CDS	-0.33594	0.04954	-6.78088	0.00000
0.4	Intercept	7.59350	0.32054	23.68971	0.00000
	Log CDS	-0.36004	0.04521	-7.96411	0.00000
0.5	Intercept	7.95531	0.28030	28.38178	0.00000
	Log CDS	-0.36548	0.04029	-9.07219	0.00000
0.6	Intercept	8.49326	0.26724	31.78114	0.00000
	Log CDS	-0.40083	0.03742	-10.71079	0.00000
0.7	Intercept	9.08676	0.28572	31.80286	0.00000
	Log CDS	-0.44001	0.04007	-10.98103	0.00000
0.8	Intercept	9.70859	0.34560	28.09197	0.00000
	Log CDS	-0.47106	0.04905	-9.60348	0.00000
0.9	Intercept	11.36497	0.42722	26.60187	0.00000
	Log CDS	-0.61925	0.05964	-10.38366	0.00000

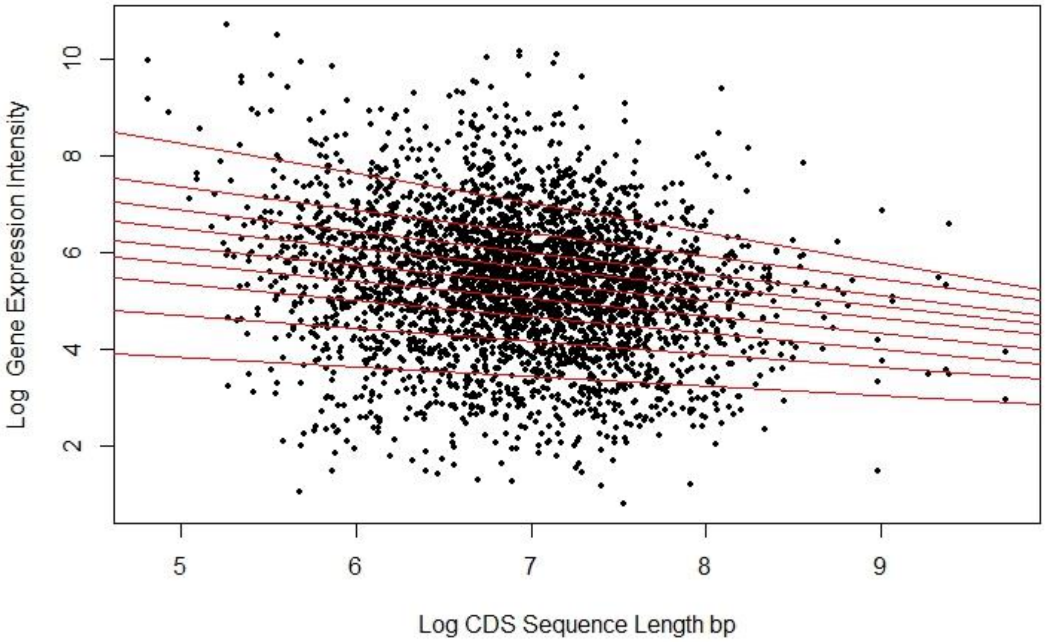


Figure 8.9 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.6 Quantile regression analysis results on Drosophila melanogaster between the log of 3' UTR sequence length and the log of gene expression (GSE42255 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	4.03542	0.08225	49.06280	0.00000
	Log 3' UTR	0.35700	0.01546	23.09641	0.00000
0.2	Intercept	4.48795	0.08257	54.35470	0.00000
	Log 3' UTR	0.31840	0.01614	19.72846	0.00000
0.3	Intercept	4.85297	0.06886	70.47384	0.00000
	Log 3' UTR	0.28407	0.01273	22.30680	0.00000
0.4	Intercept	5.03650	0.06313	79.77872	0.00000
	Log 3' UTR	0.27350	0.01290	21.20361	0.00000
0.5	Intercept	5.15329	0.05986	86.09041	0.00000
	Log 3' UTR	0.27983	0.01173	23.85901	0.00000
0.6	Intercept	5.37147	0.06384	84.13507	0.00000
	Log 3' UTR	0.26200	0.01199	21.84826	0.00000
0.7	Intercept	5.61639	0.06897	81.43580	0.00000
	Log 3' UTR	0.24204	0.01329	18.21441	0.00000
0.8	Intercept	5.94198	0.07576	78.43611	0.00000
	Log 3' UTR	0.20878	0.01507	13.85648	0.00000
0.9	Intercept	6.14204	0.09774	62.84273	0.00000
	Log 3' UTR	0.21414	0.01879	11.39432	0.00000

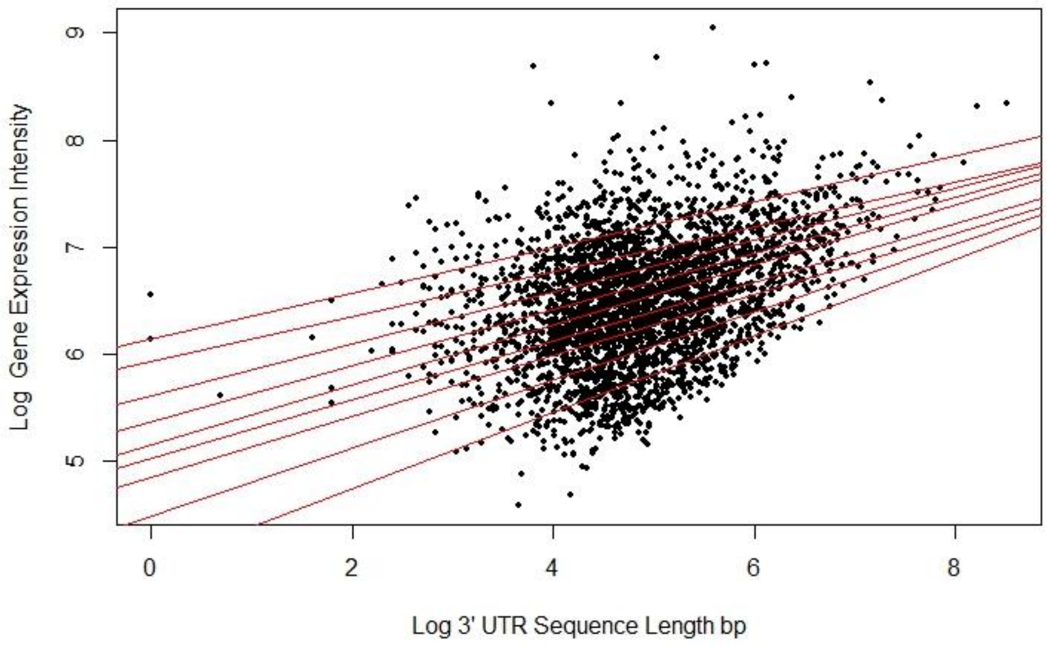


Figure 8.10 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.7 Quantile regression analysis results on Arabidopsis thaliana between the log of 5' UTR sequence length and the log of gene expression (GDS3933 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	0.76451	0.06266	12.20094	0.00000
	Log 5' UTR	0.13061	0.01399	9.33390	0.00000
0.2	Intercept	1.10719	0.05709	19.39225	0.00000
	Log 5' UTR	0.12840	0.01203	10.67434	0.00000
0.3	Intercept	1.46345	0.04631	31.60327	0.00000
	Log 5' UTR	0.09359	0.00930	10.06655	0.00000
0.4	Intercept	1.76655	0.02690	65.66343	0.00000
	Log 5' UTR	0.05411	0.00541	10.00009	0.00000
0.5	Intercept	1.91977	0.02053	93.49082	0.00000
	Log 5' UTR	0.03789	0.00419	9.03235	0.00000
0.6	Intercept	2.04182	0.01720	118.70693	0.00000
	Log 5' UTR	0.02580	0.00356	7.23725	0.00000
0.7	Intercept	2.12815	0.01635	130.17843	0.00000
	Log 5' UTR	0.01962	0.00340	5.77582	0.00000
0.8	Intercept	2.22796	0.01691	131.74460	0.00000
	Log 5' UTR	0.01224	0.00355	3.44724	0.00057
0.9	Intercept	2.34085	0.01530	153.04522	0.00000
	Log 5' UTR	0.00692	0.00338	2.04747	0.04064

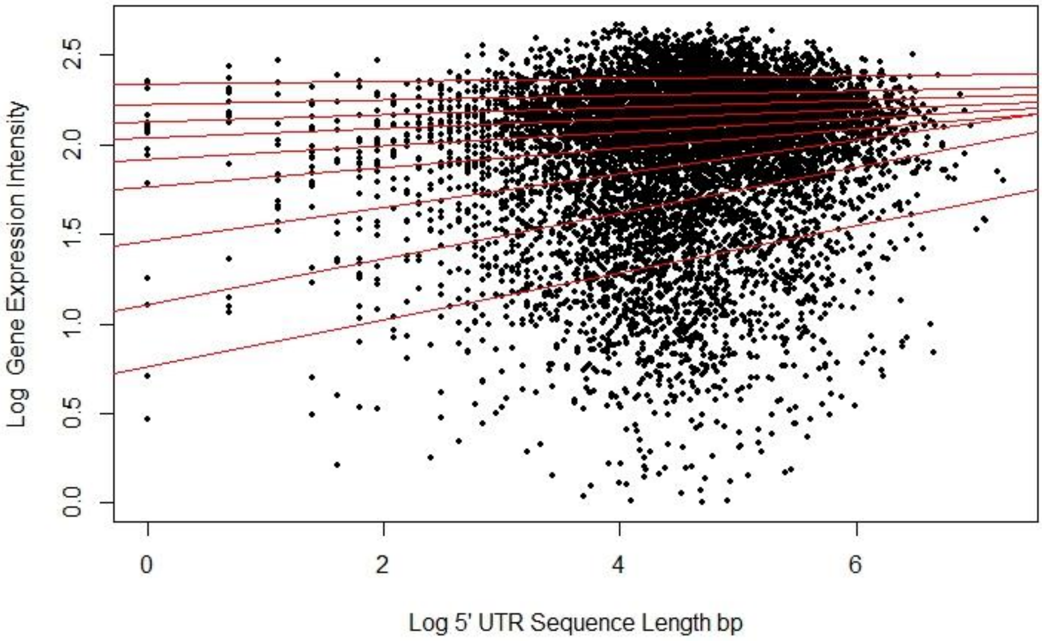


Figure 8.11 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.8 Quantile regression analysis results on Arabidopsis thaliana between the log of CDS sequence length and the log of gene expression (GDS3933 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	-0.19868	0.15081	-1.31744	0.18772
	Log CDS	0.22601	0.02155	10.48761	0.00000
0.2	Intercept	0.44103	0.11273	3.91235	0.00009
	Log CDS	0.17926	0.01574	11.39117	0.00000
0.3	Intercept	1.22926	0.08138	15.10571	0.00000
	Log CDS	0.09495	0.01100	8.63142	0.00000
0.4	Intercept	1.89539	0.04992	37.97150	0.00000
	Log CDS	0.01738	0.00666	2.61013	0.00907
0.5	Intercept	2.21277	0.03694	59.90485	0.00000
	Log CDS	-0.01660	0.00509	-3.25824	0.00113
0.6	Intercept	2.39485	0.03619	66.17765	0.00000
	Log CDS	-0.03350	0.00508	-6.59286	0.00000
0.7	Intercept	2.58925	0.03022	85.68926	0.00000
	Log CDS	-0.05279	0.00423	-12.48248	0.00000
0.8	Intercept	2.72644	0.03013	90.48563	0.00000
	Log CDS	-0.06385	0.00433	-14.75102	0.00000
0.9	Intercept	2.84588	0.03396	83.79063	0.00000
	Log CDS	-0.06942	0.00490	-14.18200	0.00000

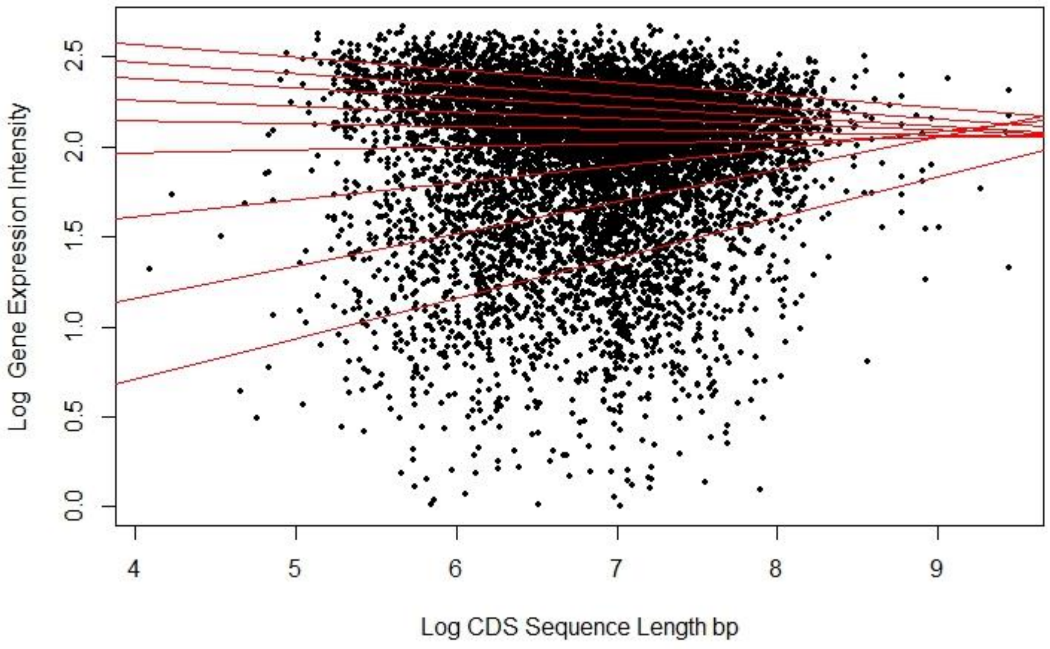


Figure 8.12 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.9 Quantile regression analysis results on Arabidopsis thaliana between the log of 3' UTR sequence length and the log of gene expression (GDS3933 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	0.026500	0.131040	0.202270	0.839710
	Log 3' UTR	0.254060	0.024960	10.177990	0.000000
0.2	Intercept	0.404410	0.120200	3.364370	0.000770
	Log 3' UTR	0.244440	0.022430	10.895530	0.000000
0.3	Intercept	0.868450	0.077930	11.144220	0.000000
	Log 3' UTR	0.195080	0.014290	13.648320	0.000000
0.4	Intercept	1.242830	0.048620	25.563150	0.000000
	Log 3' UTR	0.145700	0.008920	16.330550	0.000000
0.5	Intercept	1.450620	0.040690	35.652010	0.000000
	Log 3' UTR	0.121350	0.007440	16.313380	0.000000
0.6	Intercept	1.614860	0.033970	47.538100	0.000000
	Log 3' UTR	0.102630	0.006330	16.220570	0.000000
0.7	Intercept	1.722880	0.028790	59.848790	0.000000
	Log 3' UTR	0.093360	0.005410	17.256770	0.000000
0.8	Intercept	1.816540	0.034440	52.742530	0.000000
	Log 3' UTR	0.087470	0.006430	13.599130	0.000000
0.9	Intercept	1.942250	0.034110	56.939020	0.000000
	Log 3' UTR	0.080820	0.006430	12.572240	0.000000

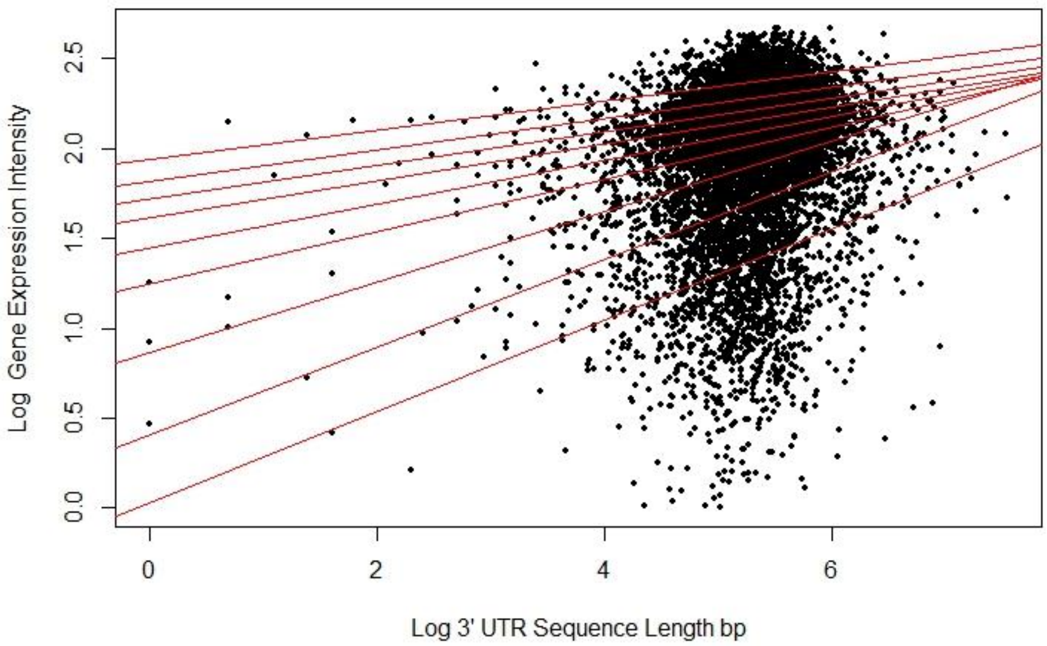


Figure 8.13 Quantile regression plot for Arabidopsis thaliana with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.10 Quantile regression analysis results on Drosophila melanogaster between the log of 5' UTR sequence length and the log of gene expression (GSE36507 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	1.28497	0.06779	18.95641	0.00000
	Log 5' UTR	0.07333	0.01374	5.33611	0.00000
0.2	Intercept	1.59194	0.04478	35.54779	0.00000
	Log 5' UTR	0.05687	0.00909	6.25711	0.00000
0.3	Intercept	1.73679	0.02932	59.23962	0.00000
	Log 5' UTR	0.05415	0.00614	8.82520	0.00000
0.4	Intercept	1.82558	0.02286	79.85883	0.00000
	Log 5' UTR	0.05229	0.00465	11.24806	0.00000
0.5	Intercept	1.90830	0.01715	111.28562	0.00000
	Log 5' UTR	0.04659	0.00343	13.60097	0.00000
0.6	Intercept	1.99646	0.01926	103.68233	0.00000
	Log 5' UTR	0.03888	0.00388	10.01241	0.00000
0.7	Intercept	2.06843	0.01618	127.85772	0.00000
	Log 5' UTR	0.03482	0.00320	10.86927	0.00000
0.8	Intercept	2.16414	0.01809	119.63192	0.00000
	Log 5' UTR	0.02680	0.00357	7.50320	0.00000
0.9	Intercept	2.33342	0.02376	98.21304	0.00000
	Log 5' UTR	0.00693	0.00468	1.48068	0.13875

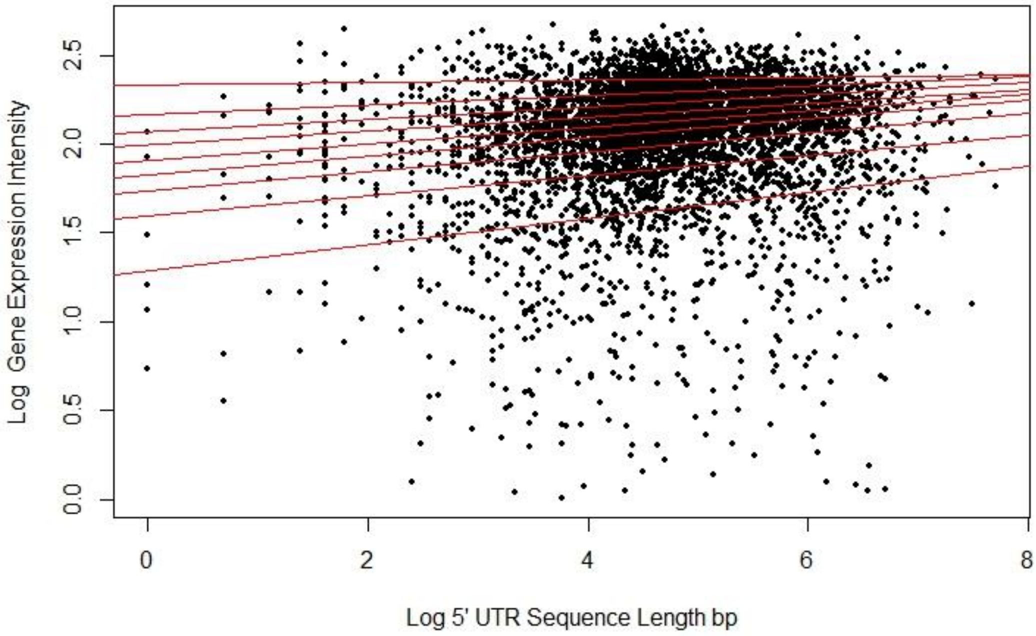


Figure 8.14 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.11 Quantile regression analysis results on Drosophila melanogaster between the log of CDS sequence length and the log of gene expression (GSE36507 gene expression experiment data)

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	1.41356	0.14825	9.53503	0.00000
	Log CDS	0.02975	0.02106	1.41256	0.15784
0.2	Intercept	1.81659	0.08947	20.30396	0.00000
	Log CDS	0.00639	0.01276	0.50060	0.61667
0.3	Intercept	1.99122	0.06902	28.85156	0.00000
	Log CDS	-0.00075	0.00975	-0.07669	0.93887
0.4	Intercept	2.11710	0.04834	43.79536	0.00000
	Log CDS	-0.00646	0.00688	-0.93966	0.34744
0.5	Intercept	2.23380	0.03819	58.48915	0.00000
	Log CDS	-0.01500	0.00536	-2.79907	0.00514
0.6	Intercept	2.36145	0.03839	61.51257	0.00000
	Log CDS	-0.02540	0.00533	-4.76824	0.00000
0.7	Intercept	2.45361	0.03457	70.97384	0.00000
	Log CDS	-0.03070	0.00489	-6.27982	0.00000
0.8	Intercept	2.58566	0.03616	71.50751	0.00000
	Log CDS	-0.04147	0.00506	-8.19982	0.00000
0.9	Intercept	2.73350	0.03934	69.48483	0.00000
	Log CDS	-0.05262	0.00554	-9.49641	0.00000

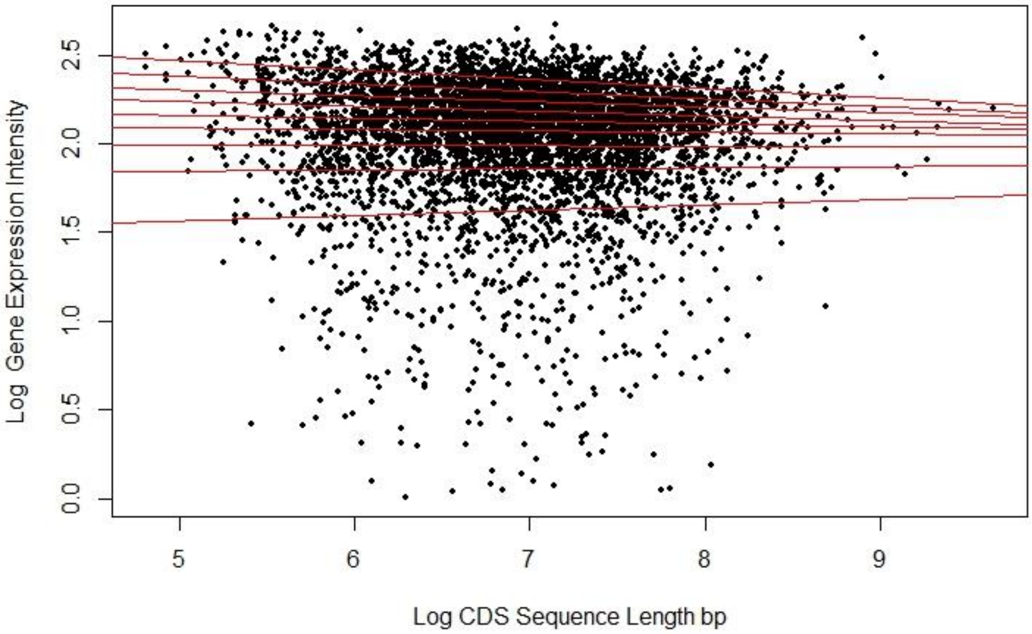


Figure 8.15 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

Table 8.12 Quantile regression analysis results on Drosophila melanogaster between the log of 3' UTR sequence length and the log of gene expression

Quantile		Value	Std. Error	t value	Pr (>|t|)
0.1	Intercept	1.53054	0.08075	18.95458	0.00000
	Log 3' UTR	0.01735	0.01533	1.13164	0.25784
0.2	Intercept	1.79567	0.04229	42.46001	0.00000
	Log 3' UTR	0.01303	0.00840	1.55044	0.12110
0.3	Intercept	1.89908	0.03645	52.09484	0.00000
	Log 3' UTR	0.01735	0.00720	2.40887	0.01604
0.4	Intercept	1.95487	0.02397	81.56678	0.00000
	Log 3' UTR	0.02306	0.00468	4.92510	0.00000
0.5	Intercept	1.99859	0.01908	104.74509	0.00000
	Log 3' UTR	0.02653	0.00382	6.95355	0.00000
0.6	Intercept	2.03241	0.01928	105.41200	0.00000
	Log 3' UTR	0.02976	0.00376	7.90712	0.00000
0.7	Intercept	2.08155	0.01642	126.79634	0.00000
	Log 3' UTR	0.03018	0.00309	9.75937	0.00000
0.8	Intercept	2.16193	0.01891	114.31813	0.00000
	Log 3' UTR	0.02546	0.00356	7.15683	0.00000
0.9	Intercept	2.27943	0.02397	95.09426	0.00000
	Log 3' UTR	0.01735	0.00458	3.79243	0.00015

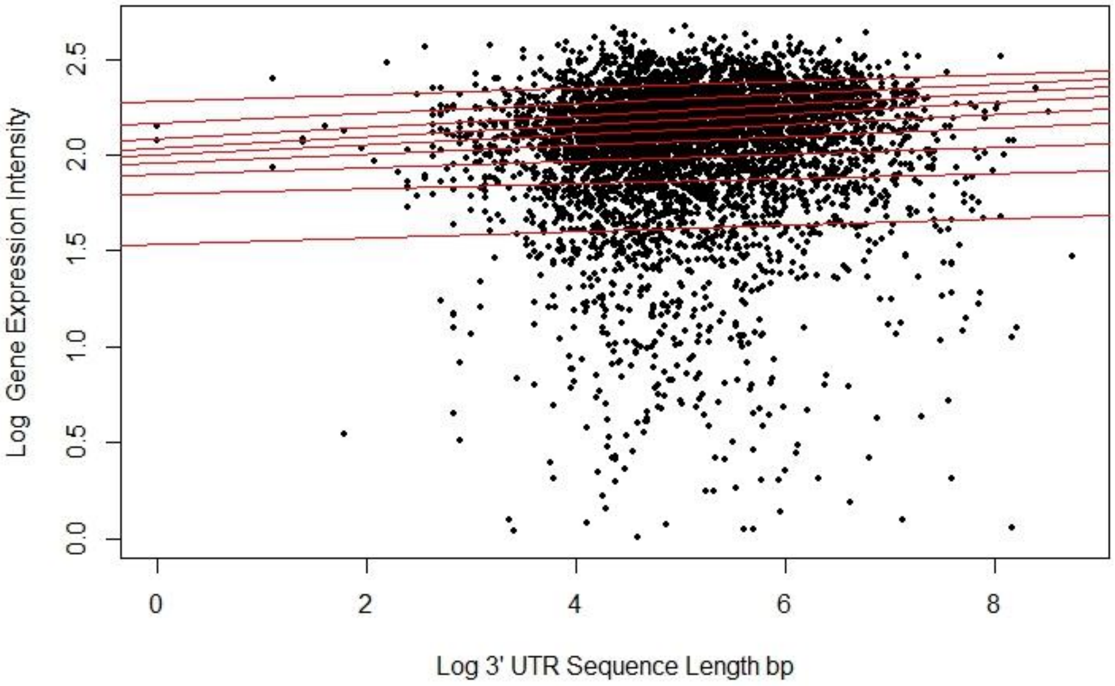


Figure 8.16 Quantile regression plot for Drosophila melanogaster with quantiles range from 0.1 to 0.9 in increments of 0.1, respectively.

The quantile regression statistical analyses was again applied to the second set of gene expression data to substantiate this method under different gene expression experiments. The results show very similar patterns to the previous gene expression experiment, indicating the model is robust in studying the relationship between gene expression and the length of coding and noncoding regions in different species (Tables 8.7-8.12 / Figures 8.11-8.16). Both gene expression datasets showed statistical significance across all quantile groups, indicating a relationship between the coding and noncoding length and gene expression in animal and plant species.

The observed expression trends in both experimental datasets suggests that there are differences between animal and plant species when considering CDS length and that the noncoding regions show similar patterns of positive correlation to gene expression.

8.7 Discussion

We aimed to develop an understanding of the relationship between the coding and noncoding sequence length in association with gene expression between an animal and plant species. In brief the findings suggest from the quantile regression analysis: (i) the patterns seen between the CDS length and gene expression intensity in *Arabidopsis* and *Drosophila* are different, the plant species showing both positive and negative correlation dependent on the quantile whilst the animal species showing a consistent negative correlation among all quantiles; (ii) in both the animal and plant species the 3’ UTR length and gene expression exhibit positive correlation.

The current research has confirmed our previous findings with the *Arabidopsis* (Caldwell, et al., 2010) and is also consistent with previous research, where it was found that highly expressed genes have larger primary transcripts [15]. Extensive studies with *Arabidopsis* has inferred that multistimuli response genes (genes that are differentially expressed in response to a large number of different external stimuli) have significantly longer upstream intergenic regions and are generally shorter (Walther, et al., 2007). A more recent study investigating the translational efficiency in *Arabidopsis* has proposed that the sequence context immediately upstream from the AUG initiation codon in plant genes are critical in determining translational efficiency (Kim, et al., 2014). Other studies investigating the role of the 5’ UTR in translational regulation found that nucleotide composition, length, potential secondary structure and the presence of uAUGs have a considerable effect on ribosome loading in *Arabidopsis* (Kawaguchi and Bailey-Serres, 2005). Furthermore, additional studies have focused on the GC content showing large variability among species, ~20 to 60% variation in eukaryotes (Lynch, 2007). Based on findings from Duret and Stoletzki GC3-rich genes tend to be shorter than GC3-poor genes (Duret and Mouchiroud, 1999; Stoletzki, 2011). To investigate the hypothesis of synonymous codon usage (SCU), which is described as highly expressed genes undergoing stronger translational selection, for example higher GC content, in seeded plants, Serres-Giardi et al tested GC3-rich and GC-poor genes against expression. It was found that in 154 plant species tested, expression was significantly and positively correlated with GC3 (Serres-Giardi, et al., 2012). The results from these studies are interesting with respect to our results, and may support and extend the understanding of gene architecture and gene expression in plants.

In addition, the patterns found in the coding sequences for *Drosophila* is consistent with previous research with animals. A study on *Gallus gallus* (chicken), found that the coding sequence length is negatively correlated with expression level (Rao, et al., 2010) as shown in the *Drosophila* in this study. In other animal investigations, the research also reported that in highly expressed genes the length of the coding sequence and protein lengths were small (Raghava and Han, 2005; Subramanian and Kumar, 2004). A popular bioinformatics technique used to detect subtle variations in sequences was used to identify differences between the 3' UTR and protein coding sequences in the *Drosophila*. Interestingly, the study found greater number of segments in the 3' UTR, suggesting greater functional complexity in the 3' UTRs than in the coding sequence (Algama, et al., 2014). This could explain the differences in the CDS and 3' UTR patterns found in this study. Genome size is also another important aspect in determining variability between organisms. A *Drosophila melanogaster* study has shown that genomes are subjected to constant change not only in their size but in their composition (Boulesteix, et al., 2006).

Identification of similarities and differences in genomes, particularly between animals and plants that might result in speciation has had a great deal of interest, with gene families, gene loss and gene amplification being the focus of these studies (Ball and Cherry, 2001). The genomes of *Arabidopsis* and *Drosophila* are of similar size, however the number of genes identified vary, ~26,000 for *Arabidopsis* and ~14,000 for *Drosophila*. Differences start to emerge when gene families are examined, *Arabidopsis* appear to have 11,000 gene families, which have more than five members, in contrast to *Drosophila* which encode fewer genes (Initiative, 2000). Understanding the genome structure of these organisms before examining the finer details of the genome itself is an important strategy.

When the coding sequence is examined in association with gene expression there seems to be divergence in *Arabidopsis* and *Drosophila*, although we cannot yet conclude and refer in general to the difference between animal and plant genomes. Differences seen in the animal and plants species may be described by differences in life strategies (Kejnovsky, et al., 2009). Plant genomes appear much more dynamic (Murat, et al., 2012), due to the sessile nature and response to adverse conditions through biochemical complexity and developmental plasticity (Wilczek, et al., 2009). In contrast, animal genomes are more conserved and stable, attributable to the ability to avoid adverse conditions (Murat, et al., 2012). There has been overwhelming evidence that natural selection appears to support the compactness of highly expressed genes in both animal and plant species (Castillo-Davis, et al., 2002; Eisenberg and Levanon, 2003; Rao, et al., 2010; Stenoien, 2007; Yang, 2009). These results may elucidate to the theory on reduction costs of energy with shorter proteins and sequences, contributing to minimizing the cost of synthesis (Vilaprinyo, et al., 2010). However it is important to highlight that the length of the coding region is only one of several factors that contribute to the complex nature of natural selection, species complexity and gene regulation.

Furthermore, the noncoding untranslated sequences have been identified as important components in the regulation of transcription and translation, influencing translation initiation, stability, elongation, and the termination of the mRNA translation (Barrett, et al., 2012). Modification to the lengths of the 5’ UTR and 3’ UTR sequences may contribute to the selective constraints between animal and plants species, and may be influenced by environmental conditions (Chen, et al., 2011). For the 3’ UTR regions, the results of this study have shown similarities in the patterns between *Arabidopsis* and *Drosophila*, that is, positive correlation between length and gene expression. This is in agreement with our previous research for *Arabidopsis* (Caldwell, et al., 2010).

The regulation of many genes has been known to be controlled primarily by 3’ UTR’s, particularly those involved in development (Merritt, et al., 2008). Other research has found that there was positive correlation with transposon and simple sequence repeats (SSRs), with these elements affecting the length and variation of both the 5’ and 3’ UTRs (Liu, et al., 2012). Differing lengths of the untranslated regions could also be affected by either selection or genetic drift (Chen, et al., 2011). These results may enforce the concept that these untranslated regions are prone to a higher level of environmental and evolutionary constraints compared to the coding sequences and it is plausible that selection shapes these lengths. However, Chen *et al* (2011) looked at over 15 species and found that the elongation of 5’ UTR alone cannot lead to the emergence of organismal complexity (Chen, et al., 2011), indicating that the untranslated regions may not be a true indication of organism evolution, thus supporting the similarities found in this research in the untranslated regions.

Furthermore, recent experimental studies have shed light on the complex ceRNA network dynamics in prostate cancer using the alternative cleavage and polyadenylation (APA). This study concluded that long 3’ UTRs tend to harbour more microRNA response elements (MREs) which in turn would influence biological process when the 3’ UTR length is modified. The understanding of 3’ UTR shortening has great potential in creating prognostic markers for oncogene expression (Li, et al., 2014). Other research in mammalian brain development proposes that lengthening of 3’ UTRs offers considerable versatility in biological processes (Miura, et al., 2013). The findings in this study have amplified the importance of the noncoding 5’ and 3’ UTR regions, and has shown differences in these regions compared to the coding sequence.

At a global scale, the picture emerging is that animal and plant species show similarities and divergences when comparisons are made with gene expression and the length distributions of the coding and noncoding regions. However, studying the association between expression levels and length can be intricate to interpret, including sample size variation between organisms, statistical methodology and data transformation. It was our intention to take advantage of available genomic data to identify general responses and relations. Using the available technologies and data our results have shown some interesting correlation between gene expression and the basic gene architecture, length, especially in the 3’ UTR region. Further research is required to explore more details in the gene length distribution variations of different genes and different organisms, including known highly expressed genes such as heat shock protein genes (HSPs).

Conclusion and Future Research

9 Conclusion and Future Research

9.1 Conclusion

"….. bioinformatics, defined as the computational handling and processing of genetic information, has become one of the most highly visible fields of modern science." (Ouzounis and Valencia, 2003)

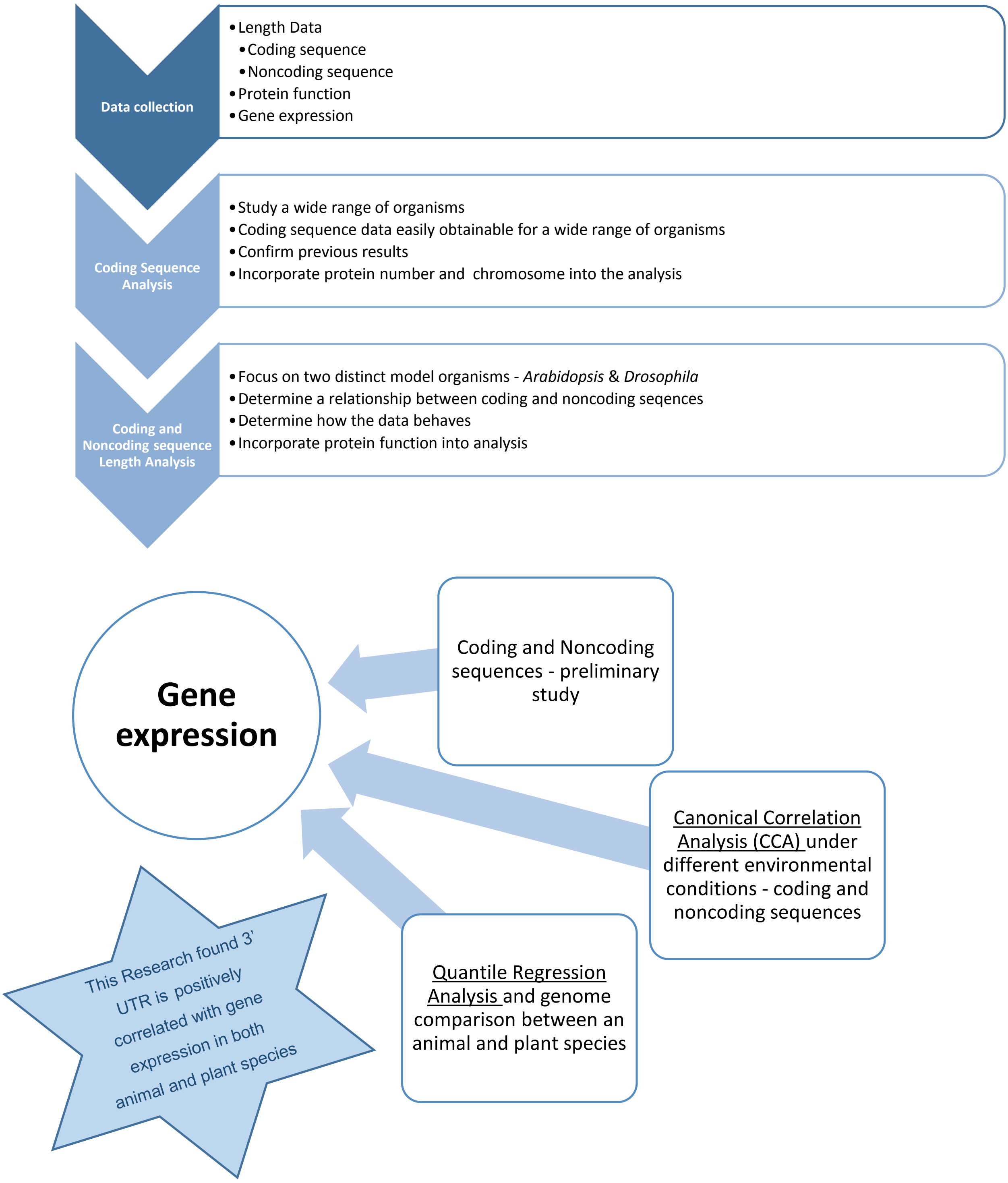

Data collection
•Length Data
•Coding sequence
•Noncoding sequence
•Protein function
•Gene expression
Coding Sequence Analysis
•Study a wide range of organisms
•Coding sequence data easily obtainable for a wide range of organisms
•Confirm previous results
•Incorporate protein number and chromosome into the analysis
Coding and Noncoding sequence Length Analysis
•Focus on two distinct model organisms - Arabidopsis & Drosophila
•Determine a relationship between coding and noncoding seqences
•Determine how the data behaves
•Incorporate protein function into analysis
Gene expression
Coding and Noncoding sequences - preliminary study
Canonical Correlation Analysis (CCA) under different environmental conditions - coding and noncoding sequences
Quantile Regression Analysis and genome comparison between an animal and plant species
This Research found 3' UTR is positively correlated with gene expression in both animal and plant species

Figure 9.1 Workflow summary of the main research points conducted in this thesis

The research in this thesis has validated the data that is publicly available on the web. It has scrutinized the data is available, including gene expression data and has shown patterns not discovered previously by other research studies. With the large amount of data readily available, it is important that strict guidelines on creating, storing and testing data is followed for the future science community to be confident the data is of high quality and accurate.

A big challenge in the post-genome-sequencing era, for deciphering the gene regulation networks, is to improve computational techniques that were lacking in accuracy. It has been shown that using the TSS-TLS and TLS-TSC distances, promoter prediction can be improved with the NNPP2.2 algorithm. However, this new technique does not have to be restricted to this program, but may be applied post-process to many other promoter prediction algorithms that also suffer from a high incidence of false positives.

The work in this thesis has also shown that there is a possible correlation between the coding and noncoding regions of protein coding genes using a variety of statistical methods. Other factors were also introduced, such as protein function and gene expression that have been a topic of interest for many scientists. Standard statistical models have identified interesting areas for further investigate, giving a focus and direction for this thesis, and allowing more complex models to be used to describe identify patterns and correlations, not previously found.

Using The Arabidopsis Information Resource (TAIR) database it has been possible to evaluate the relationship between the coding and noncoding sequences in relation to gene expression in the *Arabidopsis thaliana*. The research in this thesis has confirmed previous research on protein lengths. The patterns found have contributed some generalized understanding about the relationship between gene expression and protein function. Our results show excellent concordance with previous studies that have identified highly expressed genes are more compact when looking at the coding sequence length in both animal and plant species. However the noncoding sequence length show variation among animal and plant species.

Furthermore, the CCA method was fruitful in identifying associations between length distributions and gene expression. The research has successfully used CCA to categorize the relationship between the length distributions of coding and noncoding genes and gene expression exposed to various environmental factors. It is an easily accessible and customizable tool that that can boost insight into more complex relationships between gene architecture and gene expression. The analysis has found a relationship between the 5’ un-translated region and longevity, however this method is difficult to interpret and is limited by sample size. Therefore, these results should be treated with caution until they are confirmed by additional studies.

The difficulty in using model organisms is that they are often not “typical” and the results can be misleading if used to compare with other organisms. Limitations include for *Arabidopsis* has no know root symbioses; *Drosophila* are not pathogens or pest (Tagu, et al., 2014). However this research is much generalised and other research has benefited from such model organism use.

9.2 Outlook

The following section suggests future improvements to the approach, statistical analysis and data used in this thesis:

1. Other aspects of a protein, such as structure, regulation and localization are defined much more clearly and may show an obvious relationship with gene expression (Gerstein and Jansen, 2000);
2. To extend on the findings with gene expression data, it would be prudent to including in future studies controlled vocabularies, such as Gene Ontology (describing gene product characteristics and gene product annotation data) which would aid in the analysis of genome-wide response patterns;
3. Other factors which are worth considering would be tissue type which would be beneficial in broadening understanding;
4. An extension on the quantile regression model that uses the interaction of all three regions (d_1, d_2 and d_3) could show which length region has the most influence on the average gene expression intensity;
5. Focus on specific gene families, such as heat shock proteins and apply quantile regression analysis to gene expression and length data (forthcoming work related to or developing themes in this thesis).

In conclusion, in delving into the patterns of statistical properties of different gene regions and their correlation we intended to elucidate the spatial organization rules between various gene functional elements and the difference in such organizations among different living organisms and gene families. We believe that these rules and differences are the results of organism complexity and reflect the complexity differences in the regulation of gene expression. The information described in this thesis provides the basis for further exploration into gene regulation and architecture, with regard to sequence length of the coding and noncoding regions. With more organism genome-wide data becoming available to study and new methods and technologies to explore, we can look forward to a surge in genome-wide comparative research.

References

33rd_Square What Does The Exponential Growth Of Genomics Data Actually Mean?

Achaz, G., Coissac, E., Viari, A. and Netter, P. (2000) Analysis of Intrachromosomal Duplications in Yeast Saccharomyces cerevisiae: A Possible Model for Their Origin, *Molecular Biology and Evolution*, **17**, 1268-1275.

Achaz, G., Netter, P. and Coissac, E. (2001) Study of Intrachromosomal Duplications Among the Eukaryote Genomes, *Molecular Biology and Evolution*, **18**, 2280-2288.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.-H.C., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Miklos, Abril, J.F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., Pablos, B.d., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D.C., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.-Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.-F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M. and Venter, J.C. (2000) The Genome Sequence of Drosophila melanogaster, *Science*, **287**, 2185-2195.

Adams, T.E., Epa, V.C., Garrett, T.P.J. and Ward*, C.W. (2000) Structure and function of the type 1 insulin-like growth factor receptor, *CMLS, Cell. Mol. Life Sci.*, **57**, 1050-1093.

Akaike, H. (1974) A new look at the statistical model identification, *Automatic Control, IEEE Transactions on*, **19**, 716-723.

Akashi, H. (2001) Gene expression and molecular evolution, *Current Opinion in Genetics & Development*, **11**, 660-666.

Alexandrov, N., Troukhan, M., Brover, V., Tatarinova, T., Flavell, R. and Feldmann, K. (2006) Features of Arabidopsis Genes and Genome Discovered using Full-length cDNAs, *Plant Mol Biol*, **60**, 69-85.

Algama, M., Oldmeadow, C., Tasker, E., Mengersen, K. and Keith, J.M. (2014) Drosophila 3′ UTRs Are More Complex than Protein-Coding Sequences, *PLoS ONE*, **9**, e97336.

Andofatto, P. (2005) Adaptive evolution of non-coding DNA in Drosophila, *Nature*, **437**, 1149-1152.

Ashburner, M. and Bergman, C.M. (2005) Drosophila melanogaster: A case study of a model genomic sequence and its consequences, *Genome Research*, **15**, 1661-1667.

Bajic, V.B., Tan, S.L., Suzuki, Y. and Sugano, S. (2004) Promoter prediction analysis on the whole human genome, *Nat Biotech*, **22**, 1467-1473.

Ball, C.A. and Cherry, J.M. (2001) Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms, *Current opinion in chemical biology*, **5**, 86-89.

Barrett, L.W., Fletcher, S. and Wilton, S.D. (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements, *Cell Mol. Life Sci*, **69**, 3613-3634.

Bhardwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein–protein interactions within and across genomes, *Bioinformatics*, **21**, 2730-2738.

Boulesteix, M., Weiss, M. and Biémont, C. (2006) Differences in Genome Size Between Closely Related Species: The Drosophila melanogaster Species Subgroup, *Molecular Biology and Evolution*, **23**, 162-167.

Bracco, L. and Kearsey, J. (2003) The relevance of alternative RNA splicing to pharmacogenomics, *Trends in Biotechnology*, **21**, 346-353.

Bressan, R., Bohnert, H. and Zhu, J.-K. (2009) Abiotic Stress Tolerance: From Gene Discovery in Model Organisms to Crop Improvement, *Molecular Plant*, **2**, 1-2.

Briestanska, J. and Plachy, J. (1996) Influence of the transduced 3′UTR of the c-src oncogene on tumour growth induced by the v-src gene of avian sarcoma virus PR2257, *Journal of General Virology*, **77**, 1189-1192.

Brocchieri, L. and Karlin, S. (2005) Protein length in eukaryotic and prokaryotic proteomes, *Nucleic Acids Research*, **33**, 3390-3400.

Brogna, S. and Wen, J. (2009) Nonsense-mediated mRNA decay (NMD) mechanisms, *Nat Struct Mol Biol*, **16**, 107-113.

Brown, E.J. and Bachtrog, D. (2014) The chromatin landscape of Drosophila: comparisons between species, sexes, and chromosomes, *Genome Research*, **24**, 1125-1137.

Burden, S., Lin, Y.-X. and Zhang, R. (2005) Improving promoter prediction Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences, *Bioinformatics*, **21**, 601-607.

Buxbaum, E. (2007) *Fundamentals of Protein Structure and Function*. Springer, New York.

Caldwell, R., Kongcharoen, J., Lin, Y. and Zhang, R. (2010) The Length Distributions of Non-Coding and Coding Sequences in Relation to Gene Expression: A Case Study on *Arabidopsis thaliana*. In Chang, R.*, et al.* (eds), *BIOCOMP 2010*. CSREA Press, Las Vegas, USA, pp. 127-133.

Caldwell, R., Lin, Y.-X. and Zhang, R. (2008) Correlation of length distributions between non-coding and coding sequences of *Arabidopsis thaliana*. In Chen, X.-w., Hu, X. and Kim, S. (eds), *Bioinformatics and Biomedicine*. IEEE, Philadelphia, USA, pp. 72-77.

Castillo-Davis, C., Mekhedov, S., Hartl, D., Koonin, E. and Kondrashov, F. (2002) Selection for short introns in highly expressed genes, *Nat Genet*, **34**, 415-418.

Celniker, S.E. and Rubin, G.M. (2003) THE DROSOPHILA MELANOGASTER GENOME, *Annual Review of Genomics and Human Genetics*, **4**, 89-117.

Chang, Y.-F., Imam, J.S. and Wilkinson, M.F. (2007) The Nonsense-Mediated Decay RNA Surveillance Pathway, *Annual Review of Biochemistry*, **76**, 51-74.

Chen, C.-H., Lin, H.-Y., Pan, C.-L. and Chen, F. (2011) The genomic features that affect the lengths of 5' untranslated regions in multicellular eukaryotes, *BMC Bioinformatics*, **12**, 1-8.

Chen, J.-b. and Ding, J.-j. (2008) A Review of Technologies on Quantile Regression *Statistics & Information Forum*, 89-96.

Cheng, P., He, Q., Yang, Y., Wang, L. and Liu, Y. (2003) Functional conservation of light, oxygen, or voltage domains in light sensing, *Proceedings of the National Academy of Sciences*, **100**, 5938-5943.

Chhabra, R., Kolli, S. and Bauer, J.H. (2013) Organically Grown Food Provides Health Benefits to *Drosophila melanogaster*, *PLoS One*, **8**, e52988.

Chiaromonte, F., Miller, W. and Bouhassira, Eric E. (2003) Gene Length and Proximity to Neighbors Affect Genome-Wide Expression Levels, *Genome Research*, **13**, 2602-2608.

Chintapalli, V.R., Wang, J. and Dow, J.A.T. (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease, *Nat Genet*, **39**, 715-720.

Chung, B., Simons, C., Firth, A., Brown, C. and Hellens, R. (2006) Effect of 5'UTR introns on gene expression in Arabidopsis thaliana, *BMC Genomics*, **7**, 120.

Chung, B.Y.W., Simons, C., Firth, A. E., Brown, C. M. & Hellens, R. P. (2006) Effect of 5' UTR introns on gene expression in *Arabidopsis thaliana*, *BMC Genomics*, **7**.

Clark, A.G. (2001) The Search for Meaning in Noncoding DNA, *Genome Research*, **11**, 1319-1320.

Claverie, J.O., H. (2003) The insertion of palindromic repeats in the evolution of proteins, *Trends in Biochemical Sciences*, **28**, 75-80.

Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H. and Stein, L. (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective, *Trends in Genetics*, **21**, 673-682.

Comeron, J.M. (2001) What controls the length of noncoding DNA?, *Current Opinion in Genetics & Development*, **11**, 652-659.

Consortium, M.G.S. (2002) Initial sequencing and comparative analysis of the mouse genome, *Nature*, **420**, 520-562.

Consortium, T.C.e.S. (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology, *Science*, **282**, 2012-2018.

Dai, Y., Zhang, R. and Lin, Y.-X. (2006) The Probability Distribution of Distance TSS-TLS Is Organism Characteristic and Can Be Used for Promoter Prediction. In Ali, M. and Dapoigny, R. (eds), *Advances in Applied Artificial Intelligence*. Springer Berlin Heidelberg, pp. 927-934.

Daniel, W.W. (1999) *Biostatistics: A Foundation for Analysis in the Health Sciences*. John Wiley & Sons, Inc., New York.

Davis, R.H. (2004) The age of model organisms, *Nat Rev Genet*, **5**, 69-76.

de Lanerolle, P. and Cole, A.B. (2002) *Cytoskeletal Proteins and Gene Regulation: Form, Function, and Signal Transduction in the Nucleus*.

Ding, H., Feng, P.-M., Chen, W. and Lin, H. (2014) Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, *Molecular BioSystems*, **10**, 2229-2235.

Dorairaj, J.J., Salzman, D.W., Wall, D., Rounds, T., Preskill, C., Sullivan, C.A., Lindner, R., Curran, C., Lezon-Geyda, K., McVeigh, T., Harris, L., Newell, J., Kerin, M.J., Wood, M., Miller, N. and Weidhaas, J.B. (2014) A germline mutation in the BRCA1 3’UTR predicts Stage IV breast cancer, *BMC Cancer*, **14**, 1-11.

Doran, G. (2008) The short and the long of UTRs, *J RNAi Gene Silencing*, **4**, 264-265.

Dujon, B. (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution, *Trends in genetics : TIG*, **22**, 375-387.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., de Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.-M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., de Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.-F., Straub, M.-L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P. and Souciet, J.-L. (2004) Genome evolution in yeasts, *Nature*, **430**, 35-44.

Duret, L. (2001) Why do genes have introns? Recombination might add a new piece to the puzzle, *Trends in Genetics*, **17**, 172-175.

Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, *Proceedings of the National Academy of Sciences*, **96**, 4482-4487.

Ehrnstorfer, I.A., Geertsma, E.R., Pardon, E., Steyaert, J. and Dutzler, R. (2014) Crystal structure of a SLC11 (NRAMP) transporter reveals the basis for transition-metal ion transport, *Nat Struct Mol Biol*, **21**, 990-996.

Eichler, E.E. and Sankoff, D. (2003) Structural Dynamics of Eukaryotic Chromosome Evolution, *Science*, **301**, 793-797.

Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact, *Trends in Genetics*, **19**, 362-365.

Espinosa, N., Hernández, R., López-Griego, L., Arroyo, R. and López-Villaseñor, I. (2001) Differences between coding and non-coding regions in the Trichomonas vaginalis genome: an actin gene as a locus model1, *Acta Tropica*, **78**, 147-154.

Fairbanks, D.J. and Rytting, B. (2001) Mendelian controversies: a botanical and historical review, *American Journal of Botany*, **88**, 737-752.

Fernández-Ayala, D.J.M., Jiménez-Gancedo, S., Guerra, I. and Navas, P. (2014) Invertebrate Models for Coenzyme Q10 Deficiency, *Molecular Syndromology*, **5** 170–179.

Ferrier, T., Matus, J.T., Jin, J. and Riechmann, J.L. (2011) Arabidopsis paves the way: genomic and network analyses in crops, *Current Opinion in Biotechnology*, **22**, 260-270.

Fickett, J.W. and Hatzigeorgiou, A.G. (1997) Eukaryotic Promoter Recognition, *Genome Research*, **7**, 861-878.

Fong, J., Murphy, T. and Pruitt, K. (2013) Comparison of RefSeq protein-coding regions in human and vertebrate genomes, *BMC Genomics*, **14**, 654.

Franca, L.T.C., Carrilho, E. and Kist, T.B.L. (2002) A review of DNA Sequencing techniques, *Quarterly Reviews of biophysics*, **35**, 169-200.

Frith, M.C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L. & Grimmond, S. M. (2006) The Abundance of Short Proteins in the Mammalian Proteome, *PLoS Genetics*, **2**, e52.

Fuglsang, A. (2005) Analysis of 5' UTR composition and gene expression: canonical versus non-canonical start codons, *Biochemical and Biophysical Research Communications*, **16**, 71-75.

Gan, Y., Guan, J. and Zhou, S. (2012) A comparison study on feature selection of DNA structural properties for promoter prediction, *BMC Bioinformatics* **13**, 1-12.

Garcia-Hernandez, M., Berardini, T., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L., Mundodi, S., Reiser, L., Rhee, S., Scholl, R., Tacklind, J., Weems, D., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2002) TAIR: a resource for integrated Arabidopsis data, *Funct Integr Genomics*, **2**, 239-253.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M.A., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M. and Barrell, B. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum, *Nature*, **419**, 10.1038/nature01097.

Gerstein, M. and Jansen, R. (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function?, *Current Opinion in Structural Biology*, **10**, 574-584.

Gilbert, W. (1991) Towards a paradigm shift in Biology, *Nature*, **349**, 99-99.

Golden, R. and Melov, S. (2007) Gene expression changes associated with aging in C. elegans. In Driscoll, M. and Patterson, G. (eds), *WormBook*. The C. elegans Research Community, WormBook.

González-Pérez, S., Gutiérrez, J., García-García, F., Osuna, D., Dopazo, J., Lorenzo, Ó., Revuelta, J.L. and Arellano, J.B. (2011) Early Transcriptional Defense Responses in Arabidopsis Cell Suspension Culture under High-Light Conditions, *Plant Physiology*, **156**, 1439-1456.

Gonzalez, N., Beemster, G.T.S. and Inzé, D. (2009) David and Goliath: what can the tiny weed Arabidopsis teach us to improve biomass production in crops?, *Current Opinion in Plant Biology*, **12**, 157-164.

Grisdale, C.J. and Fast, N.M. (2011) Patterns of 5' Untranslated Region Length Distribution in *Encephalitozoon cuniculi*: Implications for Gene Regulation and Potential Links Between Transcription and Splicing, *The Journal of Eukaryotic Microbiology*, **58**, 68-74.

Gumus, E., Kursun, O., Sertbas, A. and Ustek, D. (2012) Application of canonical correlation analysis for identifying viral integration preferences, *Bioinformatics*, **28**, 651-655.

Hahn, S. (2004) Structure and mechanism of the RNA polymerase II transcription machinery, *Nat Struct Mol Biol*, **11**, 394-403.

Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., Nishi, R., Ohashi, C., Iida, K., Tanaka, M., Horii, Y., Kawashima, M., Matsui, K., Toyoda, T., Shinozaki, K., Seki, M. and Matsui, M. (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes, *Proceedings of the National Academy of Sciences*, **110**, 2395-2400.

Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., Nishi, R., Ohashi, C., Iida, K., Tanaka, M., Horii, Y., Kawashima, M., Matsui, K., Toyoda, T., Shinozaki, K., Seki, M. and Matsuia,

M. (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes, *Proc Natl Acad Sci*, **110**, 2395–2400.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E.M., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J. and Sternberg, P.W. (2010) WormBase: a comprehensive resource for nematode research, *Nucleic Acids Research*, **38**, D463-D467.

He, X. and Zhang, J. (2005) Gene Complexity and Gene Duplicability, *Current Biology*, **15**, 1016-1021.

Hesketh, J. (2004) 3-Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein, *Biochemical Society Transactions*, **32**, 990-993.

Hillman, R.T., Green, R. and Brenner, S. (2004) An unappreciated role for RNA surveillance, *Genome Biology*, **5**, R8.

Hong, X., Scofield, D.G. and Lynch, M. (2006) Intron Size, Abundance, and Distribution within Untranslated Regions of Genes, *Molecular Biology and Evolution*, **23**, 2392-2404.

Hu, X., Bi, J., Loh, H.H. and Wei, L.-N. (2002) Regulation of Mouse κ Opioid Receptor Gene Expression by Different 3′-Untranslated Regions and the Effect of Retinoic Acid, *Molecular Pharmacology*, **62**, 881-887.

Huang, L., Guan, R.J. and Pardee, A.B. (1999) Evolution of Transcriptional Control from Prokaryotic Beginnings to Eukaryotic Complexities, **9**, 175-182.

Hughes, T.A. (2006) Regulation of gene expression by alternative untranslated regions, *Trends in Genetics*, **22**, 119-122.

Hughes, T.T., Allen, A.L., Bardin, J.E., Christian, M.N., Daimon, K., Dozier, K.D., Hansen, C.L., Holcomb, L.M. and Ahlander, J. (2012) Drosophila as a genetic model for studying pathogenic human viruses, *Virology*, **423**, 1-5.

Human Genome Sequencing, C. (2004) Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931-945.

IBM (2010) IBM SPSS Statistics for Windows, Version 19.0 IBM Corp., Armonk, NY.

Ingelbrecht, I.L., Herman, L.M., Dekeyser, R.A., Van Montagu, M.C. and Depicker, A.G. (1989) Different 3' end regions strongly influence the level of gene expression in plant cells, *The Plant Cell Online*, **1**, 671-680.

Ingvarsson, P.K. (2007) Gene Expression and Protein Length Influence Codon Usage and Rates of Sequence Evolution in Populus tremula, *Molecular Biology and Evolution*, **24**, 836-844.

Initiative, T.A.G. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, *Nature*, **408**, 796-815.

Jackson, R.J., Hellen, C.U.T. and Pestova, T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation, *Nat Rev Mol Cell Biol*, **11**, 113-127.

Jasny, B.R., Kelner, K.L. and Pennisi, E. (2008) From Genes to Social Behavior, *Science*, **322**, 891.

Jeffares, D.C., Mourier, T. and Penny, D. (2006) The biology of intron gain and loss, *Trends in genetics : TIG*, **22**, 16-22.

Jeong, S., Trotochaud, A.E. and Clark, S.E. (1999) The Arabidopsis CLAVATA2 Gene Encodes a Receptor-like Protein Required for the Stability of the CLAVATA1 Receptor-like Kinase, *The Plant Cell*, **11**, 1925-1933.

Jones, A.M., Chory, J., Dangl, J.L., Estelle, M., Jacobsen, S.E., Meyerowitz, E.M., Nordborg, M. and Weigel, D. (2008) The Impact of Arabidopsis on Human Health: Diversifying Our Portfolio, *Cell*, **133**, 939-943.

Kanhere, A. and Bansal, M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes, *Nucleic Acids Research*, **33**, 3165-3175.

Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J. and Vivares, C.P. (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi, *Nature*, **414**, 450-453.

Kawaguchi, R. and Bailey-Serres, J. (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis, *Nucleic Acids Research*, **33**, 955-965.

Kebaara, B.W. and Atkin, A.L. (2009) Long 3′-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in Saccharomyces cerevisiae, *Nucleic Acids Research*, **37**, 2771-2778.

Keeling, P., Leander, B.S. and Simpson, A. (2009) Eukaryotes. Eukaryota, Organisms with nucleated cells in The Tree of Life Web Project.

Kejnovsky, E., Leitch, I.J. and Leitch, A.R. (2009) Contrasting evolutionary dynamics between angiosperm and mammalian genomes, *Trends in Ecology & Evolution*, **24**, 572-582.

Kejnovsky, E., Leitch, I.J. and Leitch, A.R. (2009) Contrasting evolutionary dynamics between angiosperm mammalian genomes, *Trends in Ecology and Evolution*, **24**, 572-582.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses, *The Plant Journal*, **50**, 347-363.

Kim, Y., Lee, G., Jeon, E., Sohn, E.j., Lee, Y., Kang, H., Lee, D.w., Kim, D.H. and Hwang, I. (2014) The immediate upstream region of the 5′-UTR from the AUG start codon has a pronounced effect on the translational efficiency in Arabidopsis thaliana, *Nucleic Acids Research*, **42**, 485-498.

Kitajima, S. and Sato, F. (1999) Plant Pathogenesis-Related Proteins: Molecular Mechanisms of Gene Expression and Protein Function, *Journal of Biochemistry*, **125**, 1-8.

Koonin, E., Fedorova, N., Jackson, J., Jacobs, A., Krylov, D., Makarova, K., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Rogozin, I., Smirnov, S., Sorokin, A., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J.

and Natale, D. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes, *Genome Biology*, **5**, R7.

Kozak, M. (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles, *Microbiological Reviews*, **47**, 1-45.

Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs, *Nucleic Acids Research*, **15**, 8125-8148.

Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes, *Gene*, **361**, 13-37.

Kuile, B.H.S.F.J. (2000) The Length of the Combined 3' Untranslated Region and Poly(A) Tail Does Not Control Rates of Glyceraldehyde-3-Phosphate Dehydrogenase mRNA Translation in Three Species of Parasitic Protists, *Journal of Bacterology*, **182**, 3587-3589.

Landis, G., Shen, J. and Tower, J. (2012) Gene expression changes in response to aging compared to heat stress, oxidative stress and ionizing radiation in Drosophila melanogaster, *Aging (Albany NY)*, **4**, 768-789.

Lee, T.I. and Young, R.A. (2000) TRANSCRIPTION OF EUKARYOTIC PROTEIN-CODING GENES, *Annual Review of Genetics*, **34**, 77-137.

Lemos, B., Bettencourt, B.R., Meiklejohn, C.D. and Hartl, D.L. (2005) Evolution of Proteins and Gene Expression Levels are Coupled in Drosophila and are Independently Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions, *Molecular Biology and Evolution*, **22**, 1345-1354.

Li, L., Wang, D., Xue, M., Mi, X., Liang, Y. and Wang, P. (2014) 3[prime]UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network, *Sci. Rep.*, **4**.

Li, S., Feng, L., Niu, D. (2007) Selection for the miniaturization of highly expressed genes, *Biochemical and Biophysical Research*, **6**.

Lin, K.-S. and Chien, C.-F. (2009) Cluster analysis of genome-wide expression data for feature extraction, *Expert Systems with Applications*, **36**, 3327-3335.

Lin, K. and Zhang, D.-Y. (2005) The excess of 5′ introns in eukaryotic genomes, *Nucleic Acids Research*, **33**, 6522-6527.

Lin, Z. and Li, W.-H. (2012) Evolution of 5' Untranslated Region Length and Gene Expression Reprogramming in Yeasts, *Molecular Biology and Evolution*, **29**, 81-89.

Lipman, D., Souvorov, A., Koonin, E., Panchenko, A. and Tatusova, T. (2002) The relationship of protein conservation and sequence length, *BMC Evolutionary Biology*, **2**, 20.

Liu, C. and Mehdy, M.C. (2007) A Nonclassical Arabinogalactan Protein Gene Highly Expressed in Vascular Tissues, AGP31, Is Transcriptionally Repressed by Methyl Jasmonic Acid in Arabidopsis, *Plant Physiology*, **145**, 863-874.

Liu, H., Yin, J., Xiao, M., Gao, C., Mason, A.S. and Zhao, Z. (2012) Characterization and evolution of 5' and 3' untranslated regions in eukaryotes, *Gene*, **507**, 106-111.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J.M., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M. and Hyman, R.W. (2005) The Genome of the Basidiomycetous Yeast and Human Pathogen Cryptococcus neoformans, *Science*, **307**, 1321-1324.

Londei, P. (2005) Evolution of translational initiation: new insights from the archaea, *FEMS Microbiology Reviews*, **29**, 185-200.

LÓPEZ-LASTRA, M., RIVAS, A. and BARRÍA, M.I. (2005) Protein synthesis in eukaryotes: The growing biological relevance of cap-independent translation initiation, *Biological Research*, **38**, 121-146.

Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S.K. and Johnson, T.E. Transcriptional Profile of Aging in C. elegans, *Current Biology*, **12**, 1566-1573.

Lynch, M. (2007) *The Origins of Genome Architecture*. Sinauer Associates, Inc.

Lynch, M., Scofield, D.G. and Hong, X. (2005) The Evolution of Transcription-Initiation Sites, *Molecular Biology and Evolution*, **22**, 1137-1146.

Ma, J., Campbell, A. and Karlin, S. (2002) Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures, *Journal of Bacteriology*, **184**, 5733-5745.

Mabrouk, M.S., Solouma, N.H., Youssef, A.-B.M. and Kadah, Y.M. (2008) Eukaryotic Gene Prediction by an Investigation of Nonlinear Dynamical Modeling Techniques on EIIP Coded Sequences, *International Journal of Medicine and Medical Sciences*, **3**, 225-230.

Magwire, M.M., Yamamoto, A., Carbone, M.A., Roshina, N.V., Symonenko, A.V., Pasyukova, E.G., Morozova, T.V. and Mackay, T.F.C. (2010) Quantitative and Molecular Genetic Analyses of Mutations Increasing <italic>Drosophila</italic> Life Span, *PLoS Genet*, **6**, e1001037.

Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics, *Nucleic Acids Research*, **32**, D75-D77.

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J.R., Strelets, V.B., Wilson, R.J. and Consortium, F. (2013) FlyBase: improvements to the bibliography, *Nucleic Acids Research*, **41**, D751-D757.

Mayr, C. and Bartel, D.P. (2009) Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells, *Cell*, **138**, 673-684.

Mazumder, B., Seshadri, V. and Fox, P.L. (2003) Translational control by the 3′-UTR: the ends specify the means, *Trends in Biochemical Sciences*, **28**, 91-98.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M. (1998) Arabidopsis thaliana: A Model Plant for Genome Analysis, *Science*, **282**, 662-682.

Merritt, C., Rasoloson, D., Ko, D. and Seydoux, G. (2008) 3' UTRs are the primary regulators of gene expression in the *C. elegens* germline, *Curr Biol*, **18**, 1476-1482.

Mi, H., Vandergriff, J., Campbell, M., Narechania, A., Majoros, W., Lewis, S., Thomas, P.D. and Ashburner, M. (2003) Assessment of Genome-Wide Protein Function Classification for Drosophila melanogaster, *Genome Research*, **13**, 2118-2128.

Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) Untranslated regions of mRNAs, *Genome Biology*, **3**, reviews0004.0001 - reviews0004.0010.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O. and Lai, E.C. (2013) Widespread and extensive lengthening of 3′ UTRs in the mammalian brain, *Genome Research*, **23**, 812-825.

Monaghan, P. and Metcalfe, N.B. (2001) Genome size, longevity and development time in birds, *Trends in Genetics*, **17**, 568.

Murat, F., Peer, Y.V.d. and Salse, J. (2012) Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and processes, *Genome Biology Evolution*, **4**, 917-928.

Naora, H. and Montell, D.J. (2005) Ovarian Cancer Metastasis: Integrating insights from disparate model organisms, *Nat Rev Cancer*, **5**, 355-366.

Naylor, M.G., Lin, X., Weiss, S.T., Raby, B.A. and Lange, C. (2010) Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants, *PLoS One*, **5**, e10395.

Nilsen, T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell?, *BioEssays*, **25**, 1147-1149.

Nowrousian, M. (2010) Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems, *Eukaryotic Cell* **9**, 1300-1310

Ogata, H., Fujibuchi, W. and Kanehisa, M. (1996) The size differences among mammalian introns are due to the accumulation of small deletions, *FEBS Letters*, **390**, 99-103.

Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches, *Trends in Genetics*, **17**, 56-60.

Osada, Y., Saito, R. and Tomita, M. (1999) Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes, *Bioinformatics*, **15**, 578-581.

Ouzounis, C.A. (2012) Rise and Demise of Bioinformatics? Promise and Progress, *PLoS Comput Biol*, **8**, e1002487.

Ouzounis, C.A. and Valencia, A. (2003) Early bioinformatics: the birth of a discipline—a personal view, *Bioinformatics*, **19**, 2176-2190.

Panda, S., Hogenesch, J.B. and Kay, S.A. (2002) Circadian rhythms from flies to human, *Nature*, **417**, 329-335.

Pandey, S.P. and Krishnamachari, A. (2005) Computational analysis of plant RNA Pol-II promoters, *Biosystems*, **83**, 38-50.

Pandey, U.B. and Nichols, C.D. (2011) Human Disease Models in Drosophila melanogaster and the Role of the Fly in Therapeutic Drug Discovery, *Pharmacological Reviews*, **63**, 411-436.

Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1999) The biology of eukaryotic promoter prediction--a review, *Computers & chemistry*, **23**, 191-207.

Pesole, G., Grillo, G., Larizza, A. and Liuni, S. (2000) The untranslated regions of eukaryotic mRNAs: Structure, function, evolution and bioinformatic tools for their analysis, *Briefings in Bioinformatics*, **1**, 236-249.

Pestova, T.V., Kolupaeva, V.G., Lomakin, I.B., Pilipenko, E.V., Shatsky, I.N., Agol, V.I. and Hellen, C.U.T. (2001) Molecular mechanisms of translation initiation in eukaryotes, *Proceedings of the National Academy of Sciences*, **98**, 7029-7036.

Powell, J.A.S., Allen, J. and Sutter, N.B. (2010) DOG-SPOT database for comprehensive management of dog genetic research data, *Source Code for Biology and Medicine*, **5**, 10-10.

Preiss, T. and Hentze, M.W. (2003) Starting the protein synthesis machine: eukaryotic translation initiation, *BioEssays*, **25**, 1201-1211.

Preiss, T.H., M. W. (1998) Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast, *Nature*, **392**, 516-520.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D. and Ostell, J.M. (2014) RefSeq: an update on mammalian reference sequences, *Nucleic Acids Research*, **42**, D756-D763.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, **35**, D61-D65.

Qiu, P. (2003) Computational approaches for deciphering the transcriptional regulatory network by promoter analysis, *BIOSILICO*, **1**, 125-133.

Qiu, P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, *Biochemical and Biophysical Research Communications*, **309**, 495-501.

Raghava, G. and Han, J. (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein, *BMC Bioinformatics*, **6**, 59.

Rao, Y.S., Wang, Z.F., Chai, X.W., Wu, G.Z., Zhou, M., Nie, Q.H. and Zhang, X.Q. (2010) Selection for the compactness of highly expressed genes in *Gallus gallus*, *Biology Direct*, **5**.

Recipon, H. and Makalowski, W. (1997) The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives, *Current Opinion in Biotechnology*, **8**, 115-118.

Reese, M.G. (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome, *Computers & chemistry*, **26**, 51-56.

Ren, X.-Y., Vorst, O., Fiers, M.W.E.J., Stiekema, W.J. and Nap, J.-P. (2006) In plants, highly expressed genes are the least compact, *Trends in Genetics*, **22**, 528-532.

Reuter, M., Engelstädter, J., Fontanillas, P. and Hurst, L.D. (2008) A Test of the Null Model for 5′ UTR Evolution Based on GC Content, *Molecular Biology and Evolution*, **25**, 801-804.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, *Nucleic Acids Research*, **31**, 224.

Ribeiro, A.S., Häkkinen, A. and Lloyd-Price, J. (2012) Effects of gene length on the dynamics of gene expression, *Computational Biology and Chemistry*, **41**, 1-9.

Richards, C.L., Rosas, U., Banta, J., Bhambhra, N. and Purugganan, M.D. (2012) Genome-Wide Patterns of Arabidopsis Gene Expression in Nature, *PLoS Genetics*, **8**, 1-14.

Robinson, G.E., Fernald, R.D. and Clayton, D.F. (2008) Genes and Social Behavior, *Science*, **322**, 896-900.

Robinson, S.W., Herzyk, P., Dow, J.A.T. and Leader, D.P. (2012) FlyAtlas: Database of gene expression in the tissues of Drosophila melanogaster, *Nucleic Acids Research*, 1-7.

Rosenbaum, D.M., Rasmussen, S.G.F. and Kobilka, B.K. (2009) The structure and function of G-protein-coupled receptors, *Nature*, **459**, 356-363.

Roy, S.W. (2006) Intron-rich ancestors, *Trends in genetics : TIG*, **22**, 468-471.

Roy, S.W. and Gilbert, W. (2005) Rates of intron loss and gain: Implications for early eukaryotic evolution, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5773-5778.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Miklos, G.L.G., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S.B., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J.M., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., Zhong, F., Zhong, W., Gibbs, R., Venter, J.C., Adams, M.D. and Lewis, S. (2000) Comparative Genomics of the Eukaryotes, *Science*, **287**, 2204–2215.

Russell, P.J. (2002) *iGenetics*. Benjamin Cummings, San Francisco.

Sachs, A.B., Sarnow, P. & Hentze, W. M. (1997) Starting at the Beginning, Middle, and End: Translation Initiation in Eukaryotes, *Cell*, **89**, 831-838.

Sakharkar, M.K., Perumal, B.S., Sakharkar, K.R. and Kangueane, P. (2005) An Analysis on Gene Architecture in Human and Mouse Genomes, *In Silico Biology*, **5**, 347-365.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Díaz-Peredo, E., Sánchez-Solano, F., Santos-Zavaleta, A., Martínez-Flores, I., Jiménez-Jacinto, V., Bonavides-Martínez, C., Segura-Salazar, J., Martínez-Antonio, A. and Collado-Vides, J. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions, *Nucleic Acids Research*, **34**, D394-D397.

Schmid, M., Davison, T. S., Henz. S. R., Pape, U. J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. & Lohmann, J. U (2005) A gene expression map of *Arabidopsis thaliana* development, *Nature Genetics*, **37**, 501-506.

Schubert, I. (2007) Chromosome evolution, *Current Opinion in Plant Biology*, **10**, 109-115.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002) Functional Annotation of a Full-Length Arabidopsis cDNA Collection, *Science*, **296**, 141-145.

Serres-Giardi, L., Belkhir, K., David, J. and Glémin, S. (2012) Patterns and Evolution of Nucleotide Landscapes in Seed Plants, *The Plant Cell*, **24**, 1379-1397.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science, *Nat Rev Genet*, **advance online publication**.

Sherry, A. and Henson, R.K. (2005) Conducting and interpreting canonical correlation analysis in personality research: a user-friendly primer, *J Pers Assess*, **84**, 37-48.

Shine, J. and Dalgarno, L. (1974) The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites, *Proceedings of the National Academy of Sciences*, **71**, 1342-1346.

Singh, N., Davis, J. and Petrov, D. (2005) Codon Bias and Noncoding GC Content Correlate Negatively with Recombination Rate on the Drosophila X Chromosome, *J Mol Evol*, **61**, 315-324.

Skeeles, L.E., Fleming, J.L., Mahler, K.L. and Toland, A.E. (2013) The Impact of 3′UTR Variants on Differential Expression of Candidate Cancer Susceptibility Genes, *PLoS One*, **8**, e58609.

Smith, N.G.C. and Eyre-Walker, A. (2002) Adaptive protein evolution in Drosophila, *Nature*, **415**, 1022-1024.

Sorensen, J.G., Nielsen, M.M., Kruhoffer, M., Justesen, J. and Loeschcke, V. (2005) Full genome gene expression analysis of the heat stress response in Drosophila melanogaster, *Cell Stress & Chaperones*, **10**, 312-328.

SØRensen, J.G., Nielsen, M.M. and Loeschcke, V. (2007) Gene expression profile analysis of Drosophila melanogaster selected for resistance to environmental stressors, *Journal of Evolutionary Biology*, **20**, 1624-1636.

Stenoien, H.K. (2007) Compact genes are highly expressed in the moss *physcomitrella patens*, *Journal of Evolutionary Biology*, **20**, 1223-1229.

Stoesser, G., Baker, W., van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., Nardone, F., Stoehr, P., Tuli, M.A., Tzouvara, K. and Vaughan, R. (2003) The EMBL Nucleotide Sequence Database: major new developments, *Nucleic Acids Research*, **31**, 17-22.

Stoletzki, N. (2011) The surprising negative correlation of gene length and optimal codon use - disentangling translational selection from GC-biased gene conversion in yeast, *BMC Evolutionary Biology*, **11**, 93.

Stoltzfus, A. (2004) Molecular Evolution: Introns Fall into Place, *Current biology : CB*, **14**, R351-R352.

Strasser, B. (2010) Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965, *J Hist Biol*, **43**, 623-660.

Subramanian, S. and Kumar, S. (2004) Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome, *Genetics*, **168**, 373-381.

Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004, *Nucleic Acids Research*, **32**, D78-D81.

SY, R., W, B., TZ, B., G, C., D, D., A, D., M, G.-H., E, H., G, L., M, M., N, M., LA, M., S, M., L, R., J, T., DC, W., Y, W., I, X., D, Y., J, Y. and P, Z. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, *Nucleic Acids Research*, **31**, 224.

Tagu, D., Colbourne, J.K. and Nègre, N. (2014) Genomic data integration for ecological and evolutionary traits in non-model organisms, *BMC Genomics*, **15**, 490.

Takano, J., Wada, M., Ludewig, U., Schaaf, G., Wirén, N.v. and Fujiwara, T. (2006) The Arabidopsis Major Intrinsic Protein NIP5;1 Is Essential for Efficient Boron Uptake and Plant Development under Boron Limitation, *Plant Cell*, **18**, 1498-1509.

Tan, T., Frenkel, D., Gupta, V. and Deem, M.W. (2005) Length, protein–protein interactions, and complexity, *Physica A: Statistical Mechanics and its Applications*, **350**, 52-62.

Tang, C.S. and Ferreira, M.A.R. (2012) A gene-based test of association using canonical correlation analysis, *Bioinformatics*, **28**, 845-850.

Tanguay, R.L. and Gallie, D.R. (1996) The effect of the length of the 3'-untranslated region on expression in plants, *FEBS Letters*, **394**, 285-288.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

Terwilliger, N.B. (1998) Functional adaptations of oxygen-transport proteins, *Journal of Experimental Biology*, **201**, 1085-1098.

Thompson, B. (1980) Canonical Correlation: Recent Extensions for Modelling Educational Processes. *Annual Meeting of the American Educational Research Association* Boston, MA, pp. 36.

Tu, K., Yu, H. and Li, Y.-X. (2006) Combining gene expression profiles and protein–protein interaction data to infer gene functions, *Journal of Biotechnology*, **124**, 475-485.

Twyman, R. (2002) What are 'model organisms'?

Urrutia, A.O.H., L. D. (2003) The Signature of Selection Mediated by Expression on Human Genes, *Genome Research*, **13**, 2260-2264.

Vanneste, S. and Friml, J. (2009) Auxin: A Trigger for Change in Plant Development, *Cell*, **136**, 1005-1016.

Vilaprinyo, E., Alves, R. and Sorribas, A. (2010) Minimization of Biosynthetic Costs in Adaptive Gene Expression Responses of Yeast to Environmental Changes, *PLos Computational Biology*, **6**, 1-15.

Villarreal, D.O., Wise, M.C., Walters, J.N., Reuschel, E.L., Choi, M.J., Obeng-Adjei, N., Yan, J., Morrow, M.P. and Weiner, D.B. (2014) Alarmin IL-33 Acts as an Immunoadjuvant to Enhance Antigen-Specific Tumor Immunity, *Cancer Research*, **74**, 1789-1800.

Vinogradov, A.E. (2001) Within-intron correlation with base composition of adjacent exons in different genomes, *Gene*, **276**, 143-151.

Vinogradov, A.E. (2002) Growth and decline of introns, *Trends in genetics : TIG*, **18**, 232-236.

Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design?, *Trends in Genetics*, **20**, 248-253.

Walther, D., Brunnemann, R. and Selbig, J. (2007) The Regulatory Code for Transcriptional Response Diversity and Its Relation to Genome Structural Properties in <named-content xmlns:xlink="http://www.w3.org/1999/xlink" content-type="genus-species" xlink:type="simple">A. thaliana</named-content>, *PLoS Genet*, **3**, e11.

Wang, D., Hsieh, M. & Li, W. (2005) A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms, *Molecular Biology and Evolution*, **22**, 142-147.

Wang, G., Ellendorff, U., Kemp, B., Mansfield, J.W., Forsyth, A., Mitchell, K., Bastas, K., Liu, C.-M., Woods-Tör, A., Zipfel, C., de Wit, P.J.G.M., Jones, J.D.G., Tör, M. and Thomma, B.P.H.J. (2008) A Genome-Wide Functional Investigation into the Roles of Receptor-Like Proteins in Arabidopsis, *Plant Physiology*, **147**, 503-517.

Warringer, J. and Blomberg, A. (2006) Evolutionary constraints on yeast protein size, *BMC Evolutionary Biology*, **6**.

Webb, S. (2011) A decade after the genome, bioinformatics comes of age, *BioTechniques*, **51**, 157-161.

Westergard, L., Christensen, H.M. and Harris, D.A. (2007) The cellular prion protein (PrPC): Its physiological function and role in disease, *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1772**, 629-644.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M. and Schmitt, J. (2009) Effects of Genetic Perturbation on Seasonal Life History Plasticity, *Science*, **323**, 930-934.

Xu, L., Chen, H., Hu, X., Zhang, R., Zhang, Z. and Luo, Z.W. (2006) Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms, *Molecular Biology and Evolution*, **23**, 1107-1108.

Xu, X.M. and Møller, S.G. (2011) The value of Arabidopsis research in understanding human disease states, *Current Opinion in Biotechnology*, **22**, 300-307.

Xue-Franzén, Y. (2014) Does gene length play a role? — Transient regulation of Gcn5 histone acetyltransferase under stress conditions, *Genomics Data*, **2**, 293-295.

Yang, H. (2009) In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure, *Biology Direct*, **4**, 45.

Yi, G., Qu, L., Liu, J., Yan, Y., Xu, G. and Yang, N. (2014) Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing, *BMC Genomics*, **15**, 962.

Zeng, J., Zhu, S. and Yan, H. (2009) Towards accurate human promoter recognition: a review of currently used sequence features and classification methods, *Briefings in Bioinformatics*, **10**, 498-508.

Zhang, J. (2000) Protein-length distributions for the three domains of life, *Trends in Genetics*, **16**, 107-109.

Zhu, J. and Zhang, M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae, *Bioinformatics*, **15**, 607-611.

Zhu, J.H., F., Hu, S. & Yu, J. (2008) On the nature of human housekeeping genes, *Trends in Genetics*, **24**, 481-484.

Zylka, Mark J., Simon, Jeremy M. and Philpot, Benjamin D. (2015) Gene Length Matters in Neurons, *Neuron*, **86**, 353-355.

Appendix A – Visual Basic Scripts

NCBI Excel Macro

```
' FormatXls Macro
' Macro created 22/10/2007 by Rachel Caldwell
'
    Columns("H:K").Select
    Selection.Delete Shift:=xlToLeft
    Columns("A:A").Select
    Selection.Copy
    Range("H1").Select
    ActiveSheet.Paste
    Columns("A:A").Select
    Application.CutCopyMode = False
    Selection.Delete Shift:=xlToLeft
    Range("C1").Select
    Selection.EntireColumn.Insert
    Range("F1").Select
    Selection.EntireColumn.Insert
    Range("C1").Select
    ActiveCell.FormulaR1C1 = "CDS 1"
    Range("F1").Select
    ActiveCell.FormulaR1C1 = "CDS 2"
    Range("C2").Select
    ActiveCell.FormulaR1C1 = "=RC[-1]-RC[-2]+1"
    Range("F2").Select
    ActiveCell.FormulaR1C1 = "=RC[-1]*3+3"
    Range("C2").Select
    ActiveWindow.ScrollRow = 8
    Range("C2:C1495").Select
    Selection.FillDown
    Range("F2").Select
    Range("F2:F1495").Select
    Selection.FillDown
    Range("A1:I1").Select
    Selection.Font.Bold = True
    Cells.Select
    Cells.EntireColumn.AutoFit
    Range("A1").Select
End Sub
```

TAIR Database Cleanup Script

ORDER IN VBS SCRIPT TO RUN

1. Paragraph removal + addition of < to replace
2. Paragraph addition (adds a paragraph marker in replacement to >
3. Line removal and replacement - removes space|space and replaces with a *
4. Line removal 2 removes space, with just a *

TAIR 10 contains release of 27,416 protein coding genes.
CDS file = All A*rabidopsis* coding sequences including predicted sequences. Similar to the transcript file but lacking the 5' and 3' UTRs and no introns.

```
Dim strSearchString, objFSO, objFile
Const ForReading = 1
Const ForWriting = 2

' Removes Line Feed from Text File and replaces with comma

Set objFSO = CreateObject("Scripting.FileSystemObject")
Set objFile = objFSO.OpenTextFile("C:\temp\FlyBase_FastA CDS 20111029.txt", ForReading)
strSearchString = objFile.ReadAll
objFile.Close

Set objFile = objFSO.OpenTextFile("C:\temp\FlyBase_FastA CDS 20111029.txt", ForWriting)
objFile.Write Replace(strSearchString, VbLf, ",")
objFile.Close

' Removes comma from text file

Set objFSO = CreateObject("Scripting.FileSystemObject")
Set objFile = objFSO.OpenTextFile("C:\temp\FlyBase_FastA CDS 20111029.txt", ForReading)
strSearchString = objFile.ReadAll
objFile.Close

Set objFile = objFSO.OpenTextFile("C:\temp\FlyBase_FastA CDS 20111029.txt", ForWriting)
objFile.Write Replace(strSearchString, ",", "")
objFile.Close

' Adds line fields in where there are >

Set objFSO = CreateObject("Scripting.FileSystemObject")
Set objFile = objFSO.OpenTextFile("C:\temp\FlyBase_FastA CDS 20111029.txt", ForReading)
strSearchString = objFile.ReadAll
objFile.Close

Set objFile = objFSO.OpenTextFile("C:\Temp\FlyBase_FastA CDS 20111029.txt", ForWriting)
objFile.Write Replace(strSearchString, ">", VbCrLf)
objFile.Close
```

Appendix B – MS Excel Formulas & Macros

Protein Category script – MS Excel

Formula:

=VLOOKUP(B2,'Lookup Protein Table'!$A$2:$B$26,2)

Lookup Protein Table worksheet data

Code	Category	Description
[A]	1	RNA processing and modification
[B]	1	Chromatin structure and dynamics
[C]	3	Energy production and conversion
[D]	2	Cell cycle control, cell division, chromosome partitioning
[E]	3	Amino acid transport and metabolism
[F]	3	Nucleotide transport and metabolism
[G]	3	Carbohydrate transport and metabolism
[H]	3	Coenzyme transport and metabolism
[I]	3	Lipid transport and metabolism
[J]	1	Translation, ribosomal structure and biogenesis
[K]	1	Transcription
[L]	1	Replication, recombination and repair
[M]	2	Cell wall/membrane/envelope biogenesis
[N]	2	Cell motility
[O]	2	Posttranslational modification, protein turnover, chaperones
[P]	3	Inorganic ion transport and metabolism
[Q]	3	Secondary metabolites biosynthesis, transport and catabolism
[R]	4	General function prediction only
[S]	4	Function unknown
[T]	2	Signal transduction mechanisms
[U]	2	Intracellular trafficking, secretion, and vesicular transport
[V]	2	Defense mechanisms
[W]	2	Extracellular structures
[Y]	2	Nuclear structure
[Z]	2	Cytoskeleton

Appendix C – R and SPSS Codes

Canonical Correlation SPSS Syntax

```
INCLUDE 'C:/Program Files/IBM/SPSS/Statistics/19/Samples/English/Canonical correlation.sps'.
CANCORR SET1=Cold, Constant30, ControlLine, Desiccation, Heat, KnockDown, Longevity, Starvation /
SET2=D1, D2, D3 / .

# Note if it errors you must clear all the windows before you can proceed with running the macro again.
```

Running Canonical Correlation Analysis in R

Make sure there is only 12 columns in the file – will error.

Read the file (must be csv file)

```
> mm<-read.table("D melanogaster Gene Expression Canonical Analysis small sample.csv", sep = ",", header = TRUE)

> library(fields)
```

To run stats on file:

```
> t(stats(mm))

# define the two sets of variables
> GeneLength<-mm[,2:4]
> GeneExp<-mm[,5:11]

# correlations

> library(CCA)
> matcor(GeneLength,GeneExp)
```

R Canonical Correlation Analysis

```
> cc1<-cc(GeneLength,GeneExp)
# display the canonical correlations
> cc1[1]
# raw canonical coefficients
> cc1[3:4]
# compute canonical loadings
>cc2<-comput(GeneLength, GeneExp, cc1)

#Display canonical loadings
>cc2[3:6]
> plot(cca.fit)
```

Quantile Regression in R

```
library(quantreg)
CdsData<-read.csv("DmelCDS.csv", strip.white=TRUE)

CDS<-CdsData[,1]
Control<-CdsData[,2]

Lcontrol<-log(Control)
LCDS<-log(CDS)
LCDS2<- LCDS^2
plot(LCDS, Lcontrol, cex = 0.05, type = "n",
     xlab = "LCDS", ylab = "Lcon")
points(LCDS, Lcontrol, cex = 0.05, col = "blue")
#plot(LCDS, Lcontrol, cex = 0.5, col = "blue")
#abline(rq(Lcontrol ~ LCDS + LCDS2, tau=0.9), col="red")

a0<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[1,1]
a1<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[2,1]
a2<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[3,1]
A<-a0+a1*LCDS+a2*LCDS2
points(LCDS, A,cex = 0.05,col="red")
a0
a1
a2
taus<-c(0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1)
 coeA1<-c()
for (i in 1:length(taus)) {
  a0<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[1,1]
  a1<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[2,1]
  A<-a0+a1*LCDS
  coeA1[i]<-a1
  points(LCDS, A, cex = 0.05, col= "red")
}
coeA
```

```
> library(quantreg)
real<-read.csv("AthFiveUTR.csv",strip.white=TRUE)
FiveUTR<-real[,1]
Control<-real[,2]
LControl<-log(Control)
LFiveUTR<-log(FiveUTR)
plot(LFiveUTR,LControl,cex=.5,type="p",col="black",xlab="Log 5' UTR Sequence Length bp",ylab="Log Gene Expression Intensity")
taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
f<-rq(LControl~LFiveUTR,tau=taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
summary(f)
quantreg.plot<-summary(f)
plot(quantreg.plot)
qr10<-rq(LControl~LFiveUTR,tau=0.1)
anova(qr10,qr20)
```

```
Quantile Regression Analysis of Deviance Table

Model: LControl ~ LFiveUTR
Joint Test of Equality of Slopes: tau in (Raghava and Han)

  Df Resid Df F value  Pr(>F)
1  1    36889  5.2053 0.02252 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
real<-read.csv("AthCDS.csv",strip.white=TRUE)
CDS<-real[,1]
Control<-real[,2]
LControl<-log(Control)
LCDS<-log(CDS)
plot(LCDS,LControl,cex=.5,type="p",col="black",xlab="Log CDS Sequence Length bp",ylab="Log Gene Expression Intensity")
taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
f<-rq(LControl~LCDS,tau=taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
quantreg.plot<-summary(f)
plot(quantreg.plot)
summary(f)
```

```
#TABLE of results:
fit2<-summary(rq(LControl~LCDS, tau=c(.1,.2,.3,.4,.5,.6,.7,.8,.9)))
 latex(fit2, caption="Arabidopsis thalania", transpose = TRUE)
 real<-read.csv("AthThreeUTR.csv",strip.white=TRUE)
 ThreeUTR<-real[,1]
 Control<-real[,2]
 LControl<-log(Control)
 LThreeUTR<-log(ThreeUTR)
 plot(LThreeUTR,LControl,cex=.5,type="p",col="black",xlab="Log 3' UTR Sequence Length bp",ylab="Log Gene Expression Intensity")
 taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
 f<-rq(LControl~LThreeUTR,tau=taus)
 for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
quantreg.plot<-summary(f)
plot(quantreg.plot)
summary(f)

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
real<-read.csv("DmelFiveUTR.csv",strip.white=TRUE)
FiveUTR<-real[,1]
Control<-real[,2]
LControl<-log(Control)
LFiveUTR<-log(FiveUTR)
plot(LFiveUTR,LControl,cex=.5,type="p",col="black",xlab="Log 5' UTR Sequence Length bp",ylab="Log Gene Expression Intensity")
taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
f<-rq(LControl~LFiveUTR,tau=taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
quantreg.plot<-summary(f)
plot(quantreg.plot)
summary(f)

real<-read.csv("DmelCDS.csv",strip.white=TRUE)
CDS<-real[,1]
Control<-real[,2]
LControl<-log(Control)
LCDS<-log(CDS)
plot(LCDS,LControl,cex=.5,type="p",col="black",xlab="Log CDS Sequence Length bp",ylab="Log Gene Expression Intensity")
taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
```

```
f<-rq(LControl~LCDS,tau=taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
quantreg.plot<-summary(f)
plot(quantreg.plot)
summary(f)

real<-read.csv("DmelThreeUTR.csv",strip.white=TRUE)
ThreeUTR<-real[,1]
Control<-real[,2]
LControl<-log(Control)
LThreeUTR<-log(ThreeUTR)
plot(LThreeUTR,LControl,cex=.5,type="p",col="black",xlab="Log 3' UTR Sequence Length bp",ylab="Log Gene Expression Intensity")
taus<-c(.1,.2,.3,.4,.5,.6,.7,.8,.9)
f<-rq(LControl~LThreeUTR,tau=taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="red")}
quantreg.plot<-summary(f)
plot(quantreg.plot)
summary(f)
```

Non-Linear Model – Quantile Regression

```
> a0
[1] 14.90438342
> a1
[1] -1.674027926
> a2
[1] 0.07782366078
>
> taus<-c(0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1)
>  coeA1<-c()
> for (i in 1:length(taus)) {
+   a0<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[1,1]
+   a1<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[2,1]
+   A<-a0+a1*LCDS
+   coeA1[i]<-a1
+   points(LCDS, A, cex = 0.05, col= "red")
+ }
> coeA1
[1] -0.4710644187 -0.4400077846 -0.4008260150 -0.3654767826
[5] -0.3600360019 -0.3359369747 -0.2719822981 -0.1978264703
```

```
library(quantreg)
CdsData<-read.csv("DmelCDS.csv", strip.white=TRUE)

CDS<-CdsData[,1]
Control<-CdsData[,2]

Lcontrol<-log(Control)
LCDS<-log(CDS)
LCDS2<- LCDS^2
plot(LCDS, Lcontrol, cex = 0.05, type = "n",
    xlab = "LCDS", ylab = "Lcon")
points(LCDS, Lcontrol, cex = 0.05, col = "blue")
#plot(LCDS, Lcontrol, cex = 0.5, col = "blue")
#abline(rq(Lcontrol ~ LCDS + LCDS2, tau=0.9), col="red")

a0<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[1,1]
a1<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[2,1]
a2<-summary(rq(Lcontrol ~ LCDS+LCDS2, tau = 0.9))$coefficient[3,1]
A<-a0+a1*LCDS+a2*LCDS2
points(LCDS, A,cex = 0.05,col="red")
a0
a1
a2
taus<-c(0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1)
 coeA1<-c()
for (i in 1:length(taus)) {
  a0<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[1,1]
  a1<-summary(rq(Lcontrol ~ LCDS, tau = taus[i]))$coefficient[2,1]
  A<-a0+a1*LCDS
  coeA1[i]<-a1
  points(LCDS, A, cex = 0.05, col= "red")
}
coeA1
```

```
chr<-read.csv("Ath.csv", header=TRUE, sep=",",fill=TRUE)
d1<-chr$d1
d2<-chr$d2
d3<-chr$d3
inten<-chr$inten
D1<-chr$D1
D2<-chr$D2
D3<-chr$D3
q3<-chr$q3
cate<-chr$cate

cor.test(d1,d2,method="spearman")
cor.test(d1,d3,method="spearman")
cor.test(d2,d3,method="spearman")



boxplot(inten~q3,ylim=c(0,30000),xlab="Quantile",ylab="gene expression intensity")

boxplot(d1,d2,d3)



 boxplot(len ~ dose, data = ToothGrowth,
             boxwex = 0.25, at = 1:3 - 0.2,



y<-inten
x<-d3
 library(MASS)
   lqsmodel1 <- lqs(y~x, method="lts")
   plot(x,y)
    abline(lqsmodel1,col=3)

quantile(d1,prob=c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1),na.ram=T)
bre<-c(135,525,741,885,1029,1155,1284,1431,1614,2090,10725)
table(cut(d1,bre,right=F))

#quantile(d2,prob=c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1),na.ram=T)
#bre<-c(1,51,71,85,102,121,144,182,228,314,3209)
#table(cut(d2,bre,right=F))
```

```
#quantile(d3,prob=c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1),na.ram=T)
#bre<-c(2,147,180,206,229,253,282,314,363,463,2016)
#table(cut(d3,bre,right=F))

#quantile(d1)
#bre<-c(135,816,1155,10725)
#table(cut(d1,bre,right=F))

#library(fBasics)
#kurtosis(d1)
#skewness(d1)
#kurtosis(d2)
#skewness(d2)
#kurtosis(d3)
#skewness(d3)
#kurtosis(inten)
#skewness(inten)


#plot(inten~d1,xlim=c(0,6000))

#par(mfrow = c(1,2))
#qqnorm(d1,xlab="d1",ylab="Length")
#qqnorm(inten,xlab="inten",ylab="Length")


#hist(d1,xlim=c(0,4000),xlab="length of coding region without intron")
#hist(d2,xlim=c(0,1000),xlab="length of noncoding region (5'UTR) without intron")
#hist(d3,xlim=c(0,1100),xlab="length of noncoding region (3'UTR) without intron")
#hist(inten,xlim=c(0,40000),xlab="average gene expression intensity")


library(quantreg)

        plot(d1,inten,panel.first = grid(8,8),pch = 1, cex = 1.2,xlab="length of coding region without intron",ylab="average gene expression intensity",col="grey")

        #plot(d3,inten,panel.first = grid(8,8),pch = 1, cex = 1.2,xlab="length of noncoding region (3'UTR) without intron",ylab="average gene expression intensity",col="grey",xlim=c(0,1800))
```

```
	#plot(d2,inten,panel.first = grid(8,8),pch = 1, cex = 1.2,xlab="length of noncoding region (5'UTR) without intron",ylab="average gene expression intensity",col="grey",xlim=c(0,1000))


tau_set<-seq(0.30,0.70,0.10)

for (tau_value in tau_set)
{
	#####################################
	##### estimate inten~d1+d1^2 quantile regression
	#####################################

	#print(fit.ml<-lm(inten~d3+I(d3^2)))
	#print(fit.ml.summary<-summary(fit.ml))
	#print(AIC(fit.ml))



	print(fit.l<-rq(inten~d1+I(d1^2),tau=tau_value))
	print(fit.l.summary<-summary(fit.l,se="iid"))
	#print(AIC(fit.l))

	#print(fit.l1<-rq(inten~d3+I(d3^2),tau=tau_value))
	#print(fit.l1.summary<-summary(fit.l1,se="iid"))
	#print(AIC(fit.l1))

	#print(fit.l2<-rq(inten~d2,tau=tau_value))
	#print(fit.l2.summary<-summary(fit.l2,se="iid"))
	#print(AIC(fit.l2))

	#fit.ml.value <-fit.ml$coef[1] + fit.ml$coef[2] * d3 + fit.ml$coef[3] * d3^2

	fit.l.value <-fit.l$coef[1] + fit.l$coef[2] * d1 + fit.l$coef[3] * d1^2

	#fit.l1.value <-fit.l1$coef[1] + fit.l1$coef[2] * d3 + fit.l1$coef[3] * d3^2

	#fit.l2.value <-fit.l2$coef[1] + fit.l2$coef[2] * d2

	lines(d1,fit.l.value,col="brown")
	#lines(d3,fit.l1.value,col="blue")
	#lines(d3,fit.ml.value,col="yellow")
```

```
        #lines(d2,fit.l2.value,col="red")
}

plot(summary(rq(inten~d1+I(d1^2),tau=tau_set,data=chr)), parm=1,mar=c(5,5,4,2)+0.2
,ylab="Intercept",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2),tau=tau_set,data=chr)), parm=2,mar=c(5,5,4,2)+0.2
,ylab="The length of coding region",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2),tau=tau_set,data=chr)), parm=3,mar=c(5,5,4,2)+0.2
,ylab="The length of coding region square",xlab="Quantile")



library(quantreg)

tau_set <- seq(0.3,0.7,0.1)



for (tau_value in tau_set)
{

        #####################################
        ##### estimate inten~d1+d2+d3 quantile regression
        #####################################

        print(fit.poly<-rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_value))
        print(fit.poly.summary<-summary(fit.poly,se="iid"))
        print(AIC(fit.poly))
}
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=1,mar=c(5,5,4,2)+0.2
,ylab="Intercept",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=2,mar=c(5,5,4,2)+0.2
,ylab="The length of coding region",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=3,mar=c(5,5,4,2)+0.2
,ylab="The length of coding region square",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=4,mar=c(5,5,4,2)+0.2
,ylab="The length of 5'UTR region",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=5,mar=c(5,5,4,2)+0.2
,ylab="The length of 3'UTR region",xlab="Quantile")
plot(summary(rq(inten~d1+I(d1^2)+d2+d3+I(d3^2),tau=tau_set,data=chr)), parm=6,mar=c(5,5,4,2)+0.2
,ylab="The length of 3'UTR region square",xlab="Quantile")
```

Normalisation of Expression data using MAS5 in R

```
local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
if(nchar(pkg)) library(pkg, character.only=TRUE)})
affy.data = ReadAffy()
eset.mas5 = mas5(affy.data)
exprSet.nologs = exprs(eset.mas5)
colnames(exprSet.nologs)
write.table(exprSet, file="DmResults.txt", quote=F, sep="\t")
data.mas5calls = mas5calls(affy.data)
data.mas5calls.calls = exprs(data.mas5calls)
write.table(data.mas5calls.calls, file="Dmcalls.txt", quote=F, sep="\t")
write.table(exprSet, file="DmResults.txt", quote=F, sep="\t")
exprSet = log(exprSet.nologs, 2)
write.table(exprSet, file="DmResults.txt", quote=F, sep="\t")
local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
if(nchar(pkg)) library(pkg, character.only=TRUE)})
eset.rma = JustRMA()
local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
if(nchar(pkg)) library(pkg, character.only=TRUE)})
library(made4)
Overview(eset)
overview(eset)
overview(eset.mas5)
```