2017

Comparative Study of the Distribution of Repetitive DNA in Model Organisms

Mohamed K. Aburweis
South Dakota State University

COMPARATIVE STUDY OF THE DISTRIBUTION OF REPETITIVE DNA IN MODEL ORGANISMS

BY

MOHAMED K. ABURWEIS

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science and Statistics

South Dakota State University

2017

COMPARATIVE STUDY OF THE DISTRIBUTION OF REPETITIVE DNA IN MODEL ORGANISMS

MOHAMED K. ABURWEIS

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Computational Science and Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Xijin Ge, Ph.D.
Dissertation Advisor Date

Kurt Cogswell, Ph.D.
Head, Mathematics and Statistics Date

Dean, Graduate School Date

This dissertation is dedicated to the memory of my parents,

and to my family, Yumna, Adam, Sara, and Omar

ACKNOWLEDGEMENTS

First and foremost, I offer my profuse thanks to the Almighty Allah for enabling me to finish this work. Praise and thanks to Allah.

I would like to express my sincere gratitude and thanks to my advisor, mentor and committee chair, Dr. Xijin Ge, for his suggestion of the field of study. His constant help, encouragement, fruitful discussions, and advice have been of great help. I am indebted to him for more than he knows.

I would gratefully acknowledge Dr. Kurt Cogswell, Head of the Mathematics and Statistics Department, for providing me with a graduate teaching assistantship during my graduate studies at South Dakota State University.

I also express my sincere appreciation to all my committee members, Dr. Gary Hatfield, Dr. Gemechis Djira, and Dr. Nancy Lyons, for reviewing and providing feedback during the period of my research.

My sincere thanks to my family, especially to my wife Yumna, who has shown me patience, love, endurance, and encouragement that allowed me to reach my dreams. I thank my wonderful children Adam, Sara, and our new addition, Omar. I would also like to acknowledge my brother, and my sister for their full understanding and constant encouragement during the period of my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

bp	Base pair
CV	Cross-Validation
GB	Gigabyte (1,024 megabytes)
GC	Guanine-Cytosine content
LASSO	Least Absolute Shrinkage and Selection Operator
LINEs	Long Interspersed Nuclear Elements
LTRs	Long Terminal Repeats
MIR	Mammalian-wide Interspersed Repeats
MLR	Multiple Linear Regression
MMLR	Multivariate Multiple Linear Regression
NGS	Next Generation Sequencing
OLS	Ordinary Least Squares
PLS	Penalized Least Squares
RNA-seq	RNA-Sequencing
SINEs	Short Interspersed Nuclear Elements
TB	Terabyte (1,024 Gigabytes)
TEs	Transposable Elements
TSS	Transcription Starting Site
UTRs	Untranslated Regions
VNTRs	Variable Number Tandem Repeats

ABSTRACT

COMPARATIVE STUDY OF THE DISTRIBUTION OF REPETITIVE DNA IN MODEL ORGANISMS

MOHAMED K. ABURWEIS

2017

Repetitive DNA elements are abundant in the genome of a wide range of organisms. In mammals, repetitive elements comprise about 40-50% of the total genomes. However, their biological functions remain largely unknown. Analysis of their abundance and distribution may shed some light on how they affect genome structure, function, and evolution.

We conducted a detailed comparative analysis of repetitive DNA elements across ten different eukaryotic organisms, including chicken (G. gallus), zebrafish (D. rerio), Fugu (T. rubripes), fruit fly (D. melanogaster), and nematode worm (C. elegans), along with five mammalian organisms: human (H. sapiens), mouse (M. musculus), cow (B. taurus), rat (R. norvegicus), and rhesus (M. mulatta). Our results show that repetitive DNA content varies widely, from 7.3% in the Fugu genome to 52% in the zebrafish, based on RepeatMasker data. The most frequently observed transposable elements (TEs) in mammals are SINEs (Short Interspersed Nuclear Elements), followed by LINEs (Long Interspersed Nuclear Elements). In contrast, LINEs, DNA transposons, simple repeats, and low complexity repeats are the most frequently observed repeat classes in the chicken, zebrafish, fruit fly, and nematode worm genomes, respectively. LTRs (Long Terminal Repeats) have significant genomic coverage and diversity, which may make them suitable for regulatory

roles. With the exception of the nematode worm and fruit fly, the frequency of the repetitive elements follows a log-normal distribution, characterized by a few highly prevalent repeats in each organism. In mammals, SINEs are enriched near genic regions, and LINEs are often found away from genes. We also identified many LTRs that are specifically enriched in promoter regions, some with a strong bias towards the same strand as the nearby gene. This raises the possibility that the LTRs may play a regulatory role. Surprisingly, most intronic repeats, with the exception of DNA transposons, have a strong tendency to be on the opposite DNA strand as the host gene. One possible explanation is that intronic RNAs which result from splicing may contribute to retrotransposition to the original intronic loci.

Moreover, our observations of repetitive DNA elements enrichment near genic regions and, specifically, the promoter region of genes, raise the question as to whether repetitive DNA elements have a significant impact on gene expression in both human and mouse genomes. In order to investigate the impact of these repeats on gene expression, we calculate the total number of base pairs (bp) for these repeats in two different locations upstream from the genes — namely, the 2kbp and 20kbp promoter regions. In addition to that, we quantified the gene expression levels in both human and mouse tissues using RNA-seq analysis. Then, we used different statistical modeling approaches to investigate the association between repetitive DNA elements and gene expression in two different promoter regions. Although most transposable elements are primarily involved in reduced gene expression, our model's results showed that Alu elements in both human and mouse are significantly associated with higher average expression in the promoter region. Furthermore, we found that the B2 in both mouse 2kbp and 20kbp and hAT.Charlie elements in the human 20kbp, are also significantly associated with up-regulated gene

expression in the 2kpb promoter. In addition to Alu and B2 in 2kbp, we found that the ERV1 have a significant association with higher average expression in the 20kbp promoter in mouse tissues. We also found that L1 and Simple_repeat elements are significantly associated with lower average expression in both human and mouse tissues. Furthermore, in the human, we found that the MIR is also associated with lower average expression. The effects of Alu elements in both human and mouse are stronger at 2kbp than at 20kbp. In contrast, the L1 effect at 20kbp is stronger than at 2kbp.

Our results indicate that comparative studies of repetitive DNA elements in multiple organisms can provide insights into their evolution and expansion, and lead to the elucidation of their potential functions. The non-random distribution of repeats across multiple organisms adds to the existing evidence that some repetitive DNA elements are drivers of genome evolution, rather than just “junk” DNA.

CHAPTER 1 - INTRODUCTION AND OVERVIEW

1.1 Introduction to Repetitive DNA

All living organism genomes contain both unique and repetitive DNA sequences [1]. A unique DNA sequence is a fragment of DNA present as only a single copy in a cell, [2] whereas a repetitive DNA sequence (repetitive elements, repetitive sequences, DNA repeats) is a stretch of DNA that is repeated many times in the genomes.

DNA reannealing studies in the 1960s revealed that eukaryotic genomes comprise a highly variable fraction of repetitive DNA [3]. Repetitive DNA was first recognized as a significant constituent of the eukaryotic genome [4]. Results from a series of rate renaturation experiments conducted by Britten and Kohne suggested that the repetitive content is roughly proportional to the genetic complexity [5]. Although repetitive DNA was earlier considered to be 'junk' or 'selfish' DNA that had no impact on gene expression and genome stability, recent studies have shown that the complexity of living organisms is not only caused by coding sequences. The purpose of repetitive DNA which does not encode proteins and their biological functions remain largely unclear may also play a significant role in the gene regulation [6].

These repeats are abundant in the genome of a wide range of organisms [7] and comprise up to 50% or more of an organism's DNA. More recent studies show the percentage of repetitive DNA elements are as high as two-thirds of the human genome [8]. On the other hand, vertebrate, insect, and nematode genomes vary widely in size and the amount of repetitive DNA. For example, the repetitive DNA of zebrafish (D. rerio) comprises about 52% of its genome, while in chicken (G. gallus) and Fugu (T. rubripes)

repetitive DNA comprise approximately 11% [9] and less than 10% [10], respectively. Repetitive DNA elements vary in their length and range, from a few base pairs (bp) such as microsatellites to several kilobase pairs (kbp), such as LINE1 (6 kbp) [11].

1.2 Classification of Repetitive DNA

Repetitive DNA elements are classified into two major groups based on their degree of repetitiveness, highly repetitive or moderately repetitive. Then they are grouped based on their organization and their functions into tandem repeats or dispersed (interspersed) repeats [1, 7] as shown in Figure 1.1.

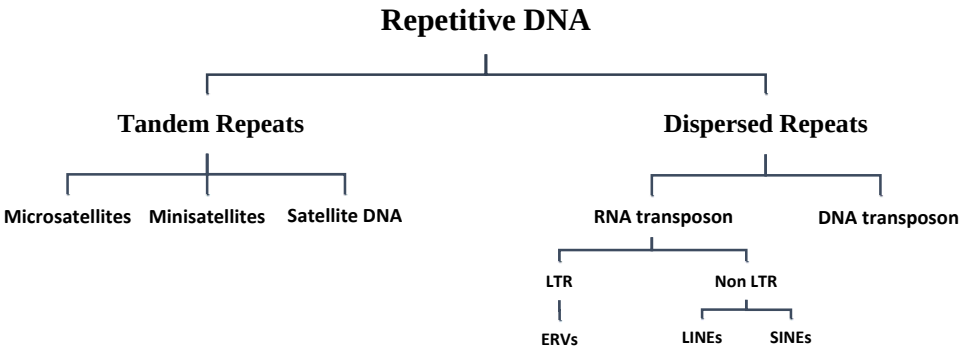

Figure 1.1: Schematic diagram of repetitive DNA classification.

1.2.1 Tandem Repeats

Tandem repeats are made of short (≥ 2bp in length) non-coding consecutive sequences, with their sequence units organized in a head to tail orientation [12]. They are the common feature of eukaryotic genomes, but are found much less frequently in prokaryotes. Tandem repeats include three subclasses: microsatellites, minisatellites, and satellite DNA; the last type is mostly found in the heterochromatin areas such as the centromeres and telomeres. Tandem repeats can be classified based on their copy number of the basic repeat units, length, and genomic location, as follows:

1.2.1.1 Microsatellites

Microsatellites, also known as short tandem repeats (STRs) by forensic geneticists or as simple sequence repeats (SSRs) by plant geneticists, consist of very small sequences ranging in length from 1 to 6 base pairs (bp) repeated 10 to100 times [13, 14]. They are distributed throughout non-coding and coding regions, including regulatory sequences of prokaryotic and eukaryotic genomes [15]. Microsatellites are useful for forensics, DNA fingerprinting, and paternity testing, because the number of repeats for a given microsatellite may differ between individuals. They are also classified as Variable Number Tandem Repeats (VNTRs) and usually made up of dinucleotide microsatellite (i.e., TATATATATA), trinucleotide (i.e., GTCGTCGTCGTCGTC), or tetranucleotide. In many organisms, dinucleotides are the premier type of microsatellite. Additional repeat units that are used for transcriptome analysis and fingerprinting are (AT)n, (GAA)n, (TCC)n, (GGAT)n, (GGCA)n, and (TTAGGG)n.

The proportion of microsatellite sequences within genomes tends to increase from invertebrates to vertebrates. For example, they comprise about 0.21% in C. elegans and 3% in human genomes [16]. The high rate of mutation of these repeats implies involvement in the regulation of gene expression which leads to phenotype changes and diseases. In human, for example, trinucleotide microsatellite sequences have been associated with several severe disorders, such as *Fragile X syndrome* and *Huntington's disease* [17].

1.2.1.2 Minisatellites

Minisatellites, also referred to as Variable Number Tandem Repeat (VNTRs), are tandemly repeated sequences of DNA composed of short repeat units ranging from 10 to 60 bp, with a total length of less than 1kbp to 15kbp. They are enriched in subtelomeric regions of chromosomes [7, 18]. Minisatellites were first described by Alec Jeffrey in 1985, based on the intronic regions of the human myoglobin gene [19]. Since then, many organism genomes have been reported with similar DNA structures. One of the minisatellites subsets comprises the highly polymorphic arrays of short tandem repeats with an unknown function, which are used as useful DNA markers [20]. Most of the minisatellites repeat are GC rich.

1.2.1.3 Satellite

Satellite DNAs are highly repetitive non-coding sequences composed of repeat units ranging from 5 to 200 bp in length and organized in long head to tail arrays comprising blocks hundreds of kilobases long. They are the primary component of heterochromatin and enriched in subtelomeric regions of chromosomes.

Early studies of satellite DNA's functional role considered them junk DNA. In contrast, recent studies have shown many functions for them, such as establishing and maintaining of the chromatin states by promoting heterochromatin assembly, influencing gene expression and contributing to the epigenetic regulatory process [1]. Satellite DNAs constitute 4.17% and higher proportions of some insect and rodent genomes [6, 21]. Several satellite DNAs families are present in each organism. For example, approximately nine families are found in the human genome, with the most abundant family being the centromeric α satellite DNA, which comprises more than half of the total satellite DNA content in the genome [22].

1.2.2 Dispersed (Interspersed) Repeats

Dispersed repeats, also known as transposable elements (TEs) or mobile elements, are identical or nearly identical DNA sequences [7] scattered within the genome. These have arisen due to transposition, having "capability to jump or switch from one locus to another in the genome"[23]. Barbara McClintock first discovered TEs in her study of corn (maize) genomes in1940s [24]. They have been found in many organisms. TEs are highly abundant in some genomes, accounting for approximately 45% of the human genome (Figure 1.2) [25] and around 85% of the maize genome [26]. TEs can both positively and negatively affect a genome; their mobilization can regulate gene expression, promote gene inactivation, or motivate illegitimate recombination. TEs are classified based on their transposition methods into class I transposable elements, also referred to as RNA transposons (retrotransposons), and class II transposable elements, called DNA transposons (Figure 1.3).

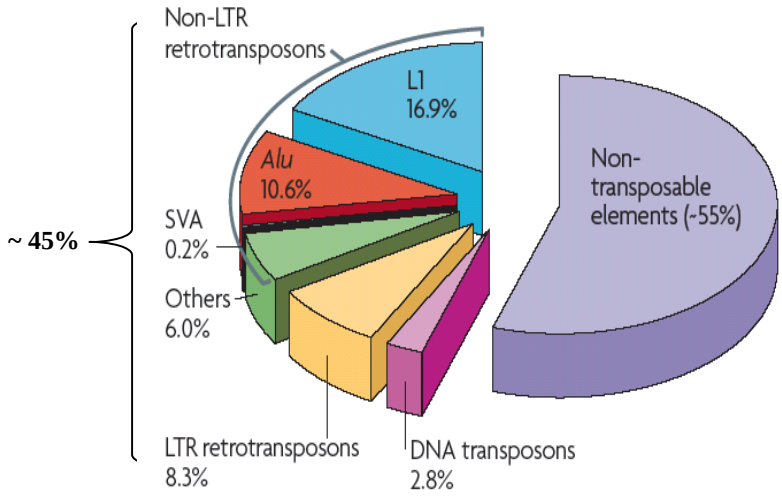

Figure 1.2: The transposable elements in the human genome (Cordaux R, and Batzer MA 2009).

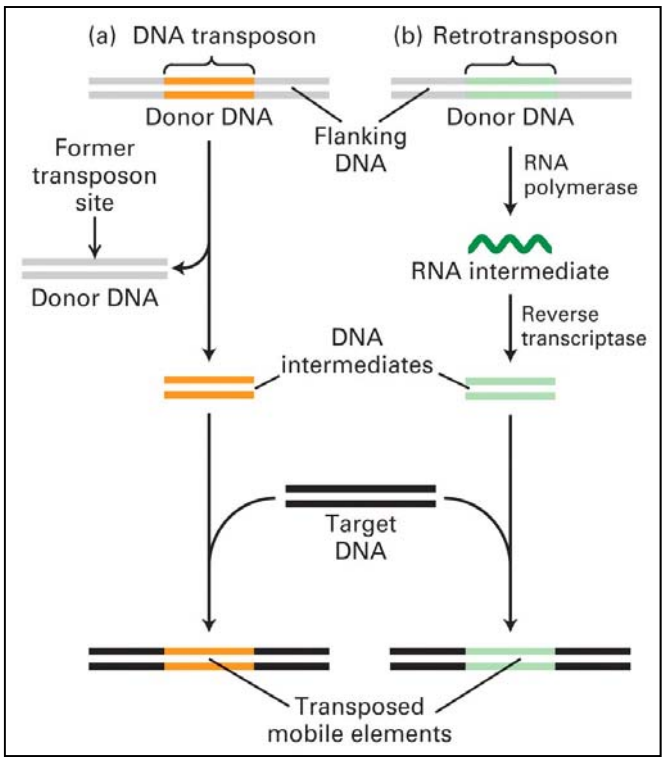

Figure 1.3: Transposition methods of transposable elements (Lodish *et al.*, Molecular Cell Biology, 7th ed).

1.2.2.1 RNA Transposons

RNA transposons (class I transposable elements or retrotransposons) are first transcribed into RNA, which are reverse transcribed before their integration at another location inside the genome via a copy-and-paste mechanism. RNA transposons are classified into two broad categories based on their structural relationship, the long terminal repeat (LTR) retrotransposons, and the non-LTR retrotransposons.

1.2.2.1.1 LTR (Long Terminal Repeats) Elements

Long Terminal Repeats (also known as endogenous retroviruses) are identical DNA sequences derived from ancient infections [12]. They repeated several hundreds of times, linked both ends of the genomes, and integrated by the reverse transcriptase of a retrovirus that manages the integration of the viral DNA into the host DNA and gene expression of

the virus. LTRs retrotransposons are responsible for many genetic variations. Copies of these fragments are much like that of a retrovirus. RNA copies are transcribed back into DNA using reverse transcription, and then inserted back into the genome. This reinsertion may have several effects: marginally modify the gene's function, completely alter the gene, or make no change whatsoever.

1.2.2.1.2 Non-LTR Elements

Non-LTR retrotransposons comprise of two broad categories: long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Non-LTR retroposons are prevalent in eukaryotic genomes.

- **LINEs (Long Interspersed Nuclear Elements)**

LINEs are autonomous retrotransposons that lack LTRs and are widespread in many eukaryotic genomes. They consist of long sequences of 6-8 kbp and comprise about 21% of the human genome. LINEs contain internal promoters for RNA polymerase III and encode a reverse transcriptase (ORF2) needed for transposition. LINEs are grouped into L1, L2, and L3 families, with the active elements belonging to the most abundant L1 family (7kbp), which alone comprises about 17% of the genome. Recent studies showed that human L1 elements have a stronger negative correlation with expression levels than the gene length and L1sequences within genes can significantly decrease transcriptional activity [27].

- **SINEs (Short Interspersed Nuclear Elements)**

SINEs are non-autonomous retrotransposons that found in the genome of most eukaryotic organisms, consisting of short sequences (<700 bp in length) [28]. They are

considered to be the largest family of repetitive DNA in the mammalian genomes [29] and comprise more than 10% of some higher eukaryotic genomes. For example, they represent about 13% of the entire human genome [30, 31]. SINEs uncommonly found in gene-rich regions and often located in transcribed regions of genes. In genes, SINEs are predominantly found in untranslated regions and introns [29]. They do not encode a reverse transcriptase (do not have reverse transcriptase gene), but instead rely on LINE-encoded enzymes for transposition. The most abundant SINE elements in the human genome are Alu elements [32] with a length of 280 bp [33]; these represent about 10% of the entire genome [34].

1.2.2.2 DNA Transposons

DNA transposons (class II transposable elements) move directly through DNA via a cut-and-paste mechanism. In eukaryotes, DNA transposons are less likely to be present than retrotransposons, representing only 3% of the human genome [35]. DNA transposons are designated by their terminal inverted repeats (TIRs) and have been grouped into superfamilies by the target site duplication (TSD), the presence or not of the DDE triad, the sequence similarities at the DNA and protein levels (e.g., Tc1/mariner, hAT).

Most DNA transposons are organized in families of autonomous and nonautonomous elements, characterized by their ability to respond to the same transposase. DNA transposons are thought to be transpositionally inactive in most mammalian genomes [36]; however, recent studies showed that DNA transposons could alter or stop the gene expression by insertion within exons, introns or regulatory regions [37].

1.3 Identifying Repetitive DNA

The search for repetitive DNA elements in the genome can be approached in various ways. This search depends on the level of knowledge of the repeats that are considered when identifying them in a genome sequence. It is possible to search for a particular element, to search for elements having structural features or to find entirely new and unknown elements solely based on their repetitive nature [13]. Many programs have been developed to identify repetitive DNA elements. In our analysis, the downloaded repetitive DNA elements datasets were defined by RepeatMasker program (www.repeatmasker.org), using consensus repeat sequences in RepBase [38].

1.4 Repetitive DNA Elements and Genome Evolution

Repetitive DNA elements contribute to genome evolution in diverse ways:

- Multiple copies of similar repetitive DNA elements may facilitate recombination, or crossing over, between the various chromosomes.
- Repetitive DNA elements insertion within a protein-coding sequence may inhibit protein production.
- Repetitive DNA elements placed in a regulatory sequence may change protein production positively or negatively.
- Repetitive DNA elements may move gene (s), singly or as groups to a different location.
- Repetitive DNA elements may also create new sites for alternative splicing in an RNA transcript.

1.5 Importance of Repetitive DNA

Repetitive DNA elements play a critical role in genome evolution and drive it in diverse ways [39]. Previous studies showed that repetitive DNA elements could affect the gene expression and genome stability [6]. Recent evidence has indicated their influence on gene expression and their responsibility for many genetic diseases, including cancer [40-42]. Some of these diseases are caused by tandem repeats and others by transposable repeats. For example, tandem repetitive DNA elements expansion can cause diseases based on their location in the genome. For example, *Fragile X Syndrome* occurs when "CGG" is repeated hundreds or even thousands of times creating a "fragile" site on the X chromosome that leads to mental retardation [20]. Also, *Huntington's disease* is caused by the trinucleotide repeat "CAG" expansion that elongates a protein of amino acid glutamine, leading to a neurological disorder that results in death [43]. TEs also cause chromatin instability and genomic rearrangements that result in a variety of genetic diseases, including, thalassemia, muscular dystrophy, and hemophilia in humans [44].

Repetitive DNA elements are an important feature of eukaryote genomes, representing the major fraction of their genomes. Thus it is important to identify the distribution and characterization of these repeats and determine their impact on the gene expression.

1.6 Objectives and Organization of the Thesis

This thesis has two major objectives. The first is to conduct a detailed comparative study of ten model organisms to investigate the distribution and characterization of repetitive DNA elements and to look for the similarities and differences between the organisms. The second objective is to investigate the association between repetitive DNA elements and gene expression levels in human and mouse genomes. Chapter two describes exploratory data analysis (EDA) to discover the essential characteristics of the abundance and the distribution of these repeats. This is followed by a study of the enrichment and the strand-preference of these repeats in different genomic contexts, defined by annotated genes. Chapter three conducts RNA-seq analysis to quantify the gene expression levels across ten different human and mouse tissues, with the resultant gene expression being used in chapter four to determine the influence of repetitive DNA elements on the gene expressions.

Chapter four builds various statistical models that quantify the association between the repetitive DNA elements and gene expression levels regarding repeat family (repFamily) and repeat name (repName) in two different promoter regions upstream the genes' 2,000 base pairs (2kbp) and 20,000 base pairs (2kbp). Then our models applied to different gene expression datasets to check model validity and results. Chapter five concludes the thesis by discussing the significance of the findings, the study's limitations and possible future work.

Chapter 2 - Exploratory Analysis of Repetitive DNA in Model Organisms

2.1 Introduction

Most eukaryotic genomes include substantial portions of repetitive DNA. In mammals, repetitive DNA is found in 40-50% of the total genomes. “Although the significance of repetitive DNA is not entirely understood, it may have both structural and functional roles, or perhaps even no essential role” [45]. Capitalizing on the availability of whole genome sequences and annotations, in this study, we compare ten different model organism genomes, ranging from nematodes, insects, and vertebrates to mammals. We investigate the similarities and the differences between their repetitive DNA regarding abundance, distribution, and their enrichments near genes. We will investigate whether these repeats have significant effects on gene regulation by comparing their frequencies in various genomic contexts. In order to do that, we compare the frequencies of repetitive DNA in the intergenic region between those different regions near genes — the 2kb promoter sequences upstream of transcription starting site (TSS), 5′ and 3′ UTRs, intronic regions, and 2kb sequence downstream of 3′ UTR.

2.2 Methods

We downloaded the locations of repetitive DNA from the UCSC Genome Browser [46] for ten different organisms, namely human (hg19), mouse (mm10), cow (bosTau7), rat (rn5), rhesus (reheMac2), chicken (galGal4), zebrafish (danRer7), Fugu (T. rubripes), fruit fly (dm6), and nematode worm (ce10). These repeats were identified by the RepeatMasker program (www.repeatmasker.org), using consensus repeat sequences in RepBase [38] (See Appendix A1 for more details about the annotation and RepeatMasker versions).

In order to calculate the repeats coverage in each genome, we used the following Bioconductor packages "*IRanges*" (ver. 2.0.1) and "*GenomicRanges*" (ver. 1.18.4) [47] for manipulating range objects. R packages *ggplots* (ver. 2.17.0) and *lattice* (ver. 0.20.31) were used to create the exploratory plots.

Initially, we used exploratory data analysis (EDA) to discover the essential characteristics of the abundance and the distribution of these repeats. Then, we studied the enrichment and strand-preference of the same repeats in different genomic contexts defined by annotated genes. Binomial tests for proportion were used to verify whether the number of repeats observed in the promoter regions was proportional to the coverage. The false discovery rate (FDR)[48] correction was used to correct for multiple testing. See R script in APPENDIX A2 for more details about enrichment/depletion and strand-preference calculations.

2.3 Results

The results are shown in Table 2.1 below, indicating the organism, genome size, repeat counts, and coverage percentages, as well as the coverage percentage variances between organisms. The highest proportion of repetitive DNA was found in zebrafish genome (52%) [49], followed by the human and cow (47%). The proportion results for rhesus, mouse, and rat were also relatively high, at 44.5%, 44%, and 38%, respectively. In contrast, Fugu, chicken, C. elegans, and fruit fly had the lowest percentages, at 7.3%, 11%, 13%, and 21%, respectively.

Table 2.1: Organism, genome size, repeats counts, and coverage percentages.

Organism	Genome size (bp)	Repeats count	Repeat Elements	Repeat Class	Repeat Family	Genome Coverage (bp)	% Coverage
Human (hg19)	3,137,161,264	5,298,130	1,395	16	45	1,469,734,726	47%
Mouse (mm10)	2,730,871,774	5,147,736	1,554	16	47	1,200,742,631	44%
Cow (bosTau7)	2,981,119,579	5,736,928	1,163	15	41	1,394,308,710	47%
Rat (rn5)	2,909,698,938	4,854,688	1,480	16	45	1,104,228,226	38%
Rhesus (reheMac2)	2,864,106,071	4,712,585	1,337	14	35	1,273,153,100	44.5%
Chicken (galGal4)	1,046,932,099	561,199	588	13	29	112,056,744	11%
Zebrafish (danRer7)	1,412,464,843	3,632,877	1,383	13	52	735,415,286	52%
Fugu (fr3)	391,484,715	210,322	508	13	42	28,759,869	7.3%
Fruit fly (dm6)	143,726,002	137,555	9,263	12	26	30,085,242	21%
C. elegans(ce10)	100,286,070	99,857	401	11	27	13,337,367	13%

2.3.1 Prevalence by Repeat Class

According to RepBase [38], repeats are classified into different types (such as mouse B1) with consensus sequences. These repeat types belong to particular repeat families (Alu) which, in turn, are grouped to repeat classes such as SINE, LINE, LTR, DNA transposons, simple repeats. Figure 2.1 summarized the coverage for these categories in each organism′ genome.

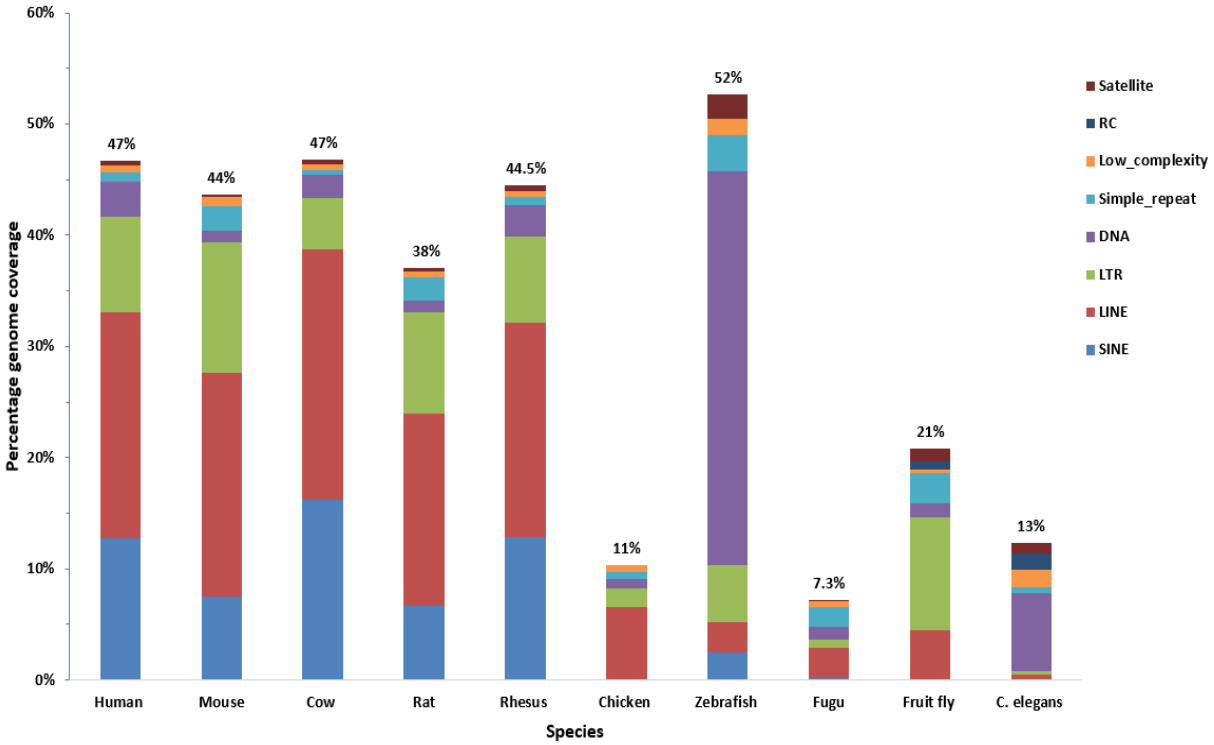

Figure 2.1: Percentage of repetitive DNA coverage in the ten model organism genomes. Repeat classes are color coded.

Repeat contents in mammalian genomes were found to be similar. Retrotransposons expansion is evident, since SINE, LINE, and LTRs constitute the majority of all repeats. In mouse, for example, the most frequently observed repeat class is SINE, followed by simple repeats, LINE, LTRs, and low complexity repeats. (Figure 2.2 A). LINEs are longer repeats covering about 20.1% of the mouse genome (Figure 2.2 B). LTRs cover 11.7% of

the genome, while SINEs and simple repeats comprise 7.5% and 2.2%, respectively. Figure 2.2 C shows that different repeat classes vary widely in diversity; for example, only 38 types of SINEs were noted, while 683 types of LTRs were found. There are also many different types of simple repeats and DNA transposons. The significant genomic coverage and the diversity of LTRs may make them candidates for regulatory roles.

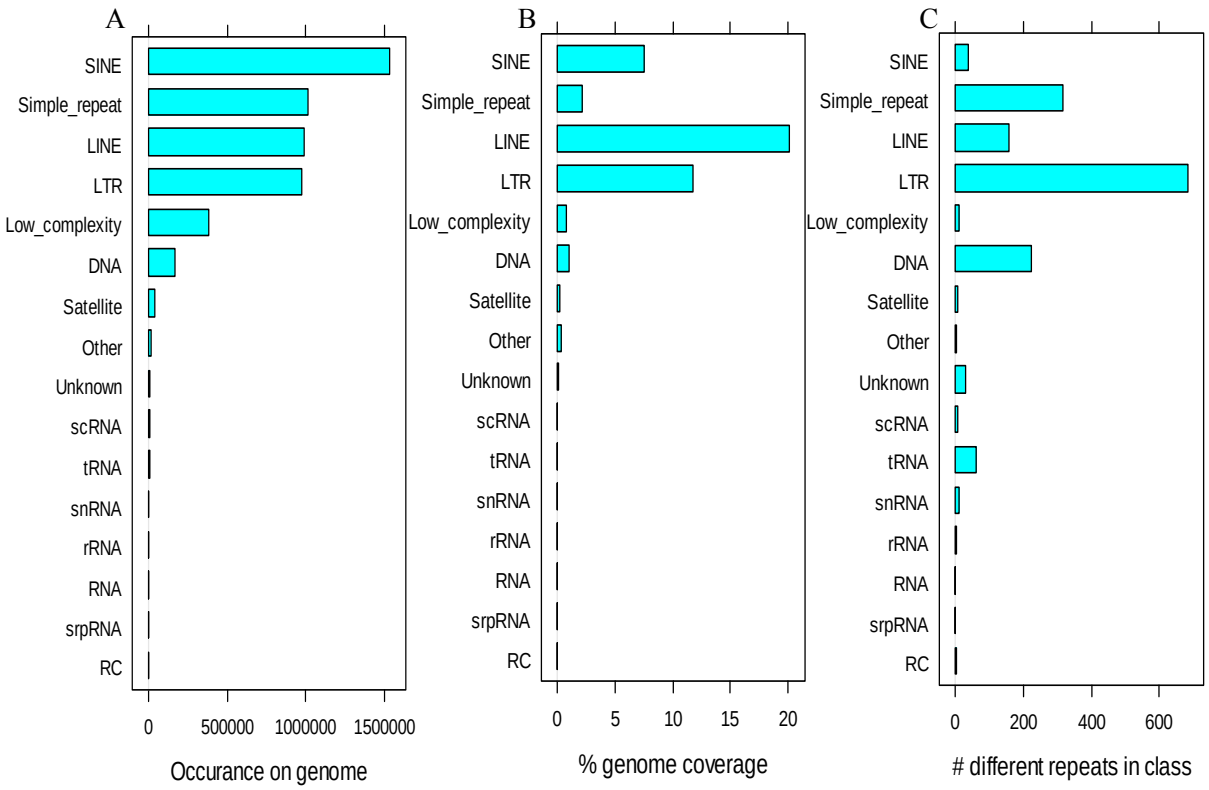

Figure 2.2: Repetitive DNA by class in the mouse genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

A similar trend is observed in the human, where the most frequently observed repeat class is SINE, followed by LINE, LTRs, and DNA repeats (Figure 2.3 A). LINEs are longer repeats included in 20.4% of the human genome (Figure 2.3 B). SINEs cover only 12.7% of the genome, while LTRs and DNA repeats comprise 8.5% and 3.2%, respectively. Figure 2.3C demonstrates that different repeat classes vary widely in

diversity; for example, there are 147 types of LINEs and 50 types of SINEs, while 505 types of LTRs were recorded. Many different types of simple repeats and DNA transposons can also be found. Similar to the mouse, the significant genomic coverage and the diversity of LTRs may make them candidates for the regulatory role.

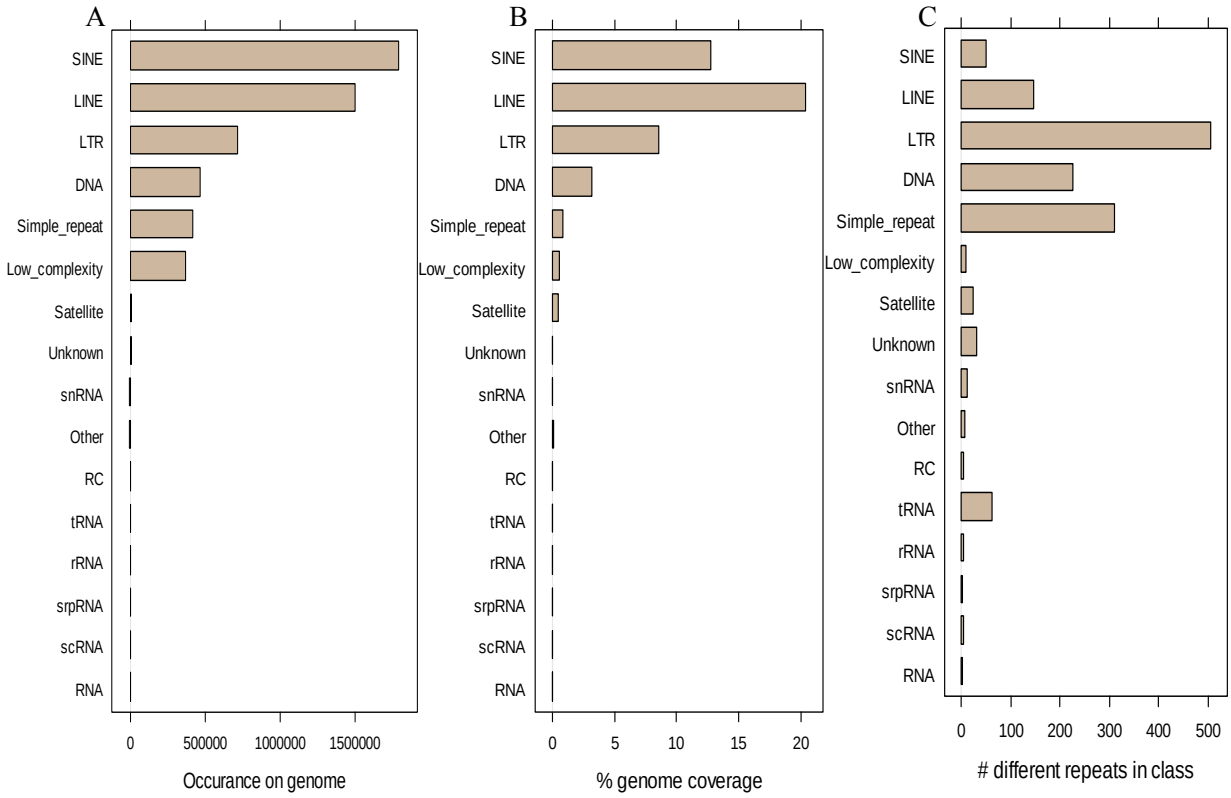

Figure 2.3: Repetitive DNA by class in the human genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

Supplementary Figures (S1-S3) in Appendix A3 show that in the three mammalian genomes, the most frequently observed repeat class is SINE, followed by LINE. In contrast, the most commonly observed repeat class in the chicken genome is LINE, followed by low complexity repeats. DNA transposons and simple repeats dominate the zebrafish and Fugu genomes, respectively. In the fruit fly genome, simple repeats are followed by LTRs, in the

C. elegans genome, low complexity is first, with DNA next (Supplementary Figures S4-S8 in Appendix A3). Major differences can be found in repeat content for the organism of different phyla, likely due to evolution.

To quantify the diversity of different repeat classes, we computed a Shannon index which is used in ecology to measure the diversity of an ecosystem [50]. Indeed, self-replicating repetitive elements, especially those from endogenous retrovirus, can be treated as an "organism" replicating on the genome.

Figure 2.4 illustrates the distribution of various repeats by prevalence and diversity based on the Shannon diversity index. Simple repeats have high diversity, as these are categorized by the exact repeat sequences, such as (CATATA)n. Among transposons, DNA transposons and LTRs are diverse, and their diversity increases as these elements expand into a different organism. This contrasts with SINEs, which have less diversity. SINEs are dominated by the rapid expansion of a few SINE elements, such as B1 and B2 elements in humans and Alu elements in primates. The significant genomic coverage and the diversity of LTRs in mammalian, fruit fly, and zebrafish genomes may suggest that they play a regulatory role. DNA transposons constitute a significant portion of zebrafish and C. elegans genomes, so their prevalence and enormous diversity may also serve as reservoirs of regulatory motifs.

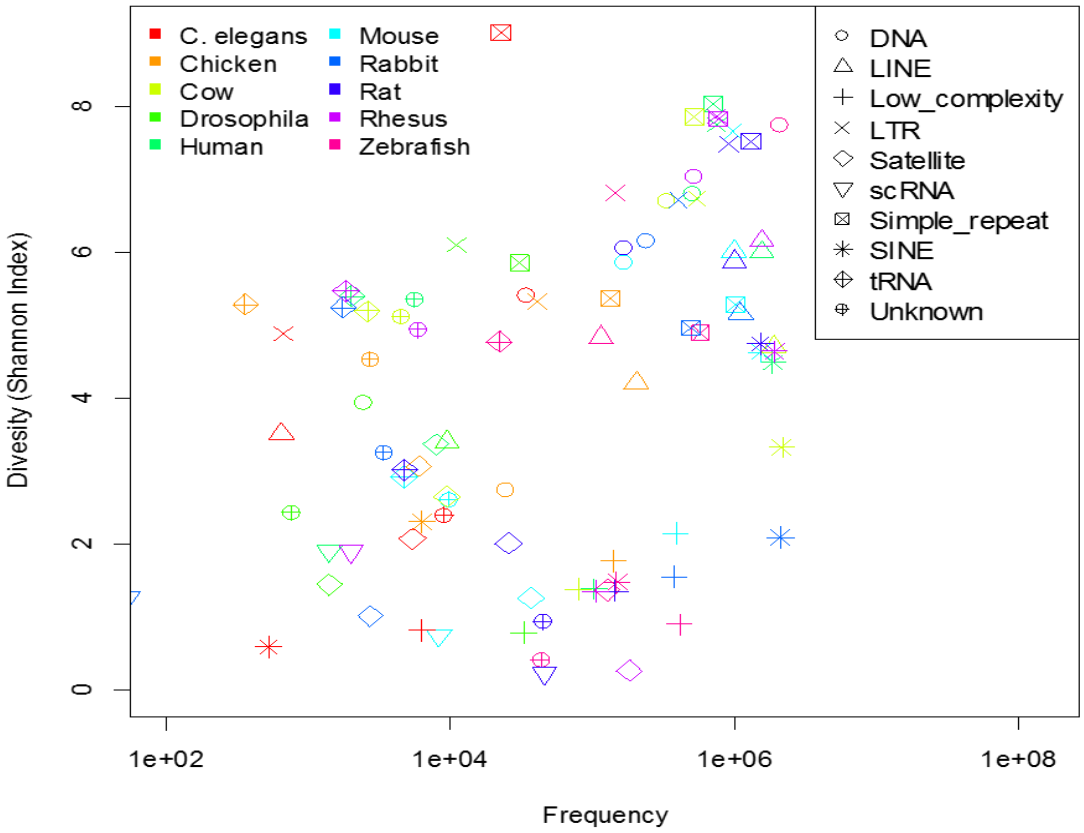

Figure 2.4: Distribution of repetitive DNA with regard to diversity and frequency.

2.3.2 Prevalence by Repeat Family

In the mouse genome (Figure 2.5 A), Alu, B4, and B2 are the most common SINEs families, while LINEs are dominated by L1 elements (Figure 2.5 B). ERVL-MaLR, ERVK, ERVL, and ERV1 are the most frequently observed repeat families in the LTR class (Figure 2.5 C). These LTR families consist of hundreds of different repeats. Similar trends are observed in the rat genome (Figure S9 in Appendix A3).

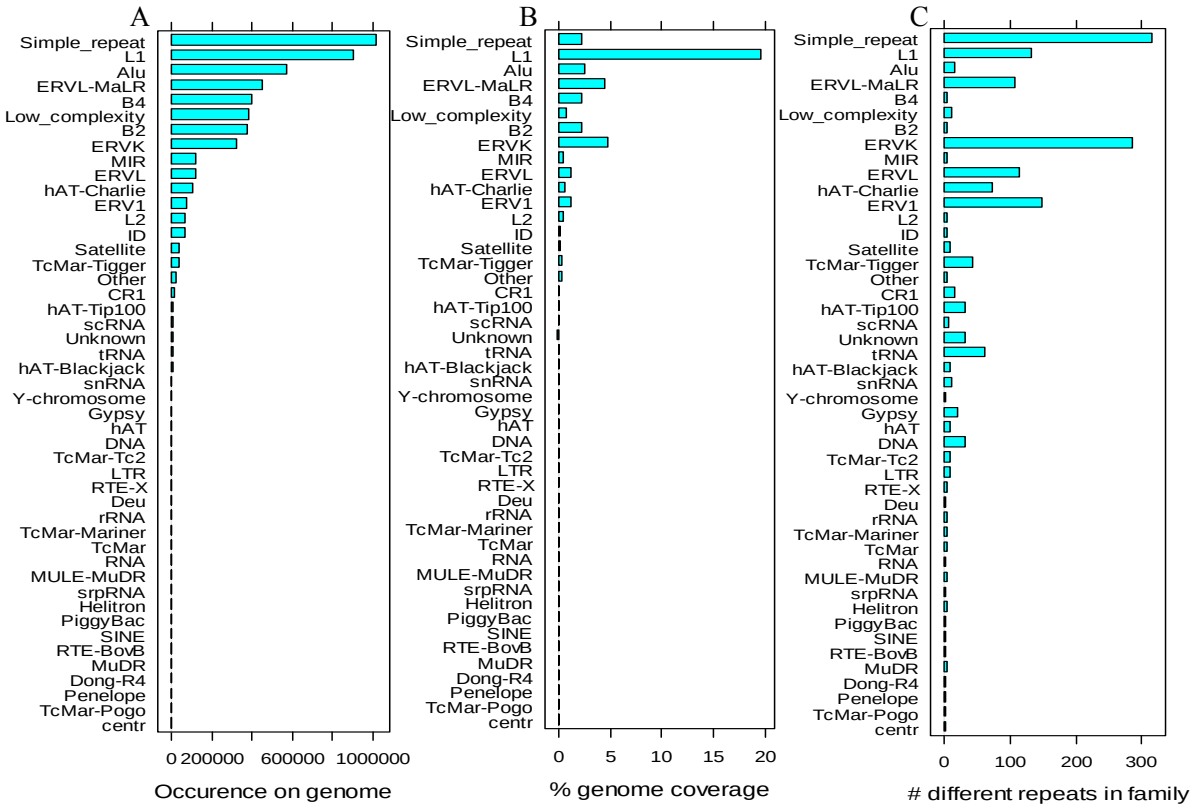

Figure 2.5: Repetitive DNA by family in the mouse genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

The human genome (Figure 2.6A) shows that the most prevalent SINEs belong to Alu and MIR families, while the most prevalent LINEs belong to L1 and L2 families. ERVL-MaLR, ERV1, and ERVL are the most frequently observed repeat families in the LTR class. Similar results are observed in the rhesus (Figure S10 in Appendix A3) and the cow genomes (Figure S11 in Appendix A3), but the cow genome also indicates that the most prevalent SINEs belong to BovA, RTE-BovB, MIR, and tRNA-Glu families.

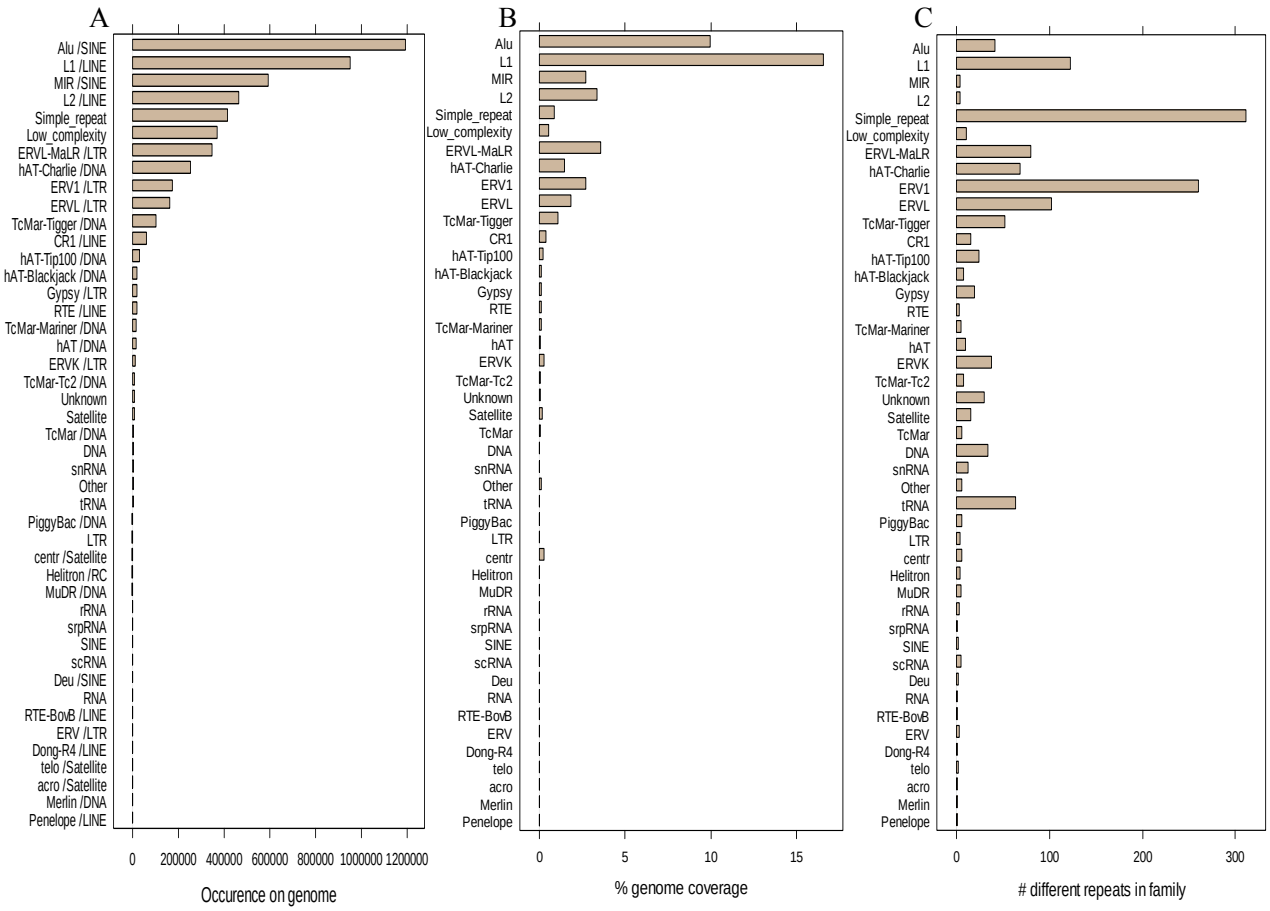

Figure 2.6: Repetitive DNA by family in the human genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

Unlike the mammalian genomes, significantly fewer retrotransposons are found in the zebrafish and Fugu genomes. In the zebrafish genome (Figure S12 in Appendix A3), the most prevalent DNA transposon belongs to the DNA, hAT, En-Spm, hAT-Charlie, and TcMar-Tc1families. The DNA transposons show remarkable diversity, with over 100 different types. Gypsy, LTR, and Nagro are the most frequently observed LTR. The Fugu genome (Figure S13 in Appendix A3) contains many simple repeats and LINEs, while the chicken genome (Figure S14 in Appendix A3) contains a significant quantity of LINEs of the CR1 family. ERVL and ERV1 are the most frequently observed repeat families in the LTR class; however, they lack the degree of diversity seen in mammals. The C. elegans genome (Figure S15 in Appendix A3) is also dominated by DNA transposons. The Helitron family repeats, of rolling-circle (RC) class, are a major type of TEs in this genome, with Pao and Gypsy being the most frequently observed repeat families in the LTR class. The fruit fly genome (Figure S16 in Appendix A3) contains many simple repeat and LTRs, with gypsy family elements being the primary type.

In summary, similar trends are observed for mammalian genomes. The most prevalent SINEs belong to Alu family, with the exception of cow BovA. LINEs are dominated by L1 and L2 elements in all the mammalian genomes. In contrast, the most prevalent LINEs in the chicken genome belong to the CR1 family, while the most prevalent DNA in zebrafish belongs to DNA, hAT, and En-Spm.

2.3.3 Frequencies Follow Log-normal Distribution

A vast difference can be seen in the prevalence of repeats among organisms. In the mouse genome, for example, some repeats are present more than 1 million times, while others only occur a few dozen times. The majority (50%) of the repeats have frequencies between 74 (first quartile) and 1363 (third quartile). This is likely the results of biased expansion during evolution. What can explain the vast difference in the frequencies of different types of repetitive elements? The distribution of repeats by their occurrence has a much longer right tail than normal distribution because of the small number of prevalent repeats. A histogram (Figure 2.7A) and a quartile-quartile (QQ) plot (Figure 2.7B) suggest the distribution is close to lognormal, which is confirmed by a Kolmogorov-Smirnov normality test (P=0.267). Figure 2.7C indicates it is not a power law distribution. The power law, or Zipf's law, is a widely observed distribution in various natural and social domains and could be expected if the more prevalent elements grow more rapidly [51]. Lognormal distribution, on the other hand, would imply that growth rate is independent of existing occurrence [52]. Since the distribution of repeats is much closer to lognormal, this suggests that the growth rates for different kinds of repetitive elements are comparable. A recent analysis shows that the distribution of the distances between repeats is similar to power-law [53], which could be expected as transposons often form clusters on the genome. After examining the distribution of all organisms, we found them to be approximately normally distributed, with the exception of C. elegans and fruit fly (Figure S17-S25 in Appendix A3).

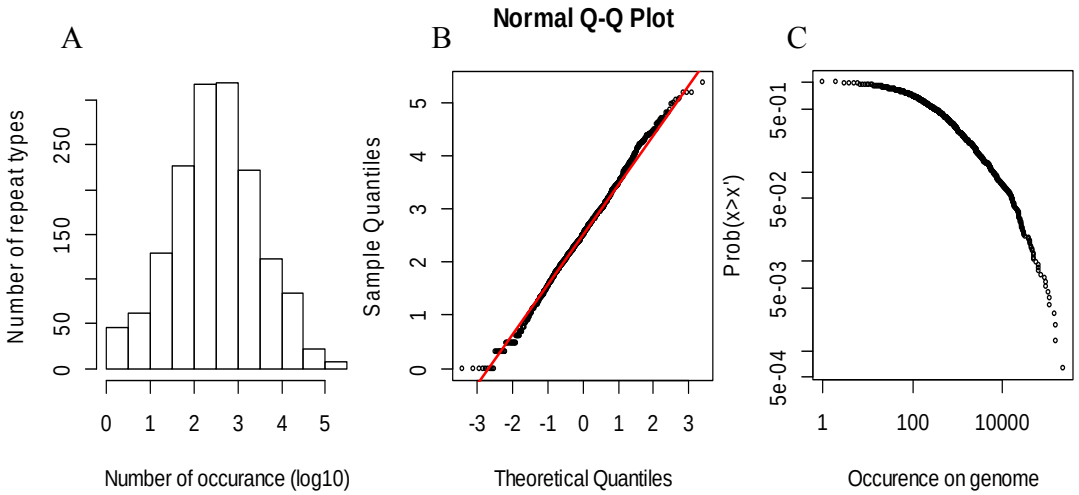

Figure 2.7: Distribution of repetitive DNA by the number of occurrences in the mouse genome. A: Distribution of the 1554 repeats according to how many times each repeat is observed in the mouse genome. After log-transformation, the distribution is bell-curved. B: The distribution is close to log-normal on a QQ plot. C: The distribution does not follow a power law.

2.3.4 Enrichment/Depletion and Strand-preference of Repetitive DNA Near Genes

In order to study the distribution of repetitive DNA in various genomic contexts, we compared the frequencies of repetitive DNA in the intergenic region with those different regions near genes — the 2kb promoter sequences upstream of transcription starting site (TSS), 5′ and 3′ UTRs, intronic regions, and 2kb sequence downstream of 3′ UTR. For example, in the mouse genome, we found that most (90%) of the repetitive DNA occurs in intergenic or intronic loci, as expected (Figure 2.8). Promoter regions also contain many repeats.

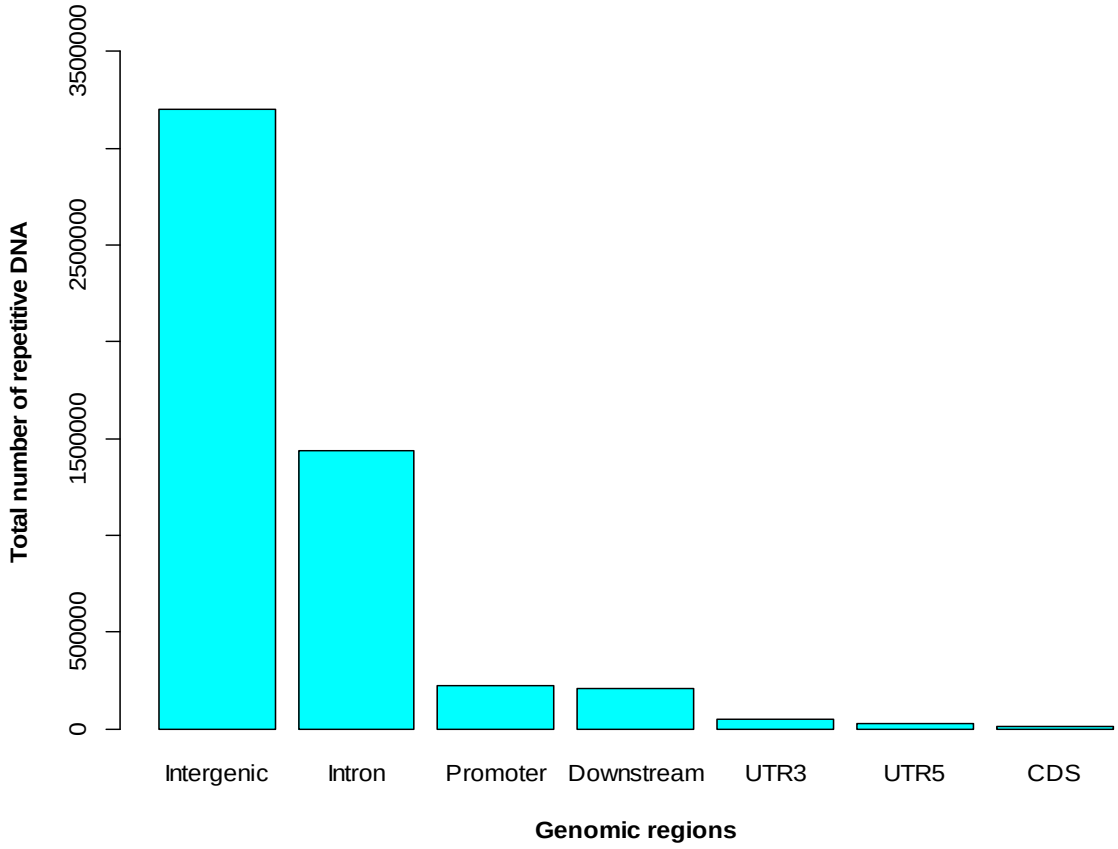

Figure 2.8: Total number of mouse repetitive DNA in different genomic contexts.

Figure 2.9 summarizes the number of unique repeats, but not the occurrence, by these regions. Promoters, introns, and downstream regions are enriched in most SINE, while they are depleted in most LINE, indicating that LINEs tend to be located away from genes. However, some LINEs enriched in intron and promoter regions such as L2, L2a, L2b, and L2c. 5'UTR, 3'UTR, and coding regions are depleted in most repetitive elements. Most LTRs are depleted from introns, but some types of LTR repeats that are enriched in the promoter region. For example, a 5317 bp mouse-specific RLTR14-int repeat, which belongs to the ERV1 family, can be found in promoter regions 287 times, which is 15.8% of all total occurrences in the genome. As the promoter regions only cover about 4% of the genome, this is a significant 3.9-fold enrichment, according to a binomial test of proportion ($P<1.2x10^{-84}$). Therefore, a significant number of LTRs are specifically enriched in promoter regions.

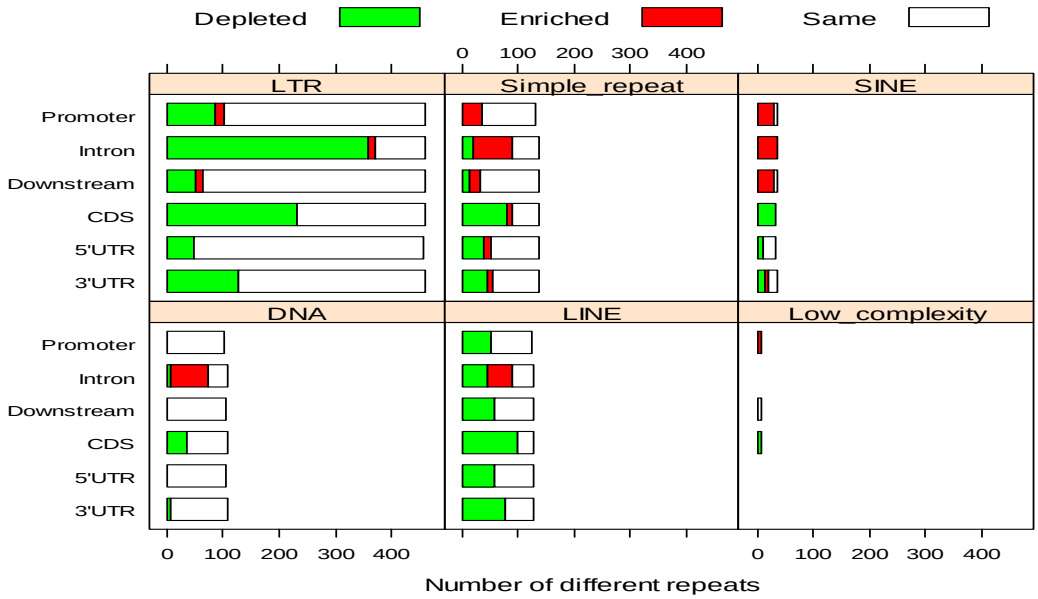

Figure 2.9: Distribution of repetitive DNA in different genomic regions of the mouse genome.

Figure 2.10 demonstrates that a large number of intronic repeats are highly strand-specific; that is, most intronic repeats are more likely to be on the opposite strand. Some repeats are depleted from promoter regions, but when they occur in these regions, they have a higher strand-bias. ORR1F, ORR1E, and MLT1B, all of which are in the ERVL-MaLR family, belong to this category. The depleted repeats might also be of interest.

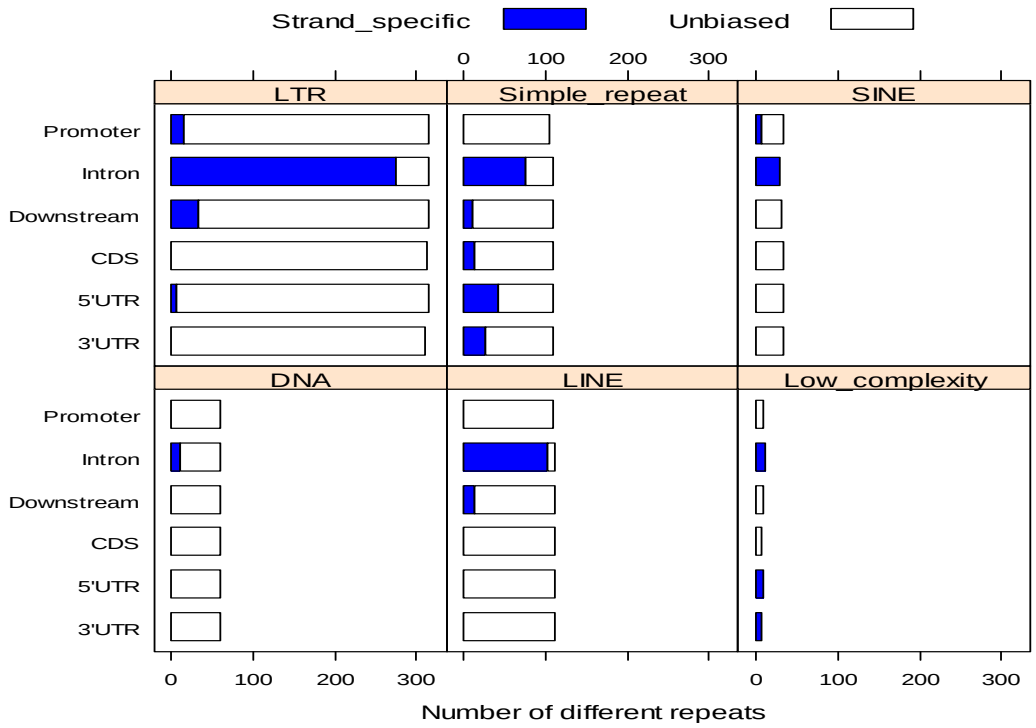

Figure 2.10: Strand-specificity of the repetitive DNA in different genomic regions of the mouse genome.

Similar trends are observed in the human, rat, rhesus, and cow genomes, where from 85%-98% of repetitive DNA occurs in the intergenic region (Figure S26-S29 in Appendix A3). Most LTRs are depleted from introns, but some types of LTRs are enriched in promoter regions, which cover from 2.6% - 43% of all total occurrences on those genomes. Since the promoter region covers approximately 0.9% of the cow genome and

6.4% of the human genome, it has a significantly 2.8 - 6.8-fold enrichment, according to the binomial test of proportion, with p-values ($P< 3.0x10^{-14}$) and ($P< 4.3x10^{-35}$).

In zebrafish and Fugu genomes, 88% and 78% of repetitive DNA occur in the intergenic region, respectively (Figure S30-S31 in Appendix A3). Promoter, intron, and downstream regions are enriched in most DNA in zebrafish, while they are enriched in very low complexity in Fugu. The promoter region has 73 different DNA types enriched. Both zebrafish and Fugu have promoters, introns, downstream, and CDS regions enriched in most simple repeat, while 5′UTR, 3′UTR, and coding regions are depleted in most repetitive elements. Intron is enriched in most LINEs in the zebrafish, while promoter is enriched in most LINEs in Fugu (Figure S41-S42 in Appendix A3)

The chicken genome shows that 92% of repetitive DNA occurs in the intergenic region (Supplementary Figure S32 in Appendix A3). The promoter region is enriched in LINEs with some repeats such as CR1-B, CR1-C, CR1-C4, CR1-D2, CR1-F0, CR1-F2, CR1-X, CR1-X1, and CR1-Y4. This is different from what is observed in mammalian genomes, where LINE elements are often found away from genes. Most LTRs are depleted from introns and promoter, intron, and downstream regions (Supplementary Figure S43 in Appendix A3).

The fruit fly genome demonstrates that 64% of repetitive DNA occurs in the intergenic region (Supplementary Figure S33 in Appendix A3). Promoters, introns, downstream, 5′UTR, 3′UTR, and coding regions are depleted in most repetitive DNA, with some exceptions in the simple repeat. Intron, 5′UTR, 3′UTR, and coding regions are enriched in a simple repeat.

Unlike other organisms, C. elegans genome results indicate that 43% of repetitive DNA occurs in the promoter region (Supplementary Figure S34 in Appendix A3). This occurs since approximately 20,000 genes exist in such a small genome size. Promoter and downstream are depleted in all repeat classes. Introns are enriched in most DNA, with 39 different types of repeats. 5′UTR, 3′UTR, and coding regions are depleted in most repetitive elements, except in the simple repeat (Supplementary Figure S42 in Appendix A3).

Overall, Figure 2.11 shows that most repetitive DNA in our organism genomes can be found in the intergenic or intronic loci, as expected except, with the exception of C. elegans, in which most of the repeats (43.5%) occur in the promoter region

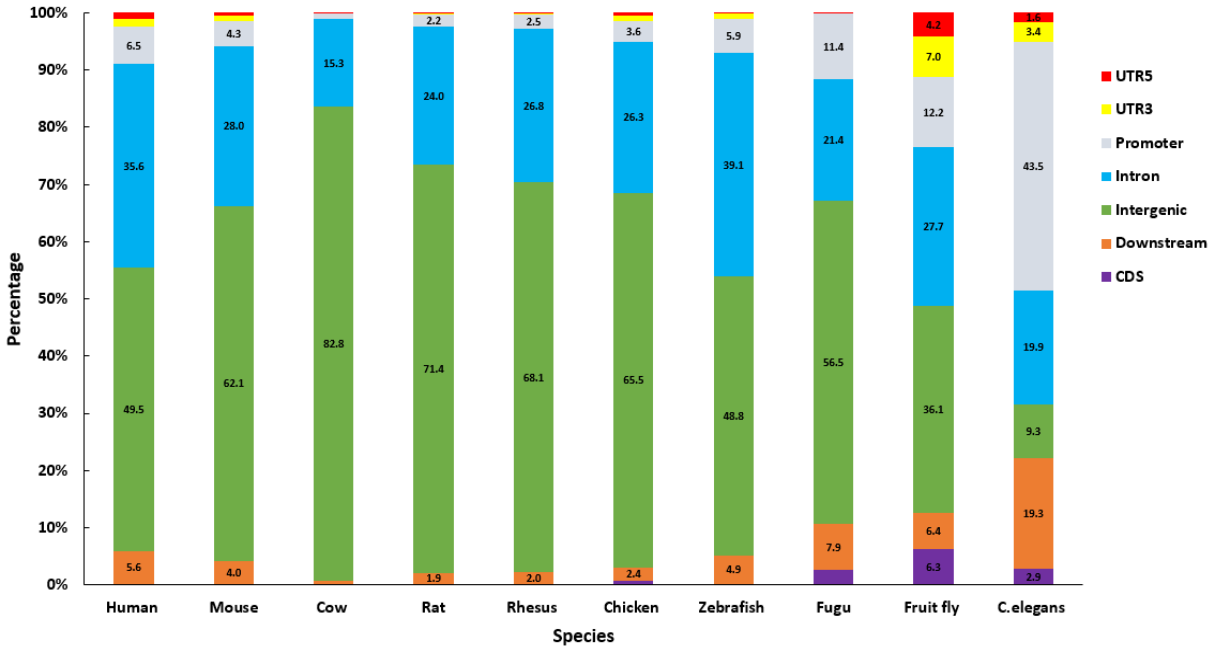

Figure 2.11: Percentage of repetitive DNA coverage in ten model organism genomes. Genomic regions are color-coded.

Figure 2.12 provides a comparison of the repetitive DNA enrichment between mammal genome (mouse), vertebrate (zebrafish and chicken), insect (fruit fly), and nematode worm (C. elegans). We can see that, in the mouse, SINEs are enriched in introns, promoters, and downstream regions. In contrast, LINEs are depleted from these regions with the expectation of some LINEs in the intronic regions. Simple repeats are enriched in both the mammal and vertebrate genome. In zebrafish, DNA transposons are enriched in introns, promoters, and downstream regions. The following DNA transposons are highly overrepresented in the promoter region: DNA-5-2_DR, Kolobok-1_DR, Kolobok-N4_DR, DNA-8-36_DR. Of these, DNA-5-2_DR is particularly interesting because of its prevalence with 10,263 copies and strand prevalence (FDR<1×10^{-11}). We did not identify any enriched repeats in the promoter regions of C. elegans and fruit fly.

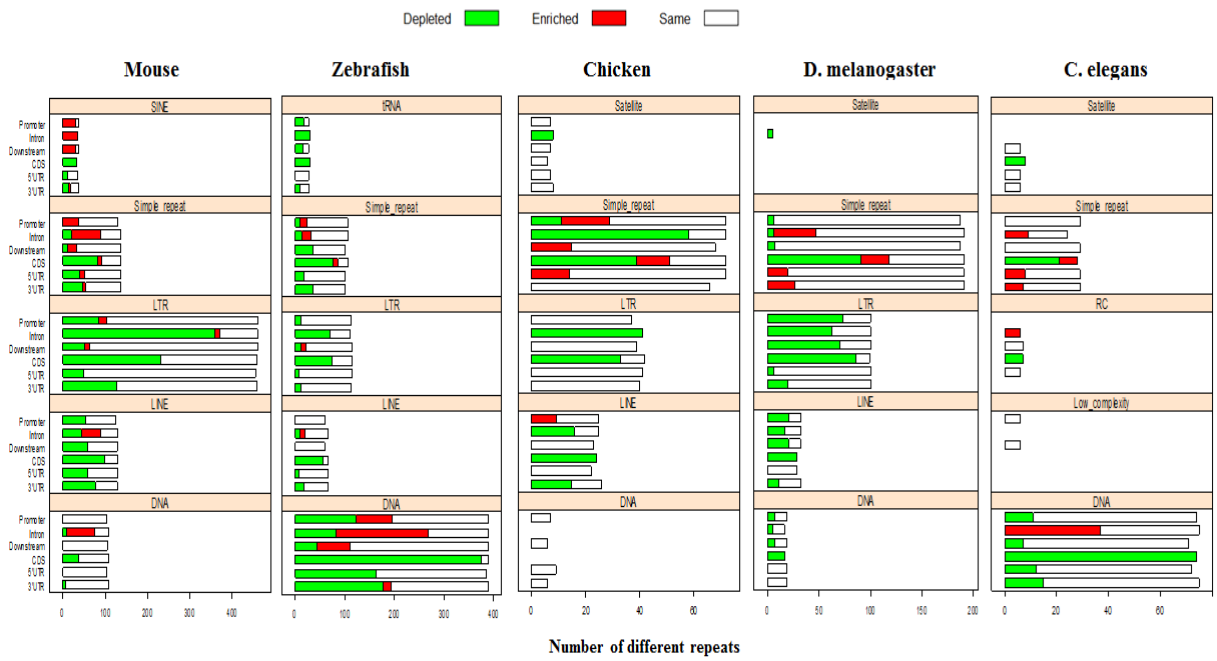

Figure 2.12: Enrichment of repetitive DNA comparison between mammal genome (mouse), vertebrate (zebrafish and chicken), insect (fruit fly), and nematode (C. elegans)

Supplementary Figures (S44-S52 in Appendix A3) show that intronic repeats are more likely found on the opposite strand of the transcripts. One possible explanation is that intronic sequences, once spliced off transcripts in the nucleus, are subjected to reverse transcription and recombination back to the genome, and subsequently give rise to intronic retrotransposons. Indeed, in animals, repeats are more likely to be found in introns.

Repetitive elements derived from small nuclear RNAs (snRNA) are enriched in the promoters and UTR regions in many genomes, including human, mouse, rhesus, chicken, and zebrafish. The enrichment is especially profound for U6, U13, U1, and U4 in 5′ UTR. For example, the 1495 copies of U6 in the rhesus genome are overrepresented in the 5′ UTR by 280-fold, compared to the genome as a whole. The U6 elements contain Pol III promoters that could drive expression of non-coding RNAs [54]. The potential role of snRNAs retrotransposition in the evolution of non-coding genes needs to be further studied.

Another feature observed across organism is the enrichment of repeats with high GC content near genes, especially in promoters. In addition to G-, C-, or GC-rich low-complexity repeats, many simple repeats [(CCCCG)n, (CCG)n, (CGG)n, (CGGGG)n] are overrepresented in promoters. Using FDR<1×10^{-5} as a cutoff, we selected 245 simple repeats enriched in promoter regions, of which 177 are from mammalian genomes, 48 from the chicken, and 20 from the zebrafish. As shown in Figure 2.13, simple repeats that are enriched in promoters are of high GC content. On the contrary, simple repeats depleted from promoters are often of low GC. CpG sites influence DNA methylation, and methylated cytosines are subject to spontaneous deamination to thymine in genomes. Expansion of these GC-rich repeats near genes might help keep the balance and participate in regulating gene expression through epigenetic mechanisms.

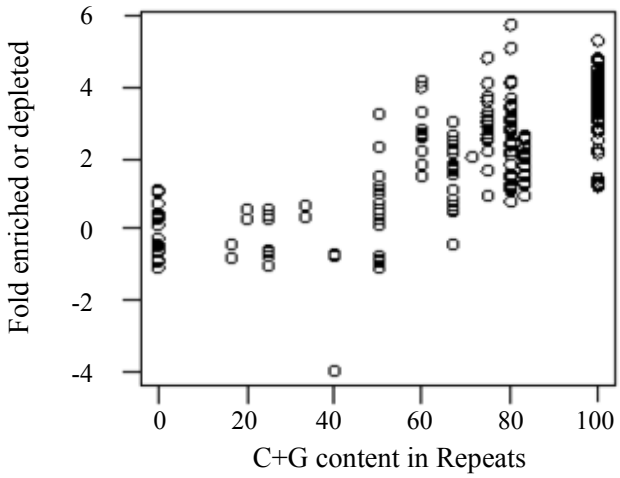

Figure 2.13: Distribution of significantly enriched or depleted simple repeats in promoters across organisms.

2.4 Conclusion and Discussion

We studied the distribution of repetitive elements in ten model organisms and found significant evidence pointing to non-randomness concerning the location, frequency, and strand-preferences of different repeats. Often found near genes are the repeats, such as the Alu family repeats in human and mouse, the GC-rich simple and low complexity repeat in most other organisms. Other repeats, such as LINEs in mammals are more frequently found away from genes. Also, some of the repeats show strong strand-bias compared to nearby genes, which indicates that these retrotransposons might be linked to the evolution of these genes. We also identified many LTRs that are specifically enriched in promoter regions, some with a strong bias towards the same strand as the nearby gene. This raises the possibility that the LTRs, may play a regulatory role. Since they have a higher degree of diversity compared to LINEs and SINEs. While the composition of different repeat classes and coverage in mammalian genomes are similar, vast differences were found among the various vertebrate genomes. Each organism exhibited examples of extremely prevalent repeats successfully fixed in the genome. The most frequently observed transposable element in mammals is SINE, compared to DNA transposons in zebrafish, LINEs in chicken, and low complexity repeats in the C. elegans genomes. These repeats may have a substantial influence on the genetic landscape of the genomes.

We have shown that repetitive DNA elements vary in their coverage among organisms, from 7.3% in the Fugu genome to 52% in zebrafish. Except for C. elegans and the fruit fly, the frequency of the TEs follows a log-normal distribution, characterized by a few highly prevalent repeats in each organism. Surprisingly, we found that most intronic repeats, with the exception of DNA transposons, have a strong tendency to be on the

opposite DNA strand as the host gene. One possible explanation is that intronic RNAs resulting from splicing may contribute to retrotransposition to the original intronic loci.

Overall, our results indicate that comparative studies of TEs in multiple organisms can lead to insights into their evolution and expansion, as well as into their potential functions. The non-random distribution of repeats across multiple organisms adds to the existing evidence that some repetitive DNA elements are drivers of genome evolution [55-58], rather than being “junk” DNA.

Chapter 3 - Quantifying Gene Expression for Human and Mouse Tissues using RNA-sequencing (RNA-seq) Analysis

3.1 Introduction

The conversion of genetic information stored in DNA to RNA and RNA to protein is the central dogma of molecular biology [59]. The information stored in DNA is called a gene, with the conversion of DNA to mRNA labeled gene expression. A gene expression pattern provides valuable information regarding the specific function of cells and organs. Instead of looking at an individual gene, analysis of gene expression at the global level is defined as transcriptomics.

The transcriptome is a complete set of transcripts present in a cell. The quantity of transcriptome determines the specific developmental stage or physiological conditions. Understanding the transcriptome of an organism plays a key role in interpreting the functional elements of the genome and the study of the molecular content of cells and tissues. Several popular methods are used to study transcriptomics, such as differential display, subtractive hybridization, Serial Analysis of Gene Expression (SAGE), DNA microarray, and RNA-Seq. Among these methods, Microarray and RNA-Seq are the most commonly used. RNA-Seq is an approach to transcriptome profiling that uses deep-sequencing technologies called next-generation sequencing (NGS) [60]. This method allows for more precise measurements of transcriptome than other methods [61]. RNA-Seq is also referred to as “Whole Transcriptome Shotgun Sequencing,” has the capacity to reveal the presence and quantity of RNA present at a given moment of time [62]. RNA-Seq analysis also has capabilities to look at different populations of RNA such as

microRNA (miRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA) [63]. Compared to the microarray, RNA-Seq provides transcriptome information at the single-base resolution, with low background signal, high dynamic range of detection, and less RNA required [61].

During the RNA-Seq experiment, RNA is extracted and converted to cDNA libraries with adapters ligated. The libraries are then sequenced using any of the sequencing technology, such as sequencing by synthesis (Illumina), sequencing by ligation (SOLiD), pyrosequencing (454). The sequenced information is retrieved in the form of nucleotide reads, which is mapped to the genome if there is a reference genome available. If no reference genome is available, denovo gene assembly is done to study the transcriptomics. This technique has many applications, including gene expression profiling, alternative expression analysis, transcript discovery and annotation, allele-specific expression, mutation detection, fusion detection, and RNA editing [64].

The main goal of this chapter is to quantify the gene expression levels in various human and mouse normal tissues through the specified pipeline analysis, with the resultant gene expression being used in the next chapter to determine the association between gene expression levels and repetitive DNA elements.

3.2 Methods and Results

Total RNA samples were purchased from Ambion (Austin, TX, USA). The RNA samples consist of ten human tissues (brain, colon, heart, kidney, liver, lung, prostate, spleen, thymus, and uterus) and ten of mouse tissues (brain, colon, embryo, heart, kidney, liver, lung, spleen, thymus, and uterus). The next generation sequencing was done at the

University of Chicago Functional Genomic Facility. The ribosomal RNA (rRNA) was removed with Ribozero Human/Mouse from Epicenter. The strand-specific RNAseq libraries were prepared with the NEXTflex™ Directional RNA-Seq Kit, dUTP method (Bioo Scientific, Austin, TX). Each library was quantitated by qPCR and sequenced on one lane 101 cycles on a HiSeq2000 using a TruSeq SBS sequencing kit version 3 and analyzed with Casava1.8.2.

3.2.1 Data Description

The raw RNA-seq datasets consisted of twenty files, ten human and ten mouse, with a total size of approximately one terabyte (1TB). Each library contained millions of reads, with 100 base pair long. Table 3.1 and Table 3.2 show the description of the raw RNA-seq reads in both mouse and human tissues, respectively as a fastq format.

Table 3.1: Description of mouse raw RNA-seq datasets.

File name	File size (GB)	Number of sequences
Mouse_brain.fastq	39.7	168,256,624
Mouse_colon.fastq	46.1	190,145,268
Mouse_7day_embryo	39.5	163,057,876
Mous _heart.fastq	45.5	193,038,010
Mouse_Kidney.fastq	44.5	188,632,130
Mouse_liver.fastq	45.1	191,130,627
Mouse_lung.fastq	44.0	186,647,729
Mouse_spleen.fastq	43.5	184,433,667
Mouse_thymus.fastq	45.1	191,405,697
Mouse_uterus.fastq	47.2	194,803,783

Table 3.2: Description of human raw RNA-seq datasets.

File name	File size (GB)	Number of sequences
Human_brain.fastq	47.6	196,243,272
Human_colon.fastq	46.0	189,915,611
Human_heart.fastq	44.1	187,083,380
Human_Kidney.fastq	43.9	186,063,722
Human_liver.fastq	44.6	189,078,160
Human_lung.fastq	44.8	189,968,197
Human_prostate.fastq	40.3	171,075,900
Human_spleen.fastq	44.4	170,936,098
Human_thymus.fastq	42.3	179,505,781
Human_uterus.fastq	48.1	203,136,931

3.2.2 RNA-sequencing (RNA-seq) Pipeline Analysis

RNA-seq has so many uses that no one type of pipeline analysis can be used in all cases [65]. In our analysis, the raw reads were analyzed using TUXEDO pipeline, which included TopHat, and Cufflinks programs [66], with the mouse (mm10) and human (hg19) genome annotations from Ensembl [67]. In order to conduct this analysis, we used a Linux cluster for research computing "High-Performance Computing (HPC)" at South Dakota State University. Figure 3.1 shows the RNA-seq analysis workflow including, TUXEDO pipeline analysis.

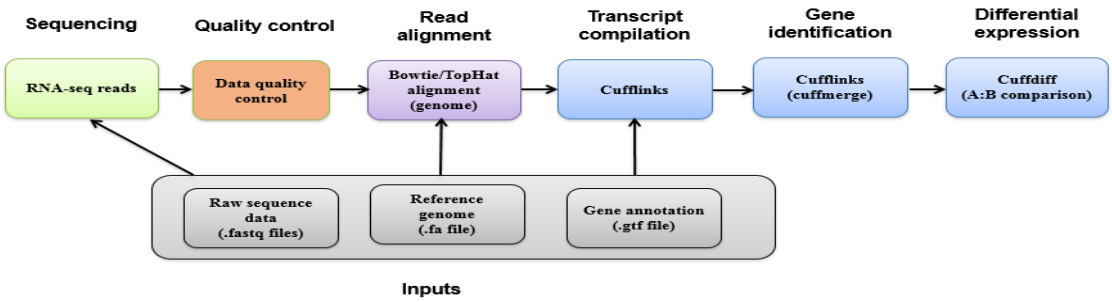

Figure 3.1: RNA-seq analysis workflow (www.bioinformatics.ca).

Quality control and filtering of raw sequence reads are the most important steps in the pre-processing of sequencing reads. Thus, before using these raw sequences to run the RNA-seq TUXEDO pipeline, their quality is checked, and they are cleaned to avoid low-quality sequences, adaptors, and contaminants.

3.2.2.1 Assessing the Sequence Reads Quality

Lower quality sequences might negatively influence the analysis by providing unreliable results, as well as erroneous sequence information. Sequence quality is affected by several factors, including the quality of library, sequencing error, Polymerase Chain reaction (PCR) artifacts or contaminations [65]. In Illumina sequencing technology, errors are more likely to occur at the 3′-ends of a read [68]. It is crucial to check the quality of the sequences before proceeding with the analysis to ensure both reliability and reproducibility of results. Several bioinformatics tools are available to check the quality of the sequence. In our analysis, FastQC (fastqc_v0.10.1) [69] software was used to check all human and mouse reads. FastQC provides a modular set of metrics, including sequence basic information, sequence quality, GC content (%GC), the presence of adapters, overrepresented *k*-mers and duplicated reads. The FastQC results can be used to provide a quick impression of whether the data has any issues of which we should be aware before doing any further downstream analysis. In the human, for example, Figure 3.2 shows the quality of the human brain RNA-seq which has a low sequence quality at the 3′-end.

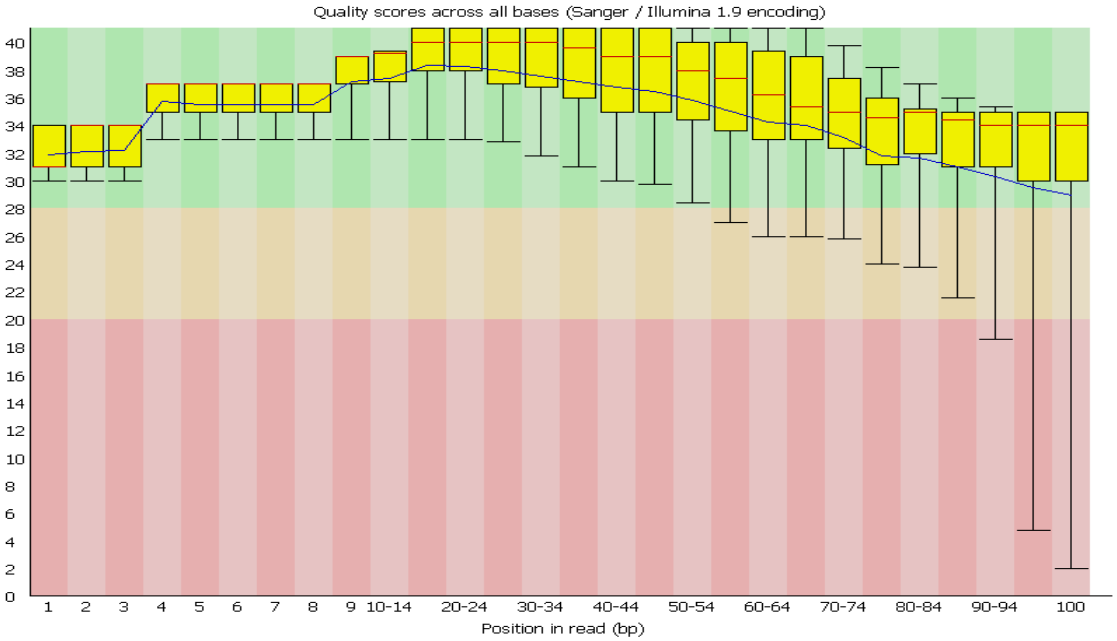

Figure 3.2: Human brain RNA sequence quality.

3.2.2.2 Filtering and Cleaning the Raw Sequence Reads

The raw RNA-seq sequences may have some regions that could be problematic. For example, some of the sequences may have some adaptor sequences left at the 3' end, and some of the sequences may have a low-quality score. To avoid these problems, we need to filter and clean the data before proceeding to the next step. FASTX-toolkit [70] was used to remove the adaptor sequences and to discard low-quality reads through *Fastx_clipper* and *fast_quality_trimmer* procedures. Precautions were taken to make the high-quality sequence but not to lose the large set of sequence, while also choosing the set of parameters in the *fast_quality_trimmer* procedure, as shown in the human case below.

```
fastx_clipper  -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAATCTCGTATGCCGTCTTCTGCTTG -Q 33 -i Human_brain_ATGTCA_L001_R1_001.fastq  -o TrimmedData/Human_brain_temp.fastq

fastq_quality_trimmer -t 16 -l 20 -Q 33  -i TrimmedData/Human_brain_temp.fastq  -o TrimmedData/Human_brain_trimmed.fastq
```

Where: [-a ADAPTER] = ADAPTER string, -Q is the quality score, [-i INFILE] = FASTA/Q input file, [-o OUTFILE] = FASTA/Q output file, -t is the quality threshold, lower quality bases are trimmed (removing the nucleotides with lower quality from the end of the

sequence), -l is the minimum length post-trimming sequence to keep (-l 20 removing the reads with a length lower than 20).

FastQC was used again to check all RNA-seq quality for all cleaned data. Figure 3.3 shows the quality of human brain data which has been cleaned, noting the sequences with quality scores of 26 or more.

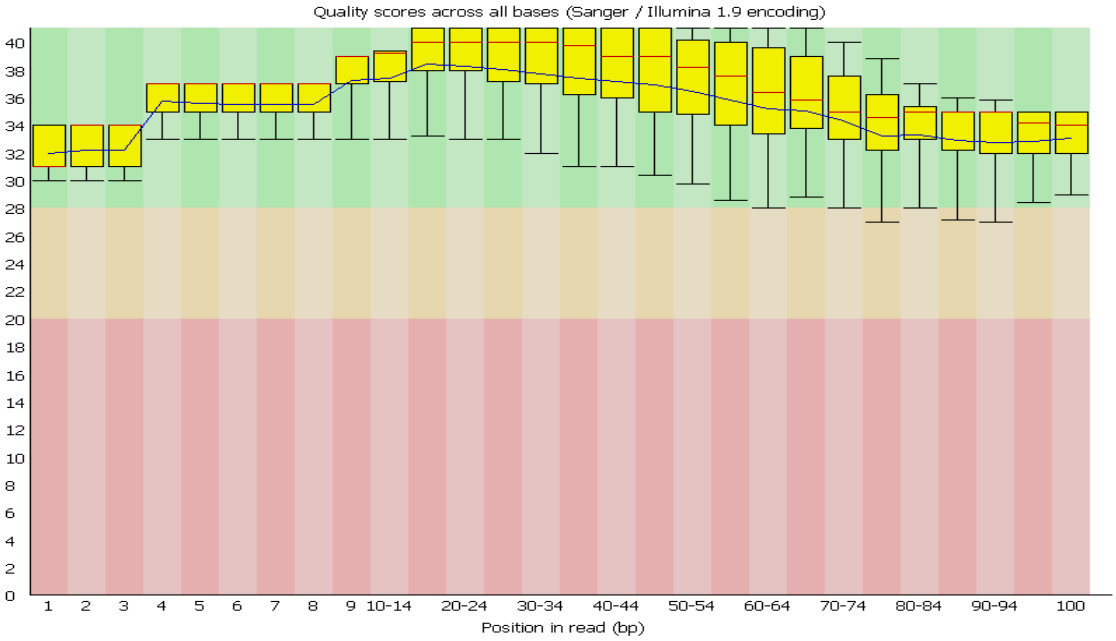

Figure 3.3: Human brain RNA sequence quality after data cleaning.

3.2.2.3 Mapping / Aligning Reads to Reference Genome

This stage considered as the initial step in most RNA-seq analysis pipelines. Thus, the accuracy of downstream analyses will be heavily dependent on it. Many algorithms and alignment tools have been developed to align reads to genomes. The main challenge when mapping RNA-seq reads is the splice junctions (exon-intron junctions) because these reads come from RNA and often cross splice junction boundaries. Thus, typical NGS aligners, such as Bowtie and BWA, are not ideal without modifying the genome sequence.

The first step to map reads is building an index for the reference genome (The human and mouse genomic sequence data used in this study are "GRCh37/hg19" and "GRCm38/mm10"). Human and mouse genomes were obtained from the UCSC Genome Browser as '**hg19.2bit**' and '**mm10.2bit'** formats. The '**hg19.fa'** and **'mm10.fa'** were extracted from the '.2bit.' files by using the utility program 'twoBitToFa.' Then Bowtie2 [71] used to create the genome index files. The last mapping step is to align the RNA-seq reads onto the indexed genomes. This is often the most time-consuming step in an RNA-seq analysis, but can be greatly expedited by using additional processing cores. Computational time increases with the genome size and the number of reads. See Appendix A2 for a complete Linux script. Tophat (Tophat-2.0.3.1.Linux_x86_64) [66] was used to map reads to both human and mouse genomes with specified parameters. For example, with the human brain data, we used the following script to run Tophat procedure.

tophat --library-type fr-firststrand -p 14 -G /disk4/aburweism/RNAseq/GeneModel/Hs_ensembl_37.gtf -o /disk4/aburweism/RNAseq1/Alignment/Human_brain /disk4/aburweism/RNAseq/Human_genome/Ensembl/GRCh37/Bowtie2Index/genome /disk4/aburweism/TrimmedData/Human_brain_trimmed.fastq &> /disk4/aburweism/RNAseq1/Alignment/Human_brain/tophat_screen_results &

Where:

library-type fr-firststrand: library type
-p 14: execute alignment with 14 cores
-G: uses known genes (Supply TopHat with a set of gene model annotations and/or known transcripts, as a GTF 2.2 or GFF3 formatted file)
Input reads: /disk4/aburweism/TrimmedData/Human_ brain_trimmed.fastq
Whole genome sequence: /disk4/aburweism/RNAseq/Human_genome/Ensembl/GRCh37/Bowtie2Index/genome
Output: -o /disk4/aburweism/RNAseq1/Alignment/Human_brain
Gene model annotations /disk4/aburweism/RNAseq/GeneModel/Hs_ensembl_37.gtf

Tophat procedure produces several results files. “Most of these files are internal, intermediate files that are generated for use within the pipeline” [72]. These output files include *accepted_hits.bam* which represents a list of the read alignments in a SAM format, *junctions.bed* represents the track of junctions reported by TopHat, *insertions.bed* and *deletions.bed*, and *Logs* files which contain information files about the process, and one of these files represent Bowtie2 alignment results. All Bowtie and Tophat results were examined and indicated that overall alignment rates were above 80%.

3.2.2.4 Quantification

Accurate quantification of the expression levels of the transcript is one of the cores of the RNA sequencing. This requires the correct identification of each isoform of a gene produced from each read. Cufflinks performs a dual function as identifying the transcripts from each of the mapping files, then merging all the transcripts to generate the master reference [66]. In our analysis, Cufflinks used the *.bam* alignment files (*accepted_hits.bam*) from TopHat output as input and assembled the transcripts. We ran Cufflinks on all human and mouse files separately, and obtained twenty GTF files. Cuffmerge was used to merge all ten files into one GTF file. After getting the *merged.gtf* file through the Cufflinks, the last step of our RNA-seq analysis is to use Cuffdiff analysis to estimate transcript abundances. Cuffdiff uses the master *merged.gtf* for the reference annotation and *.bam* mapping files, and then checks the read counts from every sample in *merged.gtf*. For reads that map to the multiple locations, Cuffdiff uses the genome to correct them. Cuffdiff provides an *isoform_exp.diff* file which provides detailed into the gene expression. Gene expression was measured by FPKM (Fragment Per Kilobase of transcript per Million mapped reads). Table 3.3 shows a small portion of the gene expression file.

Table 3.3: Gene expression data.

tracking_id	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	Brain_FPKM	Brain_conf_lo	Brain_conf_hi
TCONS_00000001	ENST00000456328	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1657	0.000263302	0	0.115491
TCONS_00000002	ENST00000450305	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1534	0.13925	0	29.3336
TCONS_00000003	ENST00000450305	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1370	0.114713	0	34.4777
TCONS_00000004	ENST00000450305	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1457	0	0	0
TCONS_00000005	ENST00000515242	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1653	0.040055	0	28.253
TCONS_00000006	ENST00000518655	XLOC_000001	DDX11L1	TSS1	1:11868-31109	1483	0	0	0
TCONS_00000007	ENST00000450305	XLOC_000001	DDX11L1	TSS2	1:11868-31109	632	0.00941883	0	1.29658
TCONS_00000008	ENST00000473358	XLOC_000002	MIR1302-10	TSS3	1:11868-31109	712	0.0102405	0	1.46084
TCONS_00000009	ENST00000469289	XLOC_000002	MIR1302-10	TSS4	1:11868-31109	535	0.017981	0	1.99526
TCONS_00000010	ENST00000408384	XLOC_000002	MIR1302-10	TSS4	1:11868-31109	138	0	0	0
TCONS_00000011	ENST00000594647	XLOC_000003	AL627309.1	TSS5	1:53048-54936	126	4.30317	0	19.9358
TCONS_00000012	ENST00000492842	XLOC_000004	OR4G11P	TSS6	1:62947-63887	940	0	0	0
TCONS_00000013	ENST00000335137	XLOC_000005	OR4F5	TSS7	1:69090-70008	918	0	0	0
TCONS_00000014	ENST00000442987	XLOC_000006	RP11-34P13.10	TSS8	1:89294-134836	3812	1.60713	0	7.36599
TCONS_00000015	ENST00000496488	XLOC_000007	RP11-34P13.9	TSS9	1:160445-161525	457	0.379247	0	1.5942
TCONS_00000016	ENST00000440038	XLOC_000008	RP4-669L17.10	TSS10	1:317719-461954	746	1.51802	0	61.6643
TCONS_00000017	ENST00000440038	XLOC_000008	RP4-669L17.10	TSS10	1:317719-461954	1564	0	0	0
TCONS_00000018	ENST00000440038	XLOC_000008	RP4-669L17.10	TSS10	1:317719-461954	1513	0	0	0
TCONS_00000019	ENST00000440038	XLOC_000008	RP4-669L17.10	TSS10	1:317719-461954	3291	0	0	0
TCONS_00000020	ENST00000426316	XLOC_000008	RP4-669L17.11	TSS10	1:317719-461954	468	0.0630354	0	3.02089
TCONS_00000021	ENST00000432964	XLOC_000008	RP4-669L17.10	TSS11	1:317719-461954	575	1.70341	0	78.9593
TCONS_00000022	ENST00000423728	XLOC_000008	RP4-669L17.10	TSS11	1:317719-461954	573	0.0211795	0	1.83326
TCONS_00000023	ENST00000440038	XLOC_000008	RP4-669L17.10	TSS11	1:317719-461954	1558	0.0713909	0	25.7691
TCONS_00000024	ENST00000601486	XLOC_000008	RP4-669L17.10	TSS12	1:317719-461954	696	0	0	0

Chapter 4 - Relationship between Repetitive DNA Elements and Gene Expression using Regression Models

4.1 Introduction

Finding the relationships among a set of variables that are subjected to random fluctuations is the ultimate goal in many statistical analyses. Regression analysis exemplifies the case in which one aims to explore the association between one or more response (dependent) variables and one or more explanatory (predictor) variables, then assess the influence of the explanatory variables on the response variables [73].

Regression analysis is a statistical technique for modeling the relationships among variables. This includes estimating the parameters of the regression model, examining the strength and direction of the relationships, and assessing the estimated model.

Regression models are divided into two major types, parametric and nonparametric. In parametric regression, the usual way of writing the regression function $f(x)$ as $f(x;\beta)$. Therefore, we are making the assumption that the functional form of the regression function f is known, except for the values of the parameters β. Thus, the word parametric comes from the fact that the regression model can only be specified using a finite number of parameters.

In general, parametric regression models are divided into two classes, linear and nonlinear. The crucial point for the linear regression models is that they are linear in the parameters, whereas the variables $x's$ can include square roots, higher powers, and other transformations of the original measurements. Additionally, an important feature of the linear regression models is that the derivative of the expectation function with respect to

any of parameters is parameter-free terms. This contrasts with the nonlinear regression models, where at least one derivative of the expectation function with respect to the parameters will depend on one or more of the parameters.

Regression models have been broadly used in various fields of science, including genetics, and their applications have significantly increased in the past few decades. Their uses include combining datasets from various sources and developing predictive models for medical and genetic research, which offer risk assessment and treatment options. Predictive models in the field of genetics have also been developed using this method.

In order to characterize the potential impact of repetitive DNA elements on the gene expression levels in human and mouse, different regression approaches were used, including standard multiple regression models, penalized regression models, and multivariate regression models. Explanatory (predictor) variables were represented by repeat families (repFamily) in the standard regression models and repeat names (repName) in the penalized regression models. The response (dependent) variable was represented by gene expression levels in different human and mouse tissues. All models were fitted based on two locations upstream from the genes (promoter region of genes) — 2,000 base pairs (2kbp) and 20,000 base pairs (20kbp). These two locations were used to evaluate the effect of the repeats on the gene expression levels based on the distance upstream from the genes because most of the long TEs, such as LINE1, are truncated and lack promoter content compared to the short TEs, such as Alu's.

4.2 Data Description

To determine the association between repetitive DNA elements and gene expression and to determine their potential impact on human and mouse gene expression, different dataset were used, including repetitive DNA locations (repeatMasker dataset), genomic regions for gene promoters, gene expressions, and Human BodyMap 2.0 gene expression dataset.

4.2.1 Repetitive DNA Locations

Repetitive DNA locations for the human genome (hg19) and mouse (mm10) were downloaded from the UCSC Genome Browser [46]. These repeats were identified by the RepeatMasker program (www.repeatmasker.org), using consensus repeat sequences in RepBase [38]. There were 5,298,130 human repetitive elements classified to 16 repeat classes (repClass), 45 repeat families (repFamily) and 1,395 repeat names (repName). In the mouse, 5,147,736 repetitive elements classified to 16 repeat classes, 47 repeat families, and 1,554 repeat names. Table 4.1 shows a small portion of mouse repetitive DNA file description.

4.2.2 Genomic Regions

The promoter regions 2kbp and 20kbp upstream of the genes for human and mouse genomes were also obtained from RepeatMasker track of the UCSC Genome Browser. Then the customized promoter regions were created based on the highly-expressed transcripts (isoforms) by merging those promoters with the gene expression data.

4.2.3 Gene Expression (Transcript Expression) Datasets

The gene expression results from Cuffdiff in RNA-seq analysis, as noted in Chapter three, were used to obtain the highly-expressed transcripts in all human and mouse tissues.

4.2.4 Human BodyMap 2.0 Dataset

The Human BodyMap 2.0 dataset (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/) was generated using Illumina HiSeq 2000 instruments. It consisted of 16 different human tissues, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. To quantify the gene expression levels in the Human BodyMap 2.0 dataset, we analyzed the dataset using the same RNA-seq pipeline that we used to create our dataset.

4.3 Data preparation

We wrote an R script (See APPENDIX A2) to create genomic ranges for both repetitive elements and promoter regions. Gene expression data from RNA-seq analysis was used to obtain the highly-expressed transcript isoform for each gene within the tissue. We also created a new customized promoter region based on the highly-expressed transcripts for each tissue. Furthermore, we created genomic ranges for the new promoters to find the overlap and count the total number of base pairs for each repeat family and repeat name. The final design matrix for regression models was created by merging the customized promoter with the gene expression dataset. We also restricted our analysis to protein-coding genes by excluding other non-coding protein from the data.

Table 4.1: Description of the mouse repetitive DNA file.

bin	swScore	milliDiv	milliDel	milliIns	genoName	genoStart	genoEnd	genoLeft	strand	repName	repClass	repFamily	repStart	repEnd	repLeft	id
607	12955	105	9	10	chr1	3000000	3002128	-192469843	-	L1_Mus3	LINE	L1	-3055	3592	1466	1
607	1216	268	31	105	chr1	3003152	3003994	-192467977	-	L1Md_F	LINE	L1	-5902	617	1	2
607	234	279	0	0	chr1	3003993	3004054	-192467917	-	L1_Mus3	LINE	L1	-6034	297	237	3
607	3685	199	21	14	chr1	3004040	3004206	-192467765	+	L1_Rod	LINE	L1	1321	1492	-4355	4
607	376	62	31	0	chr1	3004206	3004270	-192467701	+	(CAAA)n	Simple_repeat	Simple_repeat	4	69	0	5
607	3685	199	21	14	chr1	3004270	3005001	-192466970	+	L1_Rod	LINE	L1	1493	2224	-3623	4
607	1280	221	43	62	chr1	3005001	3005439	-192466532	+	L1_Rod	LINE	L1	2425	2854	-2993	4
607	4853	226	62	20	chr1	3005460	3005548	-192466423	+	Lx9	LINE	L1	6309	6394	-1250	6
607	198	0	0	0	chr1	3005548	3005570	-192466401	+	(CAAAA)n	Simple_repeat	Simple_repeat	2	23	0	7
607	4853	226	62	20	chr1	3005570	3006764	-192465207	+	Lx9	LINE	L1	6395	7644	0	6

4.4 Methods

Multiple linear regression (MLR), penalized regression including LASSO, elastic net, and multivariate multiple linear regression (MMLR) models were used to assess the potential influence of both repeat family (repFamily) and repeat name (repName) on the gene expression levels in human and mouse tissues. In order to fit these models, two different datasets were used. Gene expression levels for ten human and ten mouse tissues were used as response (dependent) variables in all of the models.

First, multiple linear regression models were used to determine the impact of the repeat families (repFamily) on the gene expression levels in the two chosen locations, 2000 base pairs (2kbp) and 20,000 base pairs (20kbp) upstream of the genes. Second, penalized regression models, including LASSO, and elastic net were used to determine the impact of repeats by using repeat names (repName) instead of repeat families. Third, multivariate multiple linear regression models were used to investigate the influence of repeats on gene expression in all tissues at the same time.

All data mining and statistical analysis were done in R language [74] using R/Bioconductor packages "*IRanges*" (ver. 2.10.2) and "*GenomicRanges*" (ver. 1.28.4) [75], "*biomaRt"* (ver. 2.32.1) [76], and R packages "*gplots"* (ver. 3.0.1) [77], "*lattice"* (ver. 0.20.35), "*car"* (ver. 2.1.5) [78], *glment* (ver. 2.0.10), *parallel* (ver. 3.4.1) [79], *doParallel* (ver. 1.0.10), *reshape2* (ver. 1.4.2) , and *stringr* (ver. 1.2.0).

4.4.1 Multiple Linear Regression Models

The multiple linear regression model with data $(X_{i1}, X_{i2}, X_{i3}, \dots, X_{ip}, Y_i)$,$i = 1, \dots, n$ where $X_{ip}{}'s$ are the explanatory (predictor) variables and Y_i is the response (dependent) variable of the i^{th} observation, as given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{1}$$

Using matrix notation, the model, can be written more concisely as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{12} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n\times(p+1)} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_n \end{bmatrix}_{n\times 1} \tag{2}$$

Alternatively,

$$\underbrace{\boldsymbol{Y}}_{(n\times 1)} = \underbrace{\boldsymbol{X}}_{n\times(p+1)} \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times 1)} \tag{3}$$

4.4.1.1 Least Squares Method

The least squares method (also known as “ordinary least squares,” “OLS”), is one of the most commonly used techniques for estimating parameters in regression models. The mathematical concept of least squares is the basis for several methods to fit particular types of curves and surfaces to data. OLS, alternately referred to as minimizer of the residual sum of squared errors (RSS)

$$RSS(\beta) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{n} X_{ij}\, \beta_j \right)^2 \tag{4}$$

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta}) \tag{5}$$

Based on the *Gauss-Markov* theorem, the OLS estimators are the Best Linear, Unbiased and Efficient estimator (BLUE), where the best is defined regarding minimum variance. We know that an OLS estimator of the unknown population parameters $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} \quad (6)$$

4.4.1.2 Assumptions for Multiple Linear Regression

The researchers must define the assumptions related to the original data before they can run a comprehensive regression analysis. Ignoring or violating these assumptions contributes to incorrect validity estimates or inaccurate results. The multiple linear regression assumptions that are identified as primary concerns in the research are linearity, multicollinearity, homoscedasticity, normality, and independence of errors. In our study, we discussed each assumption in detail by defining the assumption, detecting the violation and proposing remedial measures in case violations occur. Then all required tests were done using R to assess our model's validity.

- **Linearity Assumption**

Linearity denotes the response variable as a linear function of the explanatory variables. Some researchers contend that this assumption is the most important, as it directly relates to the bias of the results of the whole analysis. Multiple regression can accurately estimate the relationship between the response and explanatory variables when the relationship is linear. In real datasets, the chance of non-linear relationships is high; therefore, it is necessary to examine the linearity assumption [80].

Violation of linearity is highly serious because all the estimates of the regression model, including regression coefficients, standard errors, and tests of statistical

significance, may be biased. If the relationship between the response and explanatory variables is not linear, the results of the regression analysis will under - or overestimate the true relationship and increase the risk of committing Type I and Type II errors.

Residual plots presenting the standardized residuals vs. the fitted values and are very helpful in detecting violations of linearity. The residuals magnify the departures from linearity. If no departure from linearity occurs, we would expect to see a random scatter around the horizontal line. Any systematic clustering/pattern of the residuals suggests a violation. Data transformation is the remedial measure of linearity.

- **Multicollinearity Assumption**

Multicollinearity (Collinearity) refers to the assumption that the explanatory (predictor) variables are correlated. Multicollinearity appears when two or more predictor variables are moderately or highly correlated. If this assumption is not satisfied, correlation is present. Multicollinearity can result in unusual, and misleading results or inflated standard errors. Interpretations and conclusions based on the size of the regression coefficients, their standard errors or associated t-tests may be misleading due to the effects of multicollinearity. Other informal signs of multicollinearity are:

- Regression coefficients change drastically when adding or deleting an X variable.
- A regression coefficient is negative when, theoretically, the response variable should increase with increasing values of that predictor variable, or the regression coefficient is positive when, theoretically, Y should decrease with increasing values of that X variable.

- None of the coefficients have a significant t statistic, but the overall F test for the fitted model is significant.
- Coefficients have a nonsignificant t statistic, even though on theoretical grounds, that predictor variable should provide substantial information about the response variable.
- High pairwise correlations between the X variables are noted. (Exception: three or more predictor variables can be multicollinear without having high pairwise correlations).

Multicollinearity can be detected in several ways:

- Investigate the correlation matrix of the predictor variables and look for high correlation coefficients.
- Determine the tolerance levels. Tolerance measures ($Tolerance\ (\boldsymbol{T}) = 1 - R^2$) the influence of one predictor variable on all other predictor variables. Tolerance levels for correlations range from zero (no independence) to one (completely independent). Tolerance values of 0.10 or less Indicate that there may be severe multicollinearity.
- Determine the Variance Inflation Factor (VIF). The VIF (($VIF = 1/(1 - R^2)$) is an index of the amount that the variance of each regression coefficient is increased over that of uncorrelated predictors. When a strong linear association occurs between predictor variables, the associated VIF is large and is evidence of multicollinearity. “The rule of thumb for a large VIF value is ten” [80].

Multicollinearity can be fixed by using the Ridge regression, Principal component regression or Omitting the correlated variables.

- **Homoscedasticity (constant variance) Assumption**

The assumption of homoscedasticity says that the error variance is the same across all levels of the predictor variables. In other words, the degree of random noise is the same, regardless of the value of the predictors. We often have heteroscedasticity, where the variance is a function of the predictor variables.

If homoscedasticity is violated, the error variance does not bias the coefficient estimates but does affect efficiency. Most often, the standard error will be smaller than the real standard error; therefore, the t statistic and p-values will be incorrect. If heteroscedasticity causes OLS to underestimate the SE and overestimate t-statistic of the estimated coefficients, some of the estimated coefficients which are not statistically significant may incorrectly appear to be significant. The opposite case can also occur.

Homoscedasticity can be checked by visual examination or by using formal statistical tests. Residual plots showing the standardized residuals vs. the fitted values are very helpful in detecting heteroscedasticity violation. Heteroscedasticity is designated when the scatter is not even, typically appearing as fan (funnel) or butterfly shapes.

Various tests can be used to detect the heteroscedasticity, such as Levene's test, Breusch-Pagan test, White test, and Goldfeld-Quandt test. In our study, Levene's test was used because it is more robust to departures from normality assumption.

Data transformation or weighted least squares (WLS) can be used as a remedy for heteroscedasticity

- **Normality Assumption**

With large sample sizes, the normality assumption is not critical unless we use our fitted models to predict new observations. Multiple regression presumes that variables have normal distributions [81, 82]. This means that errors are normally distributed This assumption is based on the shape of the normal distribution and gives the researcher knowledge about the values to expect. Violation of normally assumption can distort relationships and significance tests.

Normality can be detected by visually using Q-Q Plots of residuals or the histogram of residuals with a superimposed normal curve that indicates distribution. Several statistical tests can also be run, such as the Kolmogorov-Smirnov test (KS test), the Shapiro-Wilk test, the Anderson-Darling test, or the Correlation test of normality.

4.4.2 Penalized Regression Models using LASSO and Elastic Net

Penalized regression methods, also called regularization or shrinkage approaches, have been developed to overcome the challenges caused by high-dimensional data [83]. High dimensional data demonstrates many practical problems and computational issues when using standard regression. To deal with these problems, variable selection, and shrinkage estimation have become popular solutions. The method of penalized least squares (PLS), which is equivalent to penalized maximum likelihood, helps to deal with these issues by putting constraints on the values of the estimated parameters. The penalized least squares method (PLS) has been shown to improve OLS estimation and prediction in the case of high dimensional data.

In general, the penalized least squares method (PLS) is said to minimize OLS subject to penalty term $P(\beta) \leq C$, where $P(\beta)$ is a specific penalty function of β, and C, is a tuning parameter. This constrained optimization problem is equivalent to the Lagrangian optimization which minimizes the residual sum of squares, as follows:

$$\text{PLS} = \text{OLS} + \text{Penalty} = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda P(\boldsymbol{\beta}),$$

where λ is the tuning parameter (model complexity) that controls the strength of shrinkage. For example, when $\lambda = 0$ (no shrinkage is performed), no penalty is applied and we have the ordinary least squares regression. When λ increases, more weight is given to the penalty term.

Penalized regression methods, such as LASSO and elastic net, have been developed to overcome the limitation of traditional variable selection methods when the number of predictors is large.

4.4.2.1 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a shrinkage method proposed by Tibshirani (1996) [84]. Unlike ridge regression, which does not provide variable selection and fails to provide a parsimonious model with few parameters, LASSO performs both estimation and variable selection simultaneously in one stage. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. It has become popular for high-dimensional estimation problems due to its statistical accuracy for model prediction and variable selection, coupled with its computational feasibility, interpretability, and numerical stability.

- **L1 Regularization**

LASSO regression based on the L1-norm penalty is useful for fitting a wide variety of models. L1 regularization adds a penalty term to the loss function. Since each non-zero coefficient adds to the penalty, it forces weak predictors to be zero as coefficients. Thus, L1 regularization produces sparse solutions, inherently performing feature selection.

- **Linear Regression Via LASSO**

Consider the multiple linear regression model with data $(X_{i1}, X_{i2}, X_{i3}, \dots, X_{ip}, Y_i)$,$i = 1, \dots, n$ and $X_{ip}{}'s$ are the explanatory (predictor) variables and Y_i is the dependent (response) variable of the i^{th} observation. The LASSO estimate is defined by

$$\widehat{\boldsymbol{\beta}}^{\,lasso} = \underset{\boldsymbol{\beta}}{\arg min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 \qquad (7)$$

$$\text{Subjected to } \lambda\sum_{j=1}^{p}\left\|\beta_j\right\|$$

We can also write the LASSO problem in the equivalent *Lagrangian* form.

$$\hat{\boldsymbol{\beta}}^{lasso} = \underbrace{\arg\min_{\boldsymbol{\beta}}}\{\underbrace{\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2}_{loss-function} + \underbrace{\lambda\sum_{j=1}^{p}\left\|\beta_j\right\|}_{\substack{Penalty \\ L1-norm}}\} \tag{8}$$

The regularization parameter lambda λ governs the degree to which coefficients are penalized. The R *glmnet* package [85] was used to fit the LASSO model, and the optimal λ obtained using cross validation.

- **Geometric Interpretation for LASSO**

In order to interpret how LASSO works, we compare it with the ridge regression. For simplicity and visualization sake, we used only two predictors, as shown in Figure 4.1

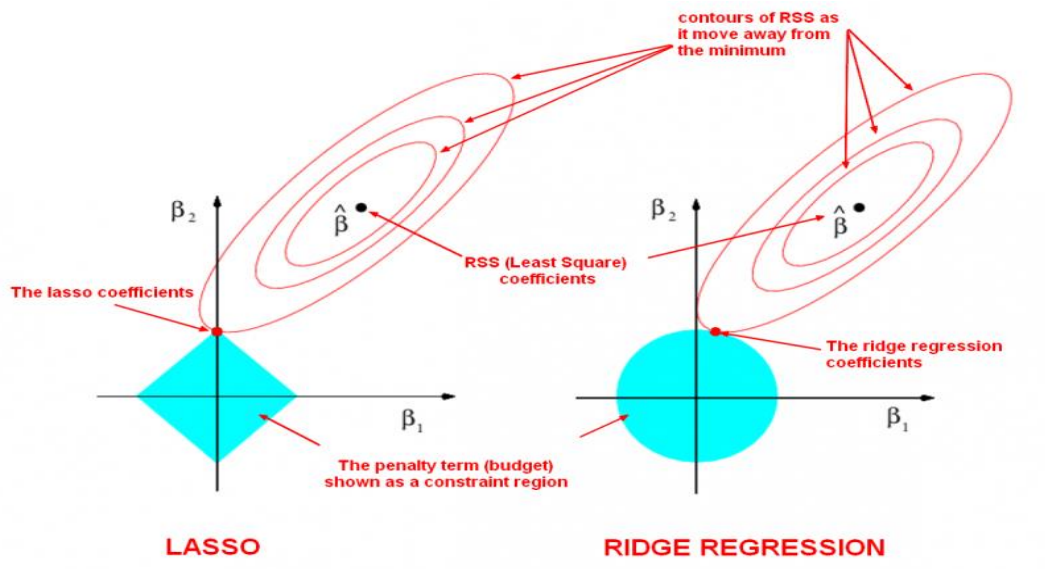

Figure 4.1: A geometrical interpretation of LASSO in two dimensions (Hastie et al. 2009)

Figure 4.1 illustrates the geometric interpretation of LASSO (left) and a ridge regression (right) in the two-dimensional case. In both panels, the center point of the ellipse is $\hat{\beta}$ (OLS

estimates). The ellipse contour corresponds to some specific residual sum of square values. The area inside the blue diamond around the origin satisfies the LASSO restriction. It denotes that $\beta = (\beta_1, \cdots, \beta_p)^T$ inside the solid blue diamond, which satisfies the constraint $\lambda \sum_{j=1}^{p} \|\beta_j\| \leq c$. Thus, minimizing the residual sum of squares according to the constraint corresponds to the contour tangent of the diamond. The LASSO solution is the first place that the contours touch the diamond; this will sometimes occur at a corner, corresponding to a zero coefficient.

- **Selection of the Model complexity (Tuning Parameter) λ**

To select the model complexity λ, we used the cross-validation (CV) method for LASSO models, as suggested by Tibshirani (1996). Cross-validation is an estimate of the expected generalization error for each λ and λ can sensibly be chosen as the minimizer of this estimate. The *cv.glmnet* function returns two values of λ. The minimizer (*lambda.min)*, and the always larger (*lambda.1se)*, which is a heuristic choice of λ producing a less complex model, rates the performance in terms of estimated expected generalization error is within one standard error of the minimum.

4.4.2.2 Elastic Net Regression

The elastic net is also a shrinkage method proposed by Zou and Hastie [2005] [86]. This method is based on a compromise between the LASSO and ridge regression penalties. Elastic net performs variable selection and dimension reduction. It uses LASSO with an L1 penalty to perform variable selection and ridge with an L2 penalty to shrink the model coefficients.

• **Linear Regression Via Elastic Net**

Consider the multiple linear regression model with data $(X_{i1}, X_{i2}, X_{i3}, \dots, X_{ip}, Y_i)$,$i = 1, \dots, n$ and $X_{ip}{}'s$ are the explanatory (predictor) variables and Y_i is the response (dependent) variable of the i^{th} observation. The elastic net estimate is defined by

$$\hat{\beta}^{\text{elastic}} = \arg\min_{\beta}\left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda P_{\alpha}(\beta)\right], \quad (9)$$

where $$P_{\alpha}(\beta) = \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\|\beta_j\|,$$

and λ is the model complexity (tuning parameter) which controls the strength of shrinkage, and $\alpha \in [0,1]$ is the penalty weight which controls the tradeoff between the LASSO and ridge penalties. When $\alpha = 1$, it reduces to the L1 or LASSO penalty, while with $\alpha = 0$, it reduces to the squared L2, corresponding to the ridge penalty.

In order to fit elastic net penalized regression model, we need to find the optimal values of both λ and α that minimize the model mean square error. We wrote an R script that use a two-layer cross-validation to simultaneously determine the best combination of λ and α.

4.4.3 Multivariate Multiple Linear Regression (MMLR)

Multivariate multiple linear regression (MMLR), in general, is an extension of the multiple linear regression (MLR). In both techniques, we try to explain and interpret the possible linear relationship between certain explanatory (predictor) variables and one or more response (dependent) variables [87]. MMLR examines each response separately in relation to a linear combination of all predictor variables without imposing any structure across the several resulting regression equations. MMLR estimates the same coefficients and standard errors as would be obtained by using separate OLS regressions for each response variable [88].

The MMLR model simultaneously tests the effect of multiple predictors on multiple responses. The advantages of using MMLR are that we can conduct tests of the coefficients across the different models. However, MMLR may also be used to test an omnibus null hypothesis and composite hypotheses for a model, which distinguish it from an OLS regression.

In the multivariate case, we consider the problem of modeling the association between k dependent (response) variables $Y_1, Y_2, Y_3, \cdots Y_k$ and a single set of explanatory (predictor) variables $X_1, X_2, X_3, \cdots X_p$. Each response variable is assumed to follow its own regression model where

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + \cdots + \beta_{p1}X_p + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}X_1 + \beta_{22}X_2 + \beta_{32}X_3 + \cdots + \beta_{p2}X_p + \varepsilon_2 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ Y_k &= \beta_{0k} + \beta_{1k}X_1 + \beta_{2k}X_2 + \beta_{3k}X_3 + \cdots + \beta_{pk}X_p + \varepsilon_k \end{aligned} \tag{10}$$

The error term $\boldsymbol{\varepsilon^T} = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_k]$ has expectation $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and variance matrix $V(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. The errors terms associated with different responses may be correlated.

Using matrix notation, the MMLR model is more concisely defined by

$$\boldsymbol{X}_{(n\times(p+1))} = \begin{bmatrix} X_{10} & X_{11} & \cdots & X_{1p} \\ X_{20} & X_{21} & & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

$$\boldsymbol{Y}_{(n\times k)} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1K} \\ Y_{21} & Y_{22} & & Y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nk} \end{bmatrix} = \left[\boldsymbol{Y}_{(1)} \vdots\ \boldsymbol{Y}_{(2)} \vdots \boldsymbol{Y}_{(3)} \vdots \cdots \vdots \boldsymbol{Y}_{(k)}\right]$$

$$\boldsymbol{\beta}_{((p+1)\times k)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0k} \\ \beta_{11} & \beta_{12} & & \beta_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pk} \end{bmatrix} = \left[\boldsymbol{\beta}_{(1)} \vdots\ \boldsymbol{\beta}_{(2)} \vdots \boldsymbol{\beta}_{(3)} \vdots \cdots \vdots \boldsymbol{\beta}_{(k)}\right]$$

$$\boldsymbol{\varepsilon}_{(n\times k)} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1k} \\ \varepsilon_{21} & \varepsilon_{22} & & \varepsilon_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nk} \end{bmatrix} = \left[\boldsymbol{\varepsilon}_{(1)} \vdots\ \boldsymbol{\varepsilon}_{(2)} \vdots \boldsymbol{\varepsilon}_{(3)} \vdots \cdots \vdots \boldsymbol{\varepsilon}_{(k)}\right]$$

Then the multivariate multiple linear regression model is

$$\underbrace{\boldsymbol{Y}}_{(n\times k)} = \underbrace{\boldsymbol{X}}_{(n\times(p+1))}\ \underbrace{\boldsymbol{\beta}}_{((p+1)\times k)} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times k)}$$

with $E\left(\boldsymbol{\varepsilon}_{(i)}\right) = \boldsymbol{0}$ and $Cov\left(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(j)}\right) = \sigma_{ij}\boldsymbol{I}\ ;\ \ i,j = 1,2,\cdots,k.$

Simply, the i[th] response $\boldsymbol{Y}_{(i)}$ follow linear regression model

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)},\ \ i = 1,2,\cdots k.$$

OLS estimates will be

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}_{(i)}.$$

The MMLR model assumptions are:

1. The errors follow a multivariate normal distribution with means equal to zero.

2. Error variances are equal across observations condition on predictors.
3. Errors have common covariance structure across observations.
4. Independent of observations.

- **Testing the Omnibus Null Hypothesis**

The omnibus null hypothesis states that all regression coefficients equal zero across all response variables. The purpose of the omnibus hypothesis test is to prevent inflation the study-wise alpha level. If separate tests are performed for each response variable, the probability of obtaining a significant false value will increase in direct proportion to the number of response variables being tested; that is, the power of the test decreases. To evaluate the omnibus null hypothesis, multivariate *F* - tests are used, which include *Wilk's Lambda*, *Pillai's trace*, *Lawley-Hotelling's trace* and *Roy's largest root*.

4.5 Results

To implement our study, we used different regression analysis approaches to investigate the association between repetitive DNA elements and gene expression.

4.5.1 Multiple Linear Regression Results

Multiple linear regression models were fitted for the average gene expression and tissue-specific gene expression to investigate the impact of repeat families (repFamily) on the gene expression levels in both 2kbp and 20kbp promoter regions for mouse and human tissues. The stepwise method was used to select the repeat families (repFamily) that have a highly significant impact on the gene expression.

4.5.1.1 Mouse regression model results for the average gene expression in the 2kbp promoter region

Table 4.2 shows the of estimated regression coefficients (unstandardized and standardized) with their corresponding p-values, *Bonferroni* correction p-values, and VIF values of the selected repeat families. Considering the large sample size, our results demonstrated that Alu and B2 elements in the promoter are significantly associated with higher average gene expression. In contrast, L1 and simple repeats results showed a significant negative association with gene expression. Model assumptions such as linearity, normality, and heteroscedasticity were checked visually using residual plot, as shown in Figure 4.2. These indicated no departure from the model assumptions. Moreover, formal tests, such as *Levene's* test for heteroscedasticity (p-value=0.07) and the *Durbin-Watson* test for autocorrelation, demonstrated no violations. Multicollinearity was also checked

using VIF, and we found that all VIF values were less than two, which demonstrates no multicollinearity between predictors.

Table 4.2: Results of fitting a multiple linear regression model to the average gene expression in the mouse 2kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	1.1×10^{-3}	0.1735	2.2×10^{-16}	1.6×10^{-15}	1.0926
B2	4.0×10^{-4}	0.0576	3.2×10^{-10}	2.6×10^{-9}	1.0794
ERV1	2.0×10^{-4}	0.0221	1.2×10^{-2}	9.9×10^{-2}	1.0068
ERVK	-2.0×10^{-4}	-0.0276	1.8×10^{-3}	1.4×10^{-2}	1.0089
ERVL-MaLR	-1.0×10^{-4}	-0.0274	1.9×10^{-3}	1.6×10^{-2}	1.0122
L1	-3.0×10^{-4}	-0.0787	2.2×10^{-16}	1.6×10^{-15}	1.0236
Satellite	-6.0×10^{-4}	-0.0223	1.2×10^{-2}	8.1×10^{-2}	1.0017
Simple_repeat	-1.5×10^{-3}	-0.1049	2.2×10^{-16}	1.6×10^{-15}	1.0088

*Unadjusted p-value cut-off 0.006 †Adjusted p-values (*Bonferroni*) cut-off 0.05

P-values less than 0.006 were deemed significant, after the *Bonferroni* adjustments.

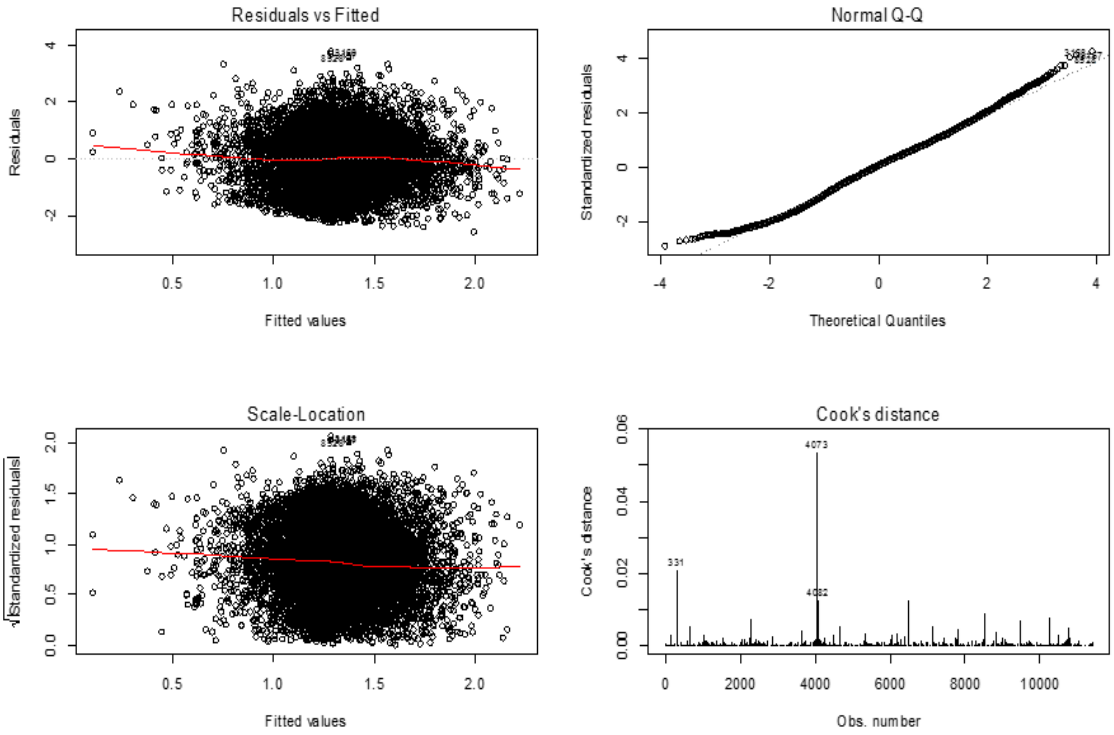

Figure 4.2: Residuals plots for the mouse average gene expression in the 2kbp promoter region.

4.5.1.2 Mouse regression model results for the average gene expression in the 20kbp promoter region

Table 4.3 shows the of estimated regression coefficients with their corresponding p-values, *Bonferroni* correction p-values, and VIF values of the selected repeat families on the average gene expressions. Our results demonstrated Alu, B2, ERV1, and tRNA elements are significantly associated with higher average expression. In contrast, L1, Low_complexity, MIR, and simple repeats results showed a negative a significant association with gene expression. Moreover, we found the effects of Alu and B2 elements decreases at 20kbp when compared with 2kbp. In contrast, the L1 effect at 20kbp is stronger than at 2kbp. Model assumptions were checked visually using residual plot, and no departure from the model assumptions was noted. Furthermore, formal tests, including VIF, *Levene's* (p=0.289), and *Durbin-Watson*, demonstrated no violations.

Table 4.3: Results of fitting a multiple linear regression model to the average gene expression in the mouse 20kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	2.0×10^{-4}	0.2171	2.2×10^{-16}	2.4×10^{-15}	2.2481
B2	1.2×10^{-4}	0.0595	6.2×10^{-7}	7.4×10^{-6}	1.9831
ERV1	5.0×10^{-5}	0.0403	2.5×10^{-6}	3.0×10^{-5}	1.0234
ERVL	4.0×10^{-5}	0.0299	5.0×10^{-4}	5.5×10^{-3}	1.0189
hAT.Charlie	1.3×10^{-4}	0.0204	1.7×10^{-2}	2.1×10^{-1}	1.0209
ID	4.0×10^{-4}	0.0329	3.0×10^{-4}	3.7×10^{-3}	1.1612
L1	-1.0×10^{-4}	-0.1109	2.2×10^{-16}	2.4×10^{-15}	1.2732
Low_complexity	-2.2×10^{-4}	-0.0336	1.0×10^{-4}	1.3×10^{-3}	1.0481
MIR	-1.0×10^{-4}	-0.0330	2.0×10^{-4}	2.4×10^{-3}	1.0991
Simple_repeat	-2.0×10^{-4}	-0.0831	2.2×10^{-16}	2.4×10^{-15}	1.0784
TcMar.Tigger	1.0×10^{-4}	0.0253	2.9×10^{-3}	3.6×10^{-2}	1.0111
tRNA	1.1×10^{-3}	0.0358	2.5×10^{-5}	2.9×10^{-4}	1.0066

*Unadjusted p-value cut-off 0.004 †Adjusted p-values (*Bonferroni*) cut-off 0.05

4.5.1.3 Mouse regression model results for the tissue-specific expression in the 2kbp promoter region

Table 4.4 and Figure 4.3 show a comparison of different repeat families using standardized regression coefficients in the mouse 2kbp tissue-specific genes. We used the standardized regression coefficients to compare the impact of repeats across all tissues. For example, the brain tissue-specific genes showed that, for each standard deviation unit increase in Alu base pairs, the association would decrease the gene expression by 0.1142 standard deviation units. We also found that Alu elements have the highest effect compared with other repeat families; they are associated with decreasing gene expression in the brain and lung tissues, when compared to the other tissues. In contrast, the L1 elements are associated with higher gene expression in the brain, colon, and liver tissues.

Table 4.4: Standardized linear regression coefficients in the mouse 2kbp tissue-specific.

	Standardized Coefficients									
RepFamily	**Brain**	**Colon**	**Embryo**	**Heart**	**Kidney**	**Liver**	**lung**	**Spleen**	**Thymus**	**Uterus**
Alu	-0.1142		0.0331				-0.07521	0.0667	0.0916	
B2	-0.0581							0.0468	0.0248	
ERV1			-0.0216			0.0207	-0.0349			-0.0303
ERVK		0.0184	-0.0276	-0.0410	0.0332	0.0449	-0.0229			-0.0344
ERVL-MaLR			-0.0316			0.0374	-0.0249		-0.0255	-0.0354
L1	0.0258	0.0699	-0.0689	-0.0292		0.1096		-0.0285	-0.0477	-0.0841
Low_complexity	0.0532	-0.0528		0.0265				-0.0314	-0.0192	
Simple_repeat	0.0545		-0.0404				0.0415	-0.0314	-0.04839	0.0447

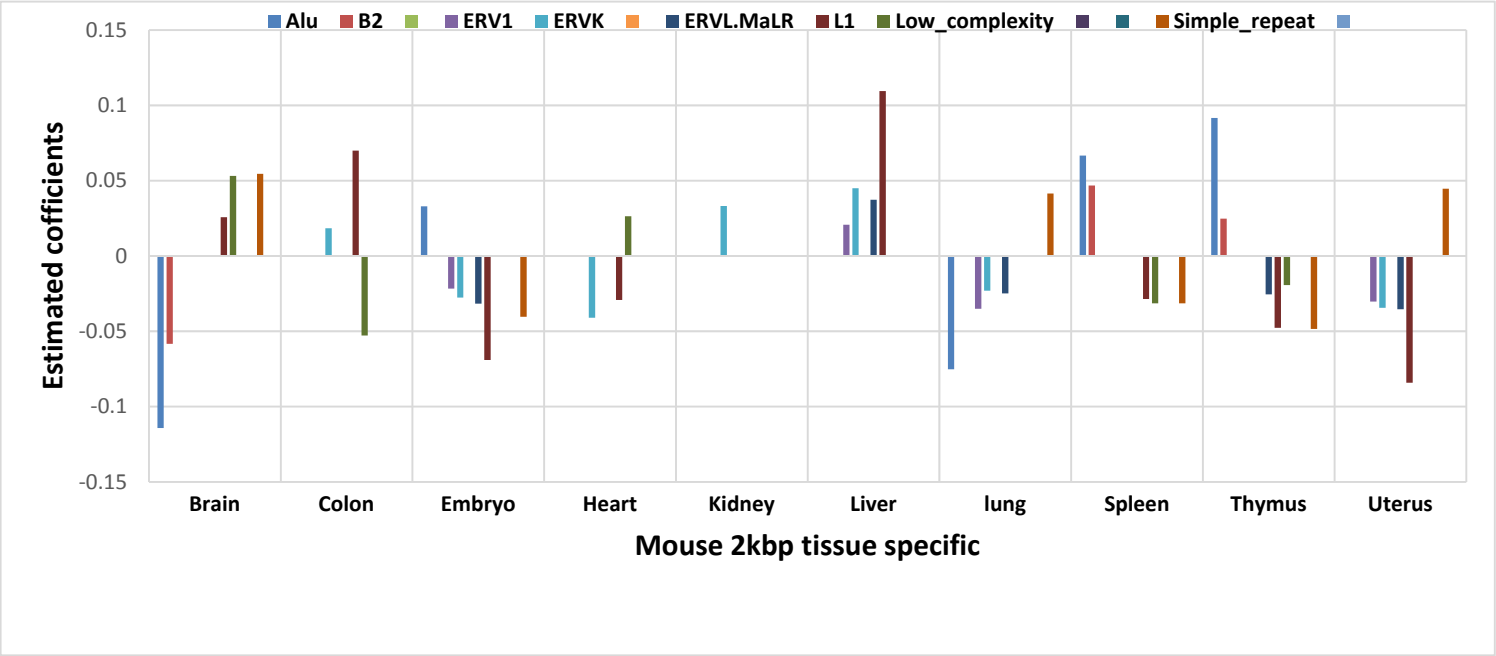

Figure 4.3: Estimated linear model coefficients for the mouse tissue-specific in 2kbp.

4.5.1.4 Mouse regression model results for the tissue-specific expression in the 20kbp promoter region

Table 4.5 and Figure 4.4 show a comparison between the different repeat families using standardized regression coefficients in the mouse 20kbp tissue-specific case. Similar to the 2kbp results, the brain tissue-specific genes demonstrated that each standard deviation unit increase in the Alu base pairs is associated with decreasing gene expression by 0.1101 standard deviation units. We also found that the L1 and Alu elements have the highest effect compared with other repeat families. The Alu elements are also associated with lower gene expression in the brain and lung tissues when compared to the other tissues. In contrast, the L1 elements are associated with higher gene expression in the brain and liver tissues.

Table 4.5: Standardized linear regression coefficients in the mouse 20kbp tissue-specific.

	Standardized Coefficients									
RepFamily	**Brain**	**Colon**	**Embryo**	**Heart**	**Kidney**	**Liver**	**lung**	**Spleen**	**Thymus**	**Uterus**
Alu	-0.1101		0.05013				-0.0868	0.0879	0.1255	
B2	-0.0274									
ERV1	-0.0267		-0.0223	-0.0211		0.0252	-0.0198	0.0213		
ERVK	-0.0249	0.0184	-0.0334	-0.0232	0.0269	0.0491	-0.0285			-0.0409
ERVL-MaLR	0.0556	0.0209		0.0218		0.0205	-0.0279	-0.0640		-0.0235
hAT.Charlie	0.0284							-0.0317		
ID			-0.0823		0.0246					-0.0297
L1	0.1384			-0.0315	0.0283	0.1365	-0.0389	-0.1109	-0.0878	-0.1123
Low_complexity	0.0741	-0.0528	-0.0241					-0.0425		-0.0227
MIR						-0.0392	0.0597		-0.0328	0.0301
Simple_repeat						-0.0252	0.0647		-0.0506	
tRNA	-0.0527		0.0249		-0.0424	-0.034		0.0429	0.0759	

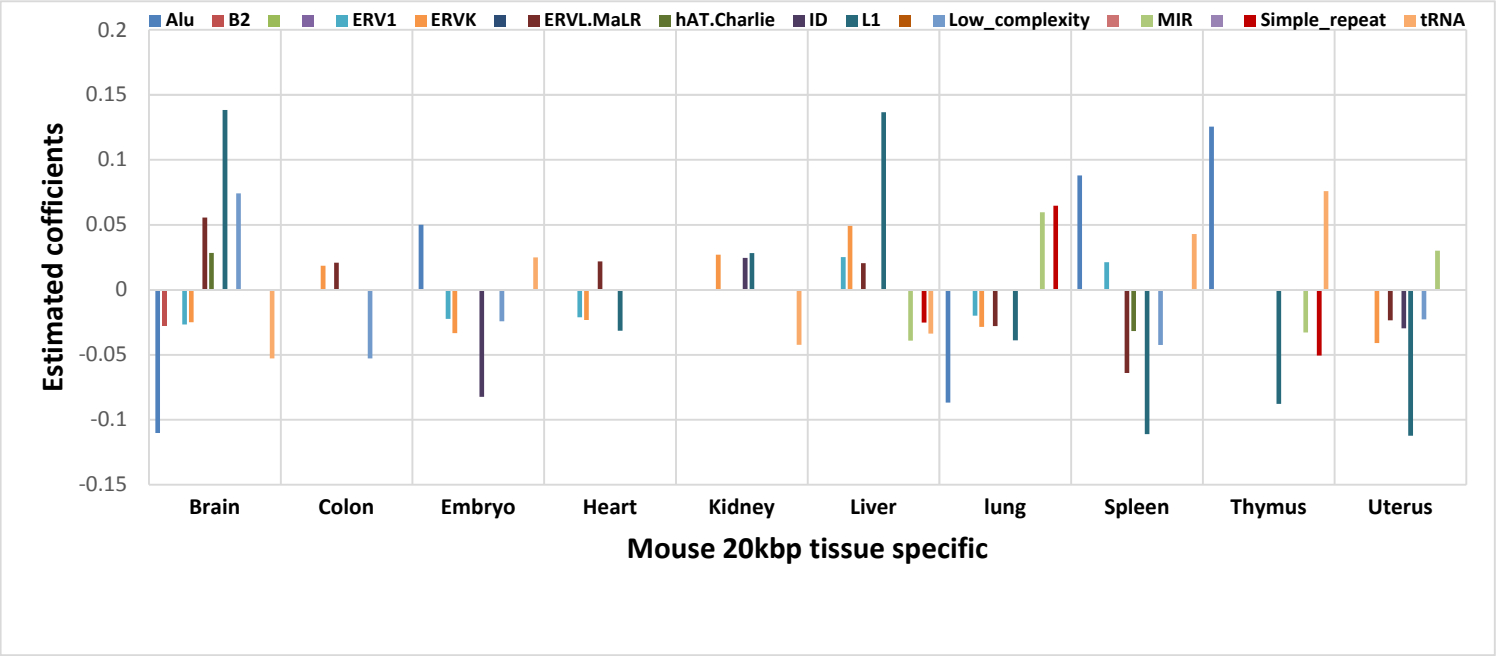

Figure 4.4: Estimated linear model coefficients for the mouse tissue-specific in 20kbp.

4.5.1.5 Human regression model results for the average gene expression in the 2kbp promoter region

Table 4.6 illustrates the estimated regression coefficients (unstandardized and standardized) with their corresponding p-values, *Bonferroni* correction p-values, and VIF values of the selected repeat families. Similar to mouse tissues, our results demonstrated a significant positive association between Alu elements and higher average gene expression. In contrast, ERVL-MaLR, L1, MIR, and simple repeats results showed a significant negative association with the gene expression. Residual plot analysis was used to check the model assumptions, and no departure found. Moreover, formal tests, including VIF, *Levene's* test (p=0.13), and *Durbin-Watson* demonstrated no violations.

Table 4.6: Results of fitting a multiple linear regression model to the average gene expression in the human 2kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	2.1×10^{-4}	0.0655	2.7×10^{-12}	3.0×10^{-11}	1.0463
ERVL	-3.3×10^{-4}	-0.0241	8.6×10^{-3}	9.5×10^{-2}	1.0034
ERVL-MaLR	-3.0×10^{-4}	-0.0372	5.2×10^{-5}	6.0×10^{-4}	1.0105
Gypsy	-6.4×10^{-4}	-0.0180	4.9×10^{-2}	5.4×10^{-1}	1.0018
hAT.Blackjack	-1.1×10^{-3}	-0.0227	1.3×10^{-2}	1.4×10^{-1}	1.0006
hAT.Charlie	2.5×10^{-4}	0.0234	1.1×10^{-2}	1.2×10^{-1}	1.0087
L1	-2.0×10^{-4}	-0.0587	2.5×10^{-10}	2.8×10^{-9}	1.0265
Low_complexity	-4.8×10^{-4}	-0.0300	1.4×10^{-3}	1.6×10^{-2}	1.0589
MIR	-3.1×10^{-4}	-0.0473	3.9×10^{-7}	4.3×10^{-6}	1.0373
Simple_repeat	-7.8×10^{-4}	-0.0515	2.6×10^{-8}	2.9×10^{-7}	1.0221
tRNA	4.0×10^{-3}	0.0262	4.2×10^{-3}	4.6×10^{-2}	1.0009

*Unadjusted p-value cut-off 0.0045 †Adjusted p-values (*Bonferroni*) cut-off 0.05

P values less than 0.0045 were deemed significant, after the *Bonferroni* adjustments.

4.5.1.6 Human regression model results for the average gene expression in the20kbp promoter region

Table 4.7 shows the estimated regression coefficients with their corresponding p-value, *Bonferroni* correction p-values, and VIF values of the selected repeat families. Again, our results demonstrated a significant positive association between Alu elements and higher average gene expression. In addition to Alu elements, we found that hAT.Charlie is also associated with higher average gene expression. In contrast, L1, Low_complexity, MIR, and simple repeats results showed a negative association. The residual plot and formal tests showed no violation of the model assumptions.

Table 4.7: Results of fitting a multiple linear regression model to the average gene expression in the human 20kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	5.1×10^{-5}	0.1499	2.2×10^{-16}	1.8×10^{-15}	1.1923
ERVL-MaLR	-2.1×10^{-5}	-0.0187	4.1×10^{-2}	3.7×10^{-1}	1.0535
hAT.Charlie	1.5×10^{-4}	0.0656	5.1×10^{-13}	4.6×10^{-12}	1.0294
L1	-2.4×10^{-5}	-0.0471	1.8×10^{-6}	1.6×10^{-5}	1.2168
Low_complexity	-2.9×10^{-4}	-0.0536	7.2×10^{-9}	6.4×10^{-8}	1.0707
MIR	-1.5×10^{-4}	-0.0912	1.8×10^{-15}	1.8×10^{-15}	1.1354
Simple_repeat	-2.8×10^{-4}	-0.0748	3.0×10^{-15}	2.7×10^{-15}	1.0442
TcMar	8.0×10^{-4}	0.0322	3.0×10^{-4}	2.9×10^{-3}	1.0033
TcMar.Tigger	5.0×10^{-5}	0.0226	1.2×10^{-2}	1.1×10^{-1}	1.0166

*Unadjusted p-value cut-off 0.005 †Adjusted p-values (*Bonferroni*) cut-off 0.05

P values less than 0.0045 were deemed significant, after the *Bonferroni* adjustments.

4.5.1.7 Human regression model results for the tissue-specific expression in the 2kbp promoter region

Table 4.8 and Figure 4.5 present a comparison between different repeat families using standardized regression coefficients in the human 2kbp tissue-specific case. For example, the brain tissue showed that each standard deviation unit increase in Alu base pairs was associated with lower gene expression by 0.0229 standard deviation units. We also found that simple_repeat, Low_complexity, and L1 elements have the highest effect on brain gene expression compared with other repeat families. In contrast, the L1 elements are associated with higher gene expression in the colon, liver, prostate, and uterus tissues.

Table 4.8: Standardized linear regression coefficients in the human 2kbp tissue-specific.

	Standardized Coefficients									
RepFamily	**Brain**	**Colon**	**Heart**	**Kidney**	**liver**	**lung**	**prostate**	**spleen**	**Thymus**	**Uterus**
Alu	-0.0229	0.0272		0.0873	0.0144	-0.0415	-0.1009	-0.0215	0.0263	-0.0889
ERVK			-0.0219	0.0273					-0.0249	-0.0265
ERVL-MaLR	-0.0328	0.0200		-0.0301			0.0276		-0.0184	0.0274
hAT.Charlie	-0.0451			-0.0195		0.0209	0.0341			0.0368
L1	-0.0679	0.0506	-0.0196	-0.0628	0.0443		0.0460		-0.0301	0.0412
Low_complexity	0.0782	-0.0195			-0.0392	-0.0502		-0.0397	-0.0277	-0.0230
MIR				0.0232			-0.0451		-0.0350	-0.0293
Simple_repeat	0.0844							-0.0192		-0.0229

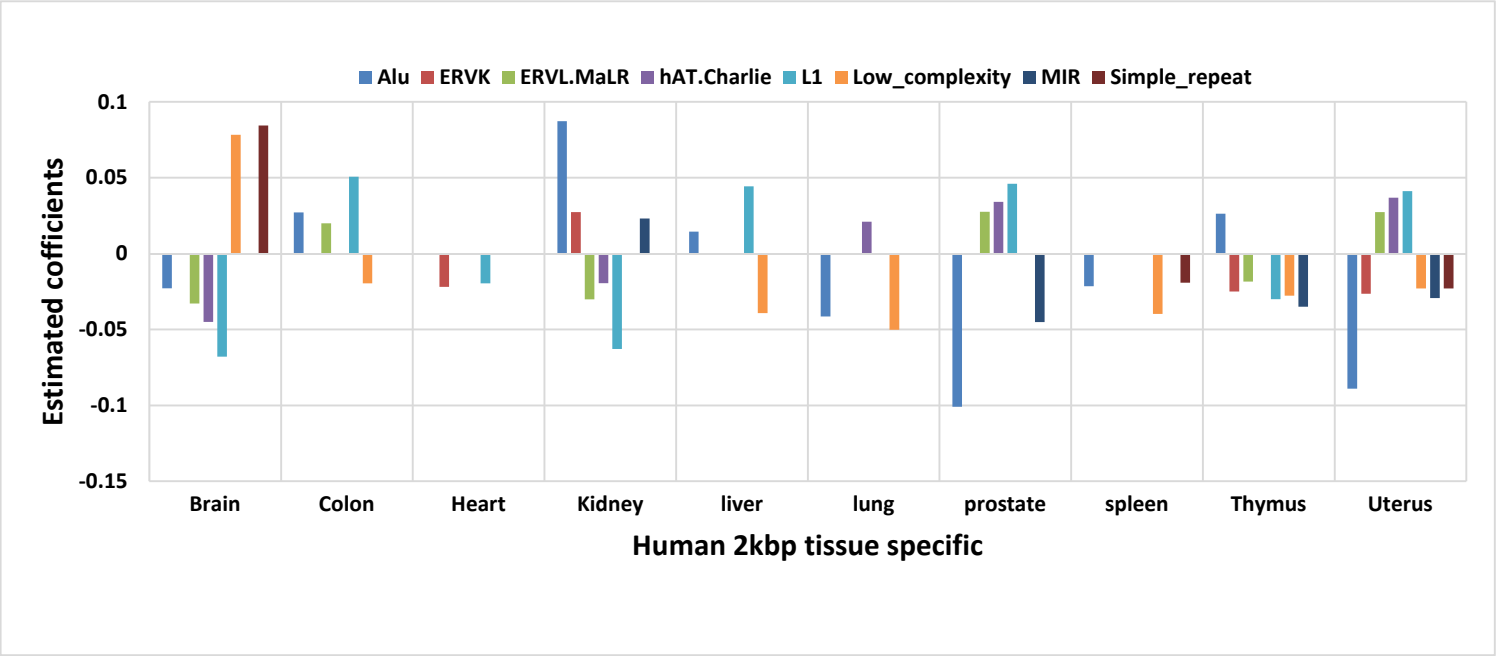

Figure 4.5: Estimated linear model coefficients for the human tissue-specific in 2kbp.

4.5.1.8 Human regression model results for the tissue-specific expression in the 20kbp promoter region

Table 4.9 and Figure 4.6 show a comparison between the different repeat families using standardized regression coefficients in the human 20kbp tissue-specific case. For example, the brain tissue showed that each standard deviation unit increase in L1 base pairs was associated with lower gene expression by 0.0910 standard deviation units. We also found that Low_complexity, simple_repeat, and MIR elements have the highest effect on brain gene expression, compared with the other repeat families.

Table 4.9: Standardized linear regression coefficients in the human 20kbp tissue-specific.

	Standardized Coefficients								
RepFamily	**Brain**	**Heart**	**Heart**	**Liver**	**Lung**	**Prostate**	**Spleen**	**Thymus**	**Thymus**
Alu	0.0153	0.0177	0.0791		-0.0378	-0.0977			-0.0968
ERVL-MaLR	-0.0256	0.0189	-0.0368			0.0302	-0.0229	-0.0644	0.0536
hAT.Charlie	-0.0576	0.0273	-0.0498			0.0757	0.0152		0.0932
L1	-0.0910	0.0194	-0.1167	0.0527	0.0441	0.1154	0.0238	-0.0608	0.1149
Low_complexity	0.0699			-0.0223	-0.0466	0.03490	-0.0188		-0.0221
MIR	0.0582		0.0587		0.0530	-0.0434	-0.0233	-0.0691	-0.0485
Simple_repeat	0.0666		0.0238		-0.0214	-0.0319			-0.0499
TcMar	-0.0664		-0.0560		0.0312		0.0334		0.0834

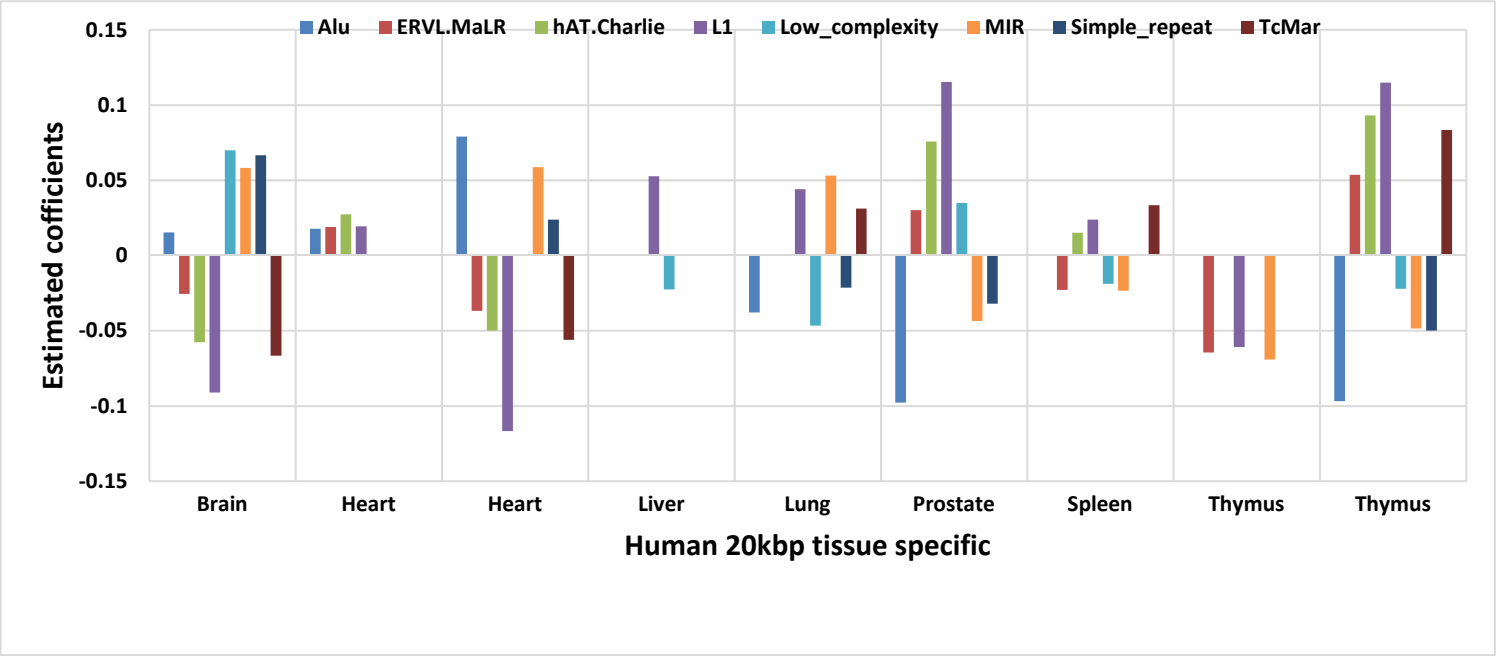

Figure 4.6: Estimated linear model coefficients for the human tissue-specific in 20kbp.

4.5.2 LASSO and Elastic Net Regression Results

LASSO and elastic net regression models were fitted to investigate the impact of repeat names (repName) on the average gene expression levels in both 2kbp and 20kbp promoter regions for both mouse and human tissues.

In order to fit our models with an accurate result from the penalized regression models using LASSO and elastic net, we must find optimal values for model parameters λ and α. In LASSO, we just need to set $\alpha = 1$ in the *glment* argument, and *glmnet* package will automatically find the optimal value of λ using cross-validation function via the cv.glmnet function. In the elastic net, we need to find the optimal α which minimize the mean square error. The *glmnet* package does not include any method to find the optimal α, particularly in the elastic net model. To tackle this issue and find the best combination between α and λ which would minimize model error, we used the *foreach* function to create a two-layer cross-validation to simultaneously find the optimal λ and α for the elastic net model. This function would require us to run *cv.glmnet* at various levels of α, but this would take a long time to perform sequentially, so parallelization was used to increase the speed. We created a vector of α that takes values ranging from 0.1 to 1 to find the optimal α in order to decrease the mean-square error (minMSE) and minMSE + 1 standard error of minMSE (minMSE+1SE). Friedman et al. [85] recommended using minMSE + 1SE to avoid overfitting.

4.5.2.1 Penalized regression model results for mouse in the 2kbp promoter region

Table 4.10 and Figure 4.7 show the results of the error values using various λ and α values. We found that that optimal α and λ values that minimize error using the one standard error methodology are 0.45 and 0.123894, respectively. Figure 4.8 shows the cross-validation curve for simultaneously fitted mouse 2kbp model using *glmnet* on the average gene expression data.

Table 4.10: Errors values using various values of λ and α simultaneously in both minimum error and one-standard-error cases in the mouse 2kbp model.

Alpha	*lambda.1se*	*error.1se*	*lambda.min*	*error.min*
0.1	0.507996	2.28707	0.182564	2.254349
0.15	0.371683	2.289891	0.133576	2.254004
0.2	0.278762	2.288308	0.100182	2.253886
0.25	0.22301	2.287329	0.080146	2.253834
0.3	0.185842	2.28667	0.066788	2.253809
0.35	0.159293	2.286198	0.057247	2.253796
0.4	0.139381	2.285843	0.050091	2.253788
0.45	**0.123894**	**2.285565**	0.044525	2.253784
0.5	0.122377	2.291318	0.040073	2.253783
0.55	0.111251	2.291123	0.03643	2.253783
0.6	0.10198	2.290962	0.033394	2.253783
0.65	0.094136	2.290825	0.030825	2.253783
0.7	0.087412	2.290708	0.028623	2.253784
0.75	0.081584	2.290605	0.026715	2.253784
0.8	0.076485	2.290517	0.025045	2.253785
0.85	0.071986	2.290438	0.023572	2.253786
0.9	0.067987	2.290368	0.022263	2.253786
0.95	0.064409	2.290306	0.021091	2.253787
1.00	0.061188	2.29025	0.020036	2.253788

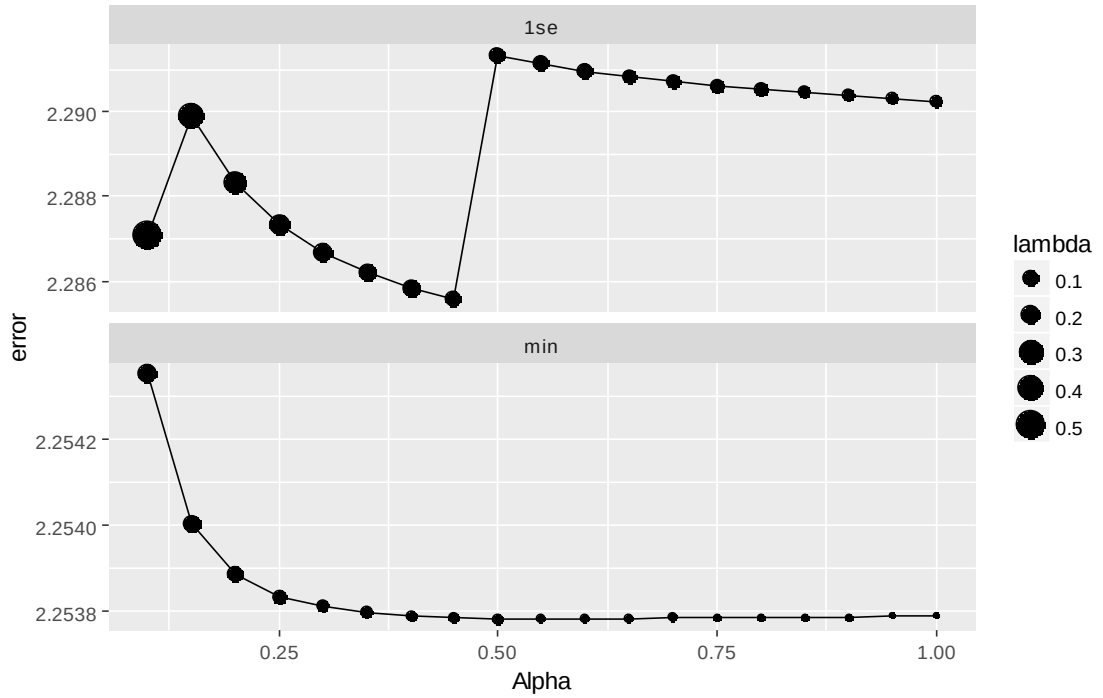

Figure 4.7: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error rule (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error for the mouse 2kbp promoter region.

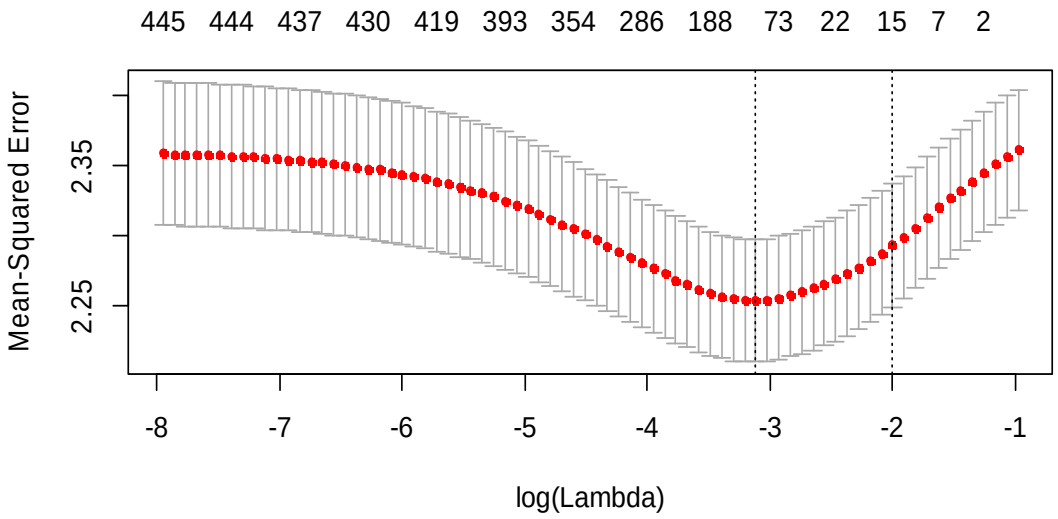

Figure 4.8: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that is within one standard error of the minimum for the mouse 2kbp promoter region.

Table 4.11 shows the estimated model coefficients using elastic-net. We found that B1, B2, B3, ID_B1, PB1D10, and PB1D9 elements which belong to the SINE class, are associated with higher average expression. In contrast, (CA)n, (GA)n, and (TG)n elements, which belong to the simple_repeat family, and L1_Mur1 which belongs to the L1 family, are associated with lower average expression. Figure 4.9 shows the coefficients profile plot of the mouse 2kbp promoter model.

Table 4.11: Results of fitting an elastic-net regression model to the average gene expression in the mouse 2kbp promoter region.

RepName	Coefficients	RepName	Coefficients
(CA)n	-1.2E-03	B1F	4.1E-04
(GA)n	-1.5E-04	B2_Mm1t	2.7E-04
(TG)n	-1.1E-03	B2_Mm2	4.9E-04
B1_Mm	5.6E-04	B3	4.0E-05
B1_Mur3	1.6E-04	ID_B1	1.7E-04
B1_Mur4	4.7E-04	PB1D10	1.0E-05
B1_Mus1	1.3E-03	PB1D9	2.1E-03
B1_Mus2	1.8E-03	L1_Mur1	-4.0E-05

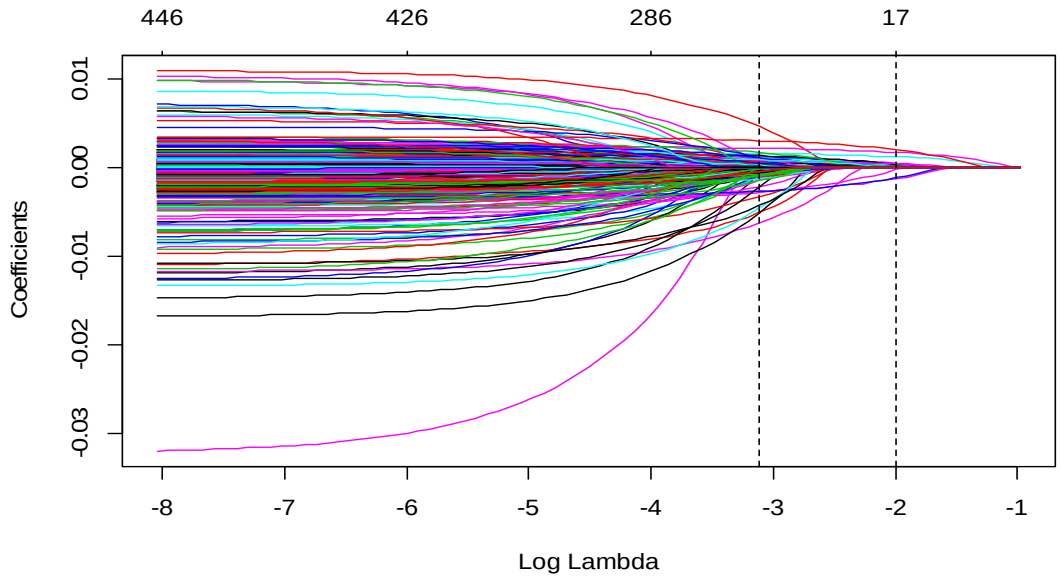

Figure 4.9: Coefficients profile plot of the fitted elastic net for the mouse 2kbp promoter with α = 0.45. Each colored line represents the coefficient value at different values of λ.

4.5.2.2 Penalized regression model results for mouse in the 20kbp promoter region

Table 4.12 and Figure 4.10 show the results of the error using different λ and α values. We found that the optimal α and λ values that minimize error using the one standard error methodology are 0.1 and 0.81917872, respectively. In the plot of α versus error, the lower error is the better results. The upper pane represents the error using the one standard error rule (lambda.1se), and the lower pane represents the error by selecting λ that minimize the error for mouse 20kbp promoter region. Figure 4.11 shows the cross-validation curve for fitted mouse 20kbp model using *glmnet* on the average gene expression data.

Table 4.12: Errors results using different values of λ and α simultaneously in both minimum error and one-standard-error cases in the mouse 20kbp model.

Alpha	*lambda.1se*	*error.1se*	*lambda.min*	*error.min*
0.1	**0.81917872**	**2.202393**	0.18489025	2.160773
0.15	0.59936509	2.204902	0.12326017	2.161681
0.2	0.44952382	2.203086	0.10145841	2.16215
0.25	0.39468147	2.207472	0.08116673	2.162453
0.3	0.32890122	2.206669	0.06763894	2.162662
0.35	0.28191534	2.206113	0.05797624	2.162813
0.4	0.24667592	2.205723	0.05072921	2.162927
0.45	0.21926748	2.205432	0.04509263	2.163017
0.5	0.19734073	2.205196	0.04058336	2.16309
0.55	0.17940067	2.205005	0.03689397	2.163153
0.6	0.16445061	2.20485	0.03381947	2.163205
0.65	0.15180057	2.20472	0.03121797	2.163249
0.7	0.14095767	2.204616	0.02898812	2.163287
0.75	0.13156049	2.204526	0.02705558	2.16332
0.8	0.12333796	2.204441	0.0253646	2.16335
0.85	0.11608279	2.20437	0.02387257	2.163377
0.9	0.10963374	2.204308	0.02254631	2.163401
0.95	0.10386354	2.204254	0.02135967	2.163422
1	0.09867037	2.204202	0.02029168	2.163441

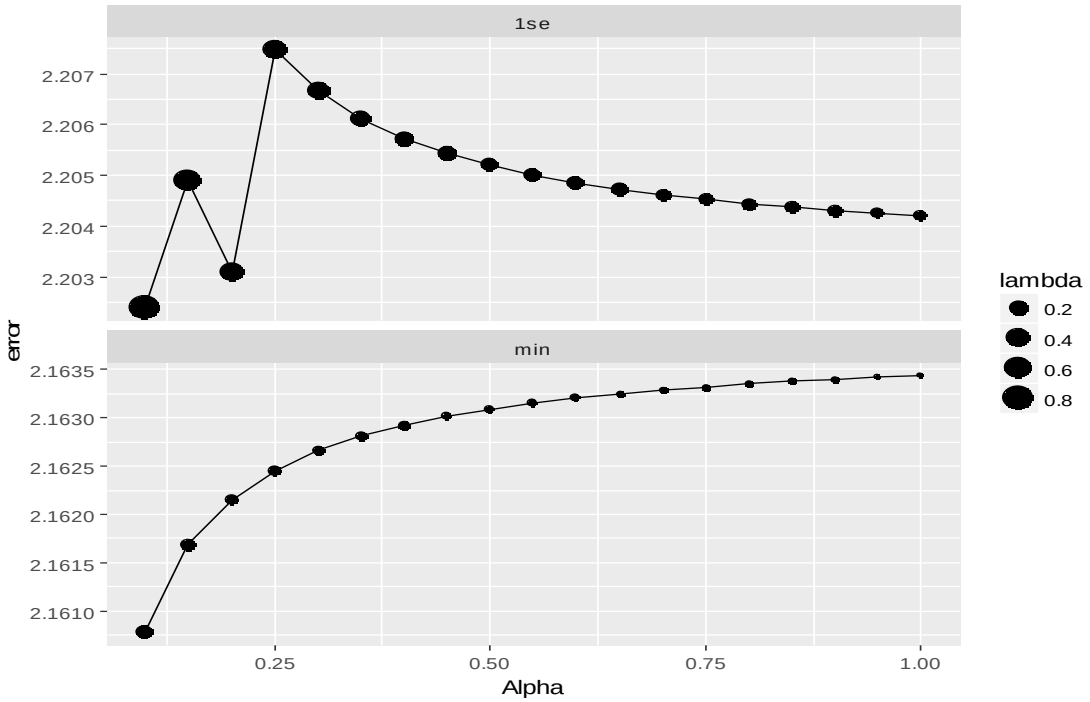

Figure 4.10: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error rule (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error for the mouse 20kbp promoter region.

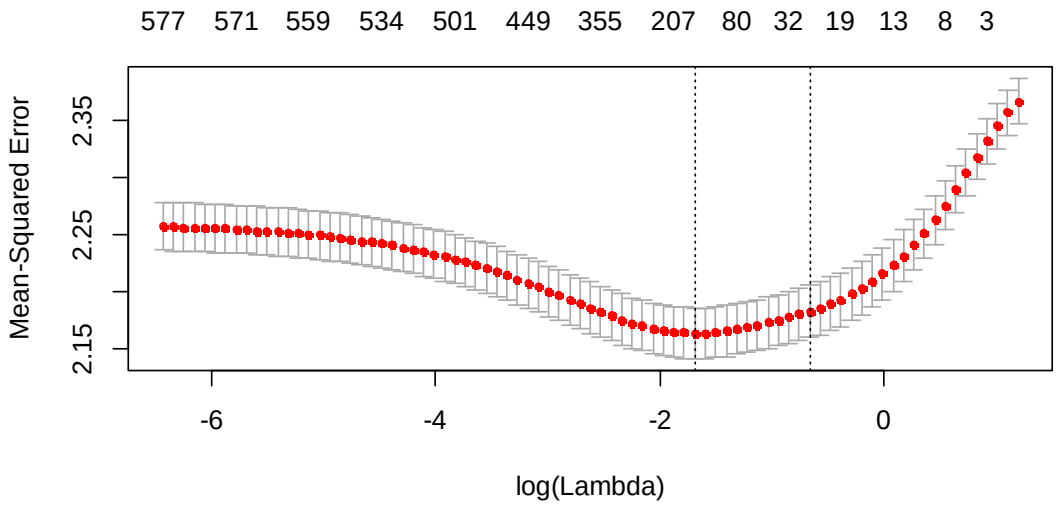

Figure 4.11: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that within one standard error of the minimum for the mouse 20kbp promoter region.

Table 4.13 shows the estimated model coefficients. Our results revealed that the most significantly associated elements are B1_Mm, B1_Mur1, B1_Mur2, B1_Mur3, B1_Mur4, B1_Mus1, and B1_Mus2. In contrast (CA)n, (TC)n, and (TCTA)n elements, which belong to the simple_repeat family, are associated with downregulation of the gene expression. Furthermore, L1_Mus1, L1_Mus3, and L1Md_F2, which belong to the L1 family are also related to gene downregulation. Figure 4.12 shows the coefficients profile plot of the mouse 2kbp promoter model.

Table 4.13: Results of fitting elastic-net regression model to the average gene expression in the mouse 20kbp promoter region

RepName	Coefficients	RepName	Coefficients
(CA)n	-1.0E-05	B2_Mm1t	5.0E-05
(TC)n	-1.1E-05	B2_Mm2	1.7E-04
(TCTA)n	-1.7E-04	B3	7.9E-05
B1_Mm	2.9E-04	ID_B1	9.3E-05
B1_Mur1	2.1E-04	ID4_	3.2E-04
B1_Mur2	1.9E-04	L1_Mus1	-8.0E-06
B1_Mur3	9.3E-05	L1_Mus3	-5.5E-05
B1_Mur4	2.0E-04	L1Md_F2	-1.5E-05
B1_Mus1	3.5E-04	PB1D10	1.3E-04
B1_Mus2	4.4E-04	PB1D9	6.5E-04
B1F	1.4E-04	RMER5	2.0E-05
B2_Mm1a	1.1E-04	RSINE1	1.7E-05

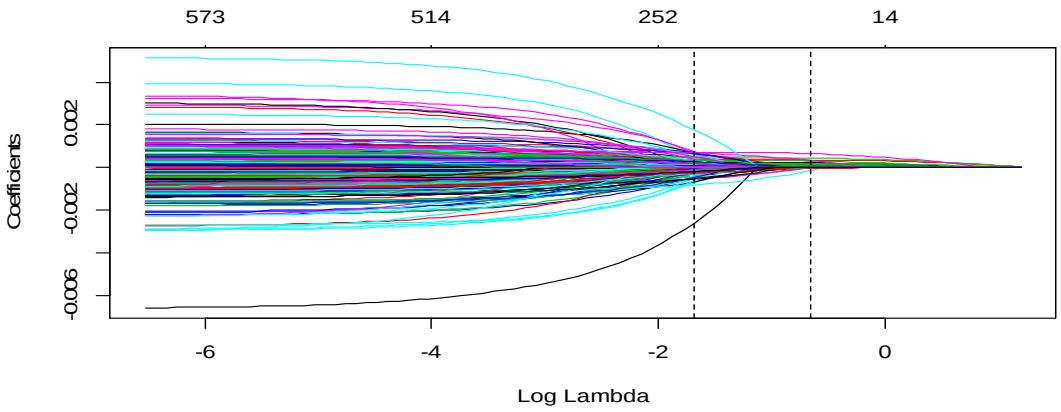

Figure 4.12: Coefficients profile plot of the mouse 20kbp promoter model fitted in with $\alpha = 0.1$. Each colored line represents the coefficient value at different values of λ.

4.5.2.3 Penalized regression model results for human in the 2kbp promoter region

Table 4.14 and Figure 4.13 show the results for various λ and α. We found that the optimal α and λ values that minimize error using the one standard error methodology are 1 (LASSO) and 0.105993, respectively. Figure 4.14 shows the cross-validation curve for fitted human 2kbp model using *glmnet* on the average gene expression data.

Table 4.14: Errors results using different values of λ and α simultaneously in both minimum error and one-standard-error cases in the human 2kbp model.

Alpha	***lambda.1se***	***error.1se***	***ambda.min***	***error.min***
0.1	1.0599303	1.798196	0.21797597	1.762142
0.15	0.7066202	1.798142	0.14531731	1.761643
0.2	0.5299651	1.798108	0.10898798	1.761449
0.25	0.4239721	1.798084	0.08719039	1.761354
0.3	0.3533101	1.798066	0.07265866	1.761303
0.35	0.3028372	1.798052	0.06227885	1.761272
0.4	0.2649826	1.798042	0.05449399	1.761252
0.45	0.2355401	1.798033	0.0484391	1.761239
0.5	0.2119861	1.798026	0.04359519	1.761231
0.55	0.1927146	1.79802	0.03963199	1.761224
0.6	0.176655	1.798015	0.03632933	1.76122
0.65	0.1630662	1.79801	0.03353476	1.761216
0.7	0.1514186	1.798006	0.03113942	1.761213
0.75	0.141324	1.798003	0.02906346	1.761211
0.8	0.1324913	1.798	0.027247	1.76121
0.85	0.1246977	1.797997	0.02564423	1.761209
0.9	0.11777	1.797995	0.02421955	1.761208
0.95	0.1115716	1.797993	0.02294484	1.761207
1	**0.105993**	**1.797991**	**0.0217976**	**1.761207**

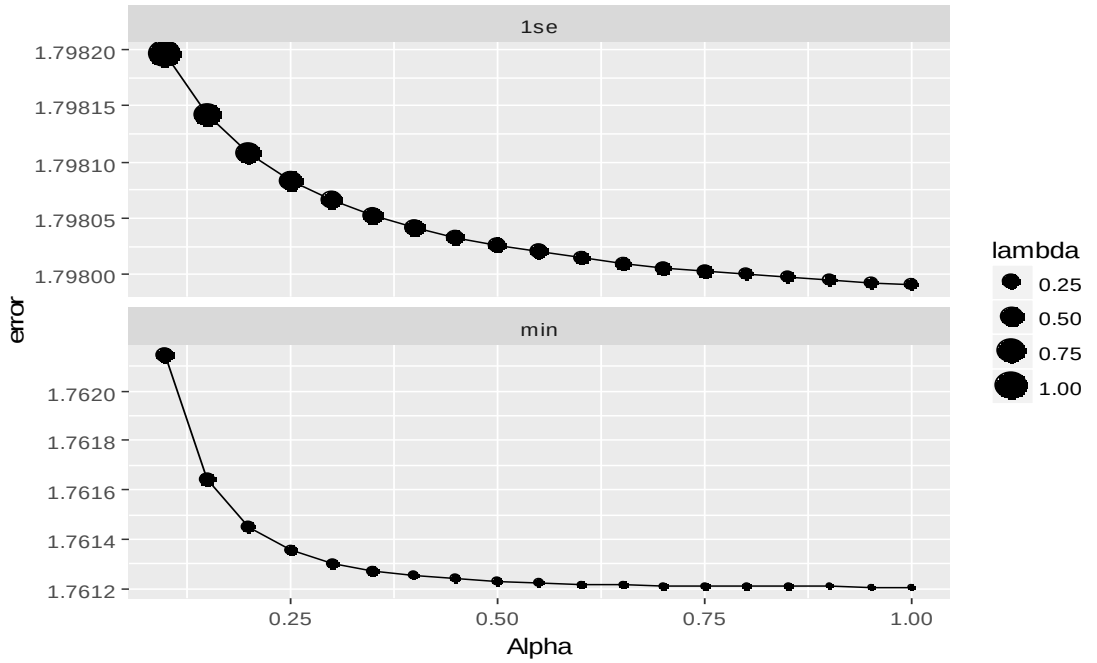

Figure 4.13: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error.

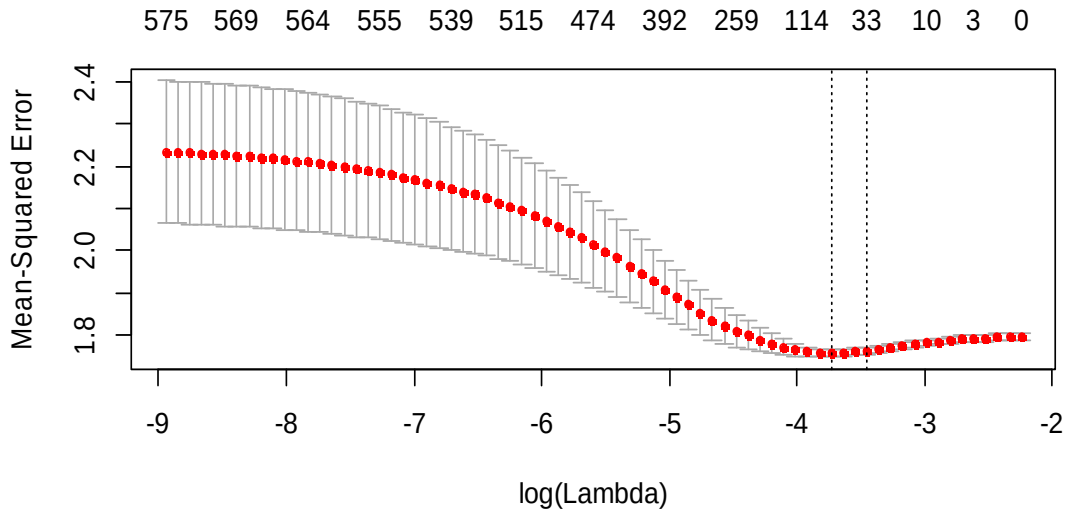

Figure 4.14: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that within one standard error of the minimum.

Table 4.15 shows the estimated model coefficients. Our results revealed that the most significantly associated elements are AluSc, AluSg, AluSg7, AluSp, AluSx, AluSz, and AluY which belong to the Alu family. In contrast, (CA)n, (CACG)n, (TA)n, (TCCC)n**,** and (TG)n elements, which belong to the simple_repeat family, are associated with down-regulation of the gene expression.

Table 4.15: Results of fitting an elastic-net regression model to the average gene expression in the human 2kbp promoter region.

RepName	**Coefficients**	**RepName**	**Coefficients**
(CA)n	-8.1E-04	L1M4c	-3.0E-05
(CACG)n	-7.9E-04	L1MEe	-2.2E-04
(TA)n	-6.5E-04	L1MEf	2.8E-03
(TAGA)n	7.0E-05	L1PA15	-8.5E-04
(TCCC)n	-1.7E-03	L1PA4	-2.2E-04
(TG)n	-5.8E-04	L3b	-7.1E-04
AluSc	2.8E-04	LTR12D	-2.3E-04
AluSg	2.1E-04	LTR42	-2.2E-04
AluSg7	8.2E-04	LTR45C	-1.4E-03
AluSp	5.1E-04	LTR73	2.1E-03
AluSx	9.0E-05	MER112	1.6E-04
AluSz	1.2E-04	MIRb	-1.1E-04
AluY	2.4E-04	MLT1E3	5.7E-03
FLAM_A	2.5E-03	MLT1K	-4.4E-04
FLAM_C	3.4E-04	MSTB1	-1.0E-05
G-rich	-6.0E-05	Ricksha_c	1.3E-03
GA-rich	-1.3E-03		

4.5.2.4 Penalized regression model results for human in the 20kbp promoter region

Table 4.16 and Figure 4.15 show the results for various λ and α values for human tissues in the 20kbp promoter region. We found that the optimal α and λ values to minimize error using the one standard error methodology are 0.4 and 0.1006811, respectively. Figure 4.16 shows the cross-validation curve for fitted human 20kbp model using *glmnet* on the average gene expression data.

Table 4.16: Errors results using different values of λ and α simultaneously in both minimum error and one-standard-error cases in the human 20kbp model.

Alpha	*lambda.1se*	*error.1se*	*ambda.min*	*error.min*
0.1	0.40272439	1.726216	0.14473164	1.688878
0.15	0.26848293	1.724605	0.09648776	1.6893
0.2	0.2013622	1.723816	0.07236582	1.68958
0.25	0.16108976	1.723334	0.05789266	1.689771
0.3	0.13424146	1.723014	0.04824388	1.68991
0.35	0.11506411	1.722789	0.0413519	1.690014
0.4	**0.1006811**	**1.722621**	0.03618291	1.690095
0.45	0.09821989	1.729162	0.03216259	1.69016
0.5	0.08839791	1.729053	0.02894633	1.690214
0.55	0.08036173	1.728963	0.02631484	1.690258
0.6	0.07366492	1.728888	0.02412194	1.690296
0.65	0.06799839	1.728825	0.02226641	1.690328
0.7	0.06314136	1.72877	0.02067595	1.690356
0.75	0.05893194	1.728723	0.01929755	1.69038
0.8	0.05524869	1.72868	0.01809146	1.690402
0.85	0.05199877	1.728643	0.01702725	1.690421
0.9	0.04910995	1.72861	0.01608129	1.690438
0.95	0.04652521	1.728581	0.01523491	1.690453
1	0.04419895	1.728554	0.01447316	1.690467

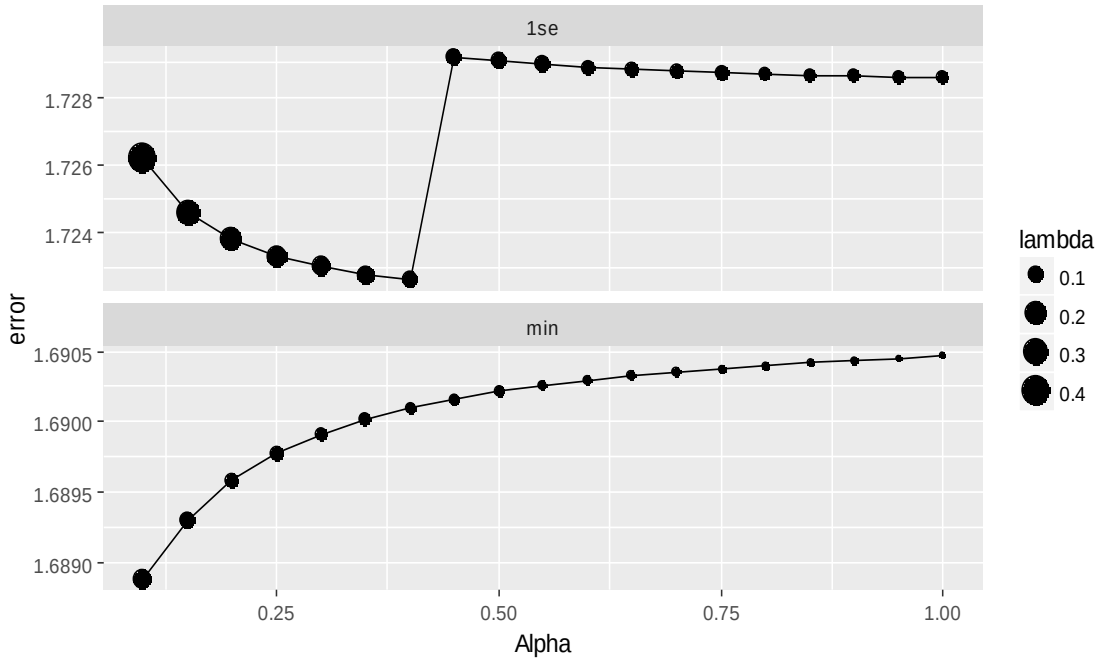

Figure 4.15: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error.

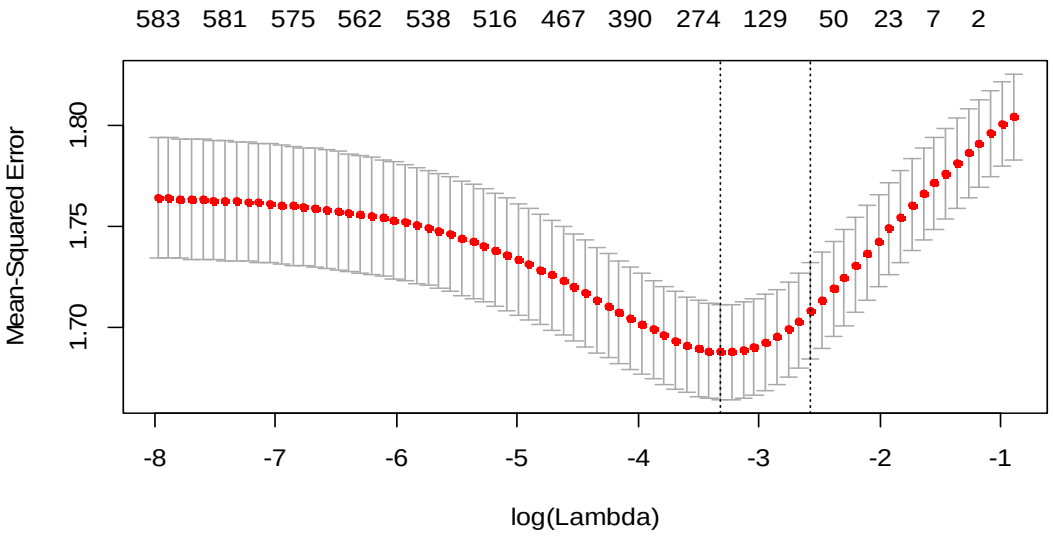

Figure 4.16: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that within one standard error of the minimum.

Table 4.17 shows the estimated model coefficients in the human 20kbp promoter region. Our results revealed that the repeats belonging to the Alu and DNA families are significantly associated with higher gene expressions. In contrast, repeats belonging to the L1, Low_complexity, and simple_repeat families are significantly associated with the decreasing gene expressions.

Table 4.17: Results of fitting an elastic-net regression model to the average gene expression in the human 20kbp promoter region.

RepName	**Coefficients**	**RepName**	**Coefficients**	**RepName**	**Coefficients**
(ATG)n	-7.7E-04	AluY	8.0E-06	LTR16C	-9.0E-05
(CA)n	-2.9E-04	AT-rich	7.7E-04	LTR32	1.6E-04
(CAGAGA)n	-3.2E-04	C-rich	-2.3E-04	LTR33B	-9.0E-05
(CTG)n	-1.6E-03	Charlie4z	2.9E-04	LTR5A	-2.3E-04
(GA)n	-1.2E-03	Charlie5	4.3E-05	MamRep137	5.5E-04
(TCCC)n	-3.4E-04	Charlie9	1.5E-04	MER105	3.1E-04
(TG)n	-3.2E-04	CT-rich	-2.9E-04	MER1B	4.0E-04
(TGAA)n	1.3E-03	FAM	1.1E-03	MER47B	8.9E-04
(TTAAA)n	-2.2E-04	FLAM_A	1.2E-04	MER5C	1.3E-04
(TTCA)n	-7.8E-04	FLAM_C	1.7E-04	MER66B	3.0E-05
(TTCC)n	-9.0E-05	G-rich	-3.3E-04	MIR3	-1.2E-04
(TTTC)n	-1.4E-03	GA-rich	-1.7E-04	MIRb	-1.9E-04
AluJb	1.3E-04	GC-rich	-4.6E-04	MIRc	-1.5E-04
AluJr4	1.8E-04	HERV16int	-1.9E-05	MLT1E3	7.1E-04
AluSc	1.8E-04	L1HS	-2.5E-05	MLT1F	1.7E-04
AluSg	1.6E-04	L1M1	-4.0E-05	MLT1F1	2.2E-04
AluSg7	1.5E-04	L1M6	-1.9E-05	MSTC	1.9E-04
AluSp	1.6E-04	L1MA4A	5.0E-06	SVA_D	9.0E-05
AluSq2	9.0E-05	L1MA8	-2.0E-06	Tigger3	2.9E-04
AluSx	1.0E-08	L1MC3	-2.4E-05	Tigger4a	5.4E-04
AluSx3	1.3E-04	L1MEe	-6.2E-05	Tigger4b	1.8E-04
AluSz	4.0E-05	L2	1.3E-05	U2	7.9E-04
AluSz6	9.0E-05	L2c	8.0E-06	X7B_LINE	7.6E-04

4.5.3 Multivariate Linear Regression Results

First, we fitted the univariate regression model for each tissue of both human and mouse in the 2kbp and 20kbp to determine the residuals for each model. Then we calculated the correlation between all model residuals in each case to examine the degree of correlation between them. Figure 4.17 shows the correlation matrix of residuals in the mouse 2kbp case. We found that the residuals of response variables are correlated. In order to test whether the repeat coefficients are significantly different across all response variables, we fitted MMLR models for each case. The results are shown in the next section.

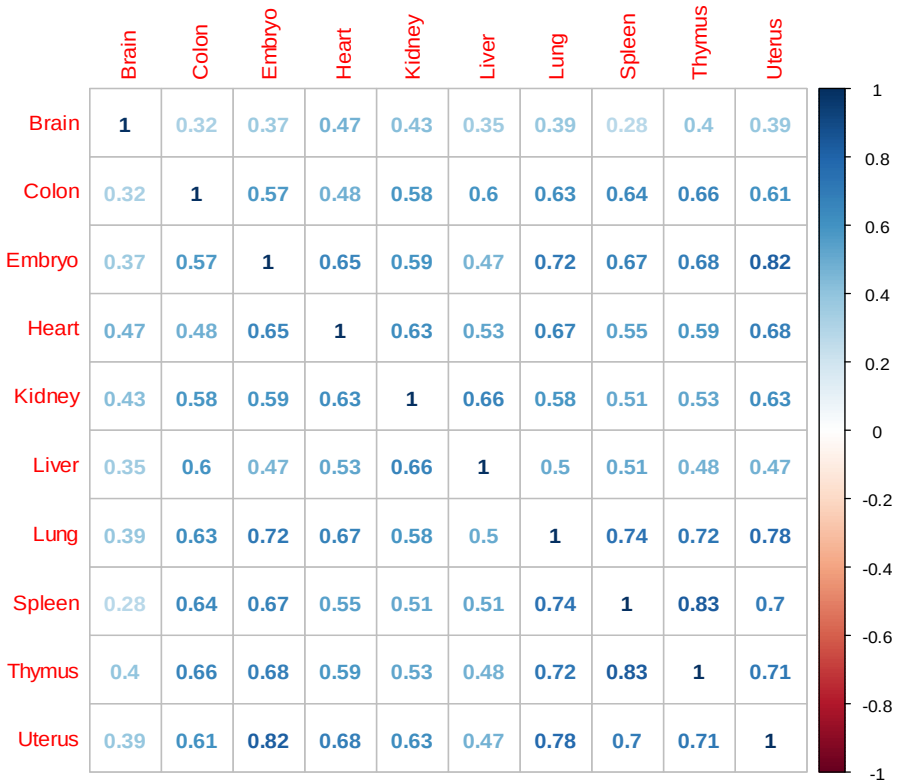

Figure 4.17: Correlation plot of residuals in mouse 2kbp models.

4.5.3.1 Mouse MMLR model results for the tissue expression in the 2kbp promoter region

Table 4.18 shows the significant estimated regression coefficients. Our results demonstrated that repeat coefficients are still significant across all response variables (Table 4.19).

Table 4.18: Results of fitting an MMLR model for all tissues expression in the mouse 2kbp promoter region.

	Brain	**Embryo**	**Heart**	**Kidney**	**Liver**
Alu	5.0E-04	2.1E-03	1.7E-03	1.8E-03	1.9E-03
B2	3.0E-04	7.0E-04	5.0E-04	6.0E-04	8.0E-04
ERVK	-5.0E-04	-6.0E-04	-6.0E-04	-1.0E-04	1.0E-04
ERVL-MaLR	-2.0E-04	-5.0E-04	-4.0E-04	-2.0E-04	1.0E-04
L1	-4.0E-04	-8.0E-04	-7.0E-04	-6.0E-04	-2.0E-04
Low_complexity	-1.1E-03	-4.0E-04	2.0E-04	-2.0E-04	-4.0E-04
MIR	-8.0E-04	-4.0E-04	-5.0E-04	-8.0E-04	-7.0E-04
Simple_repeat	-1.3E-03	-3.2E-03	-2.4E-03	-2.3E-03	-2.4E-03
	Lung	**Spleen**	**Colon**	**Uterus**	**Thymus**
Alu	1.4E-03	2.5E-03	1.7E-03	1.8E-03	2.6E-03
B2	6.0E-04	1.1E-03	7.0E-04	5.0E-04	9.0E-04
ERVK	-5.0E-04	-2.0E-04	-2.0E-04	-5.0E-04	-2.0E-04
ERVL-MaLR	-4.0E-04	-5.0E-04	-1.0E-04	-5.0E-04	-5.0E-04
L1	-6.0E-04	-5.0E-04	-1.0E-04	-8.0E-04	-7.0E-04
Low_complexity	-8.0E-04	-1.4E-03	-1.9E-03	-4.0E-04	-1.0E-03
MIR	-1.0E-04	-3.0E-04	-8.0E-04	-3.0E-04	-9.0E-04
Simple_repeat	-1.9E-03	-2.9E-03	-2.0E-03	-2.4E-03	-3.2E-03

Table 4.19: Multivariate test statistics results for mouse 2kbp promoter region.

RepFamily	**Pillai**	**Wilks**	**Roy**	**Hotelling-**
Alu	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
B2	$< 3.8\times10^{-16}$	$< 3.8\times10^{-16}$	$< 3.8\times10^{-16}$	$< 3.8\times10^{-16}$
ERVK	$< 4.1\times10^{-11}$	$< 4.1\times10^{-11}$	$< 4.1\times10^{-11}$	$< 4.1\times10^{-11}$
ERVL-MaLR	$< 2.4\times10^{-5}$	$< 2.4\times10^{-5}$	$< 2.4\times10^{-5}$	$< 2.4\times10^{-5}$
L1	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Low_complexity	$< 6.6\times10^{-9}$	$< 6.6\times10^{-9}$	$< 6.6\times10^{-9}$	$< 6.6\times10^{-9}$
MIR	$< 5.1\times10^{-6}$	$< 5.1\times10^{-6}$	$< 5.1\times10^{-6}$	$< 5.1\times10^{-6}$
Simple_repeat	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$

4.5.3.2 Mouse MMLR model results for the tissue expression in the 20kbp promoter region

Table 4.20 shows the significant estimated regression coefficients. Our results demonstrated that repeat coefficients are still significant across all response variables. Table 4.21 shows the multivariate test statistics which verify this result, with the exception of ID elements.

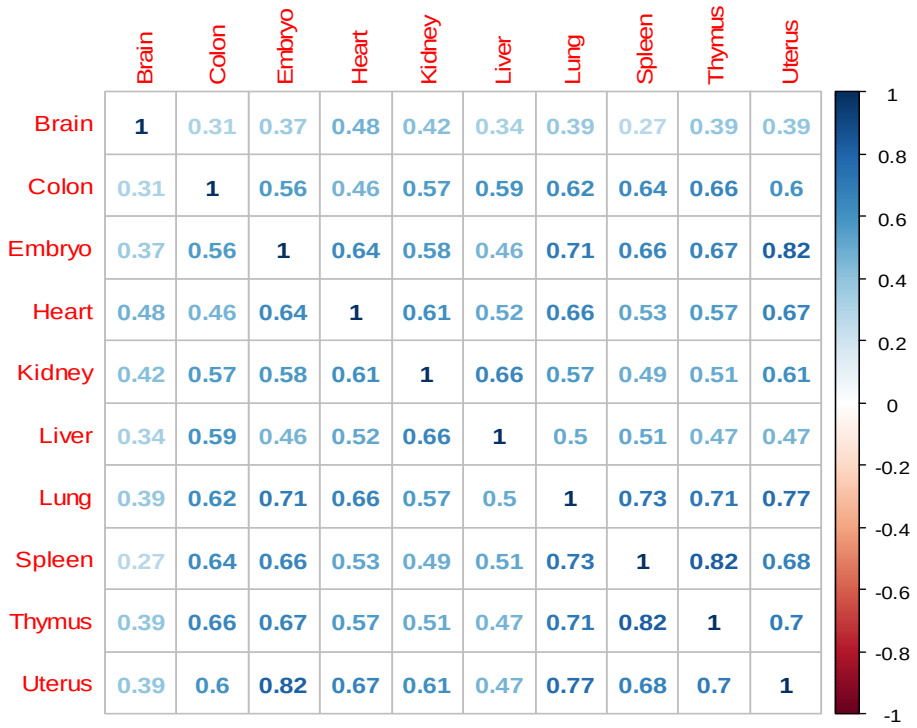

Figure 4.18: Correlation plot of residuals in mouse 20kbp models.

Table 4.20: Results of fitting an MMLR model for all tissues expression in the mouse 20kbp promoter region.

	Brain	**Embryo**	**Heart**	**Kidney**	**Liver**
Alu	1.7E-04	4.3E-04	3.3E-04	3.5E-04	3.5E-04
B2	3.3E-05	1.0E-04	1.2E-04	1.3E-04	1.4E-04
ERV1	2.5E-05	5.2E-05	4.9E-05	9.8E-05	1.3E-04
ERVK	-1.9E-05	-2.2E-05	-1.2E-05	1.9E-05	4.8E-05
ERVL-MaLR	1.2E-04	-2.0E-06	4.5E-05	3.3E-05	5.1E-05
ID	3.9E-04	4.3E-04	5.4E-04	8.2E-04	7.7E-04
L1	2.9E-05	-1.0E-04	-8.1E-05	-5.6E-05	2.0E-05
Low_complexity	3.5E-04	-4.0E-04	-3.1E-04	-2.5E-04	-2.4E-04
MIR	-1.6E-04	-1.9E-04	-1.8E-04	-4.1E-04	-6.1E-04
TcMar.Tigger	3.6E-04	2.9E-04	3.6E-04	3.6E-04	2.8E-04
Simple_repeat	-3.1E-04	-4.5E-04	-4.5E-04	-3.9E-04	-5.3E-04
tRNA	2.2E-04	2.9E-03	1.4E-03	4.6E-04	4.4E-04
	Lung	**Spleen**	**Colon**	**Uterus**	**Thymus**
Alu	2.7E-04	4.9E-04	3.3E-04	3.6E-04	4.9E-04
B2	1.2E-04	1.2E-04	1.0E-04	1.0E-04	1.3E-04
ERV1	5.1E-05	1.1E-04	1.1E-04	6.0E-05	9.0E-05
ERVK	-1.2E-05	2.0E-05	3.7E-05	-2.4E-05	4.0E-06
ERVL-MaLR	-5.0E-06	-6.3E-05	5.0E-06	1.0E-06	-6.0E-06
ID	5.1E-04	3.3E-04	5.1E-04	2.0E-04	3.7E-04
L1	-8.0E-05	-1.2E-04	-4.0E-05	-1.1E-04	-1.0E-04
Low_complexity	-3.4E-04	-6.6E-04	-3.4E-04	-4.1E-04	-4.7E-04
MIR	-1.5E-05	-1.9E-04	-4.8E-04	-1.5E-04	-4.3E-04
TcMar.Tigger	3.3E-04	1.1E-04	1.4E-04	3.0E-04	3.7E-04
Simple_repeat	-1.8E-04	-4.0E-04	-3.7E-04	-2.7E-04	-5.8E-04
tRNA	2.0E-03	3.8E-03	1.8E-03	3.2E-03	4.6E-03

Table 4.21: Multivariate test statistics results for mouse 20kbp promoter region.

RepFamily	**Pillai**	**Wilks**	**Roy**	**Hotelling-**
Alu	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
B2	0.00166	0.00166	0.00166	0.00166
ERV1	$< 2.2\times10^{-5}$	$< 2.2\times10^{-5}$	$< 2.2\times10^{-5}$	$< 2.2\times10^{-5}$
ERVK	$< 6.1\times10^{-9}$	$< 6.1\times10^{-9}$	$< 6.1\times10^{-9}$	$< 6.1\times10^{-9}$
ERVL-MaLR	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
L1	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Low_complexity	$< 4.9\times10^{-10}$	$< 4.9\times10^{-10}$	$< 4.9\times10^{-10}$	$< 4.9\times10^{-10}$
MIR	$< 6.6\times10^{-15}$	$< 6.6\times10^{-15}$	$< 6.6\times10^{-15}$	$< 6.6\times10^{-15}$
Simple_repeat	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
TcMar.Tigger	0.00740	0.00740	0.00740	0.00740
tRNA	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$

4.5.3.3 Human MMLR model results for the tissue expression in the 2kbp promoter region

Table 4.22 indicates the significant estimated regression coefficients. Our results demonstrated that repeat coefficients are still significant across all response variables. Table 4.23 shows the multivariate test statistics which verify this results, with the exception of ID elements.

Figure 4.19: Correlation plot of residuals in human 2kbp models.

Table 4.22: Results of fitting an MMLR model for all tissues expression in the human 2kbp promoter region.

	Brain	**Heart**	**Kidney**	**Liver**	**Lung**
Alu	3.0E-04	2.0E-04	8.0E-04	4.0E-04	2.0E-04
ERVL-MaLR	-1.0E-03	-6.0E-04	-8.0E-04	-2.0E-04	-4.0E-04
L1	-6.0E-04	-4.0E-04	-5.0E-04	-1.0E-04	-3.0E-04
Low_complexity	1.8E-03	-9.0E-04	-2.0E-04	-1.5E-03	-1.2E-03
MIR	-3.0E-04	-5.0E-04	-1.0E-04	-3.0E-04	-3.0E-04
Simple_repeat	-5.0E-04	-1.3E-03	-8.0E-04	-1.4E-03	-1.1E-03
tRNA	6.7E-03	9.2E-03	6.9E-03	3.1E-03	6.2E-03
	Prostate	**Spleen**	**Colon**	**Uterus**	**Thymus**
Alu	1.0E-04	3.0E-04	4.0E-04	1.0E-08	4.0E-04
ERVL-MaLR	-2.0E-04	-4.0E-04	-2.0E-04	-3.0E-04	-7.0E-04
L1	-3.0E-04	-2.0E-04	-1.0E-04	-2.0E-04	-4.0E-04
Low_complexity	-9.0E-04	-1.2E-03	-1.1E-03	-1.1E-03	-1.0E-03
MIR	-9.0E-04	-4.0E-04	-8.0E-04	-7.0E-04	-7.0E-04
Simple_repeat	-1.2E-03	-1.2E-03	-1.2E-03	-1.3E-03	-1.2E-03
tRNA	4.7E-03	4.1E-03	7.0E-03	5.5E-03	8.1E-03

Table 4.23: Multivariate test statistics results for the human 2kbp promoter region.

RepFamily	**Pillai**	**Wilks**	**Roy**	**Hotelling-Lawley**
Alu	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
ERVL-MaLR	$< 3.3\times10^{-10}$	$< 3.3\times10^{-10}$	$< 3.3\times10^{-10}$	$< 3.3\times10^{-10}$
L1	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Low_complexity	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
MIR	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Simple_repeat	$< 9.2\times10^{-16}$	$< 9.2\times10^{-16}$	$< 9.2\times10^{-16}$	$< 9.2\times10^{-16}$
tRNA	0.01240	0.01240	0.01240	0.01240

4.5.3.4 Human MMLR model results for the tissue expression in the 20kbp promoter region

Table 4.24 shows the significantly estimated regression coefficients. Our results demonstrated that repeat coefficients are still significant across all response variables, with the exception of ID elements (Table 4.25).

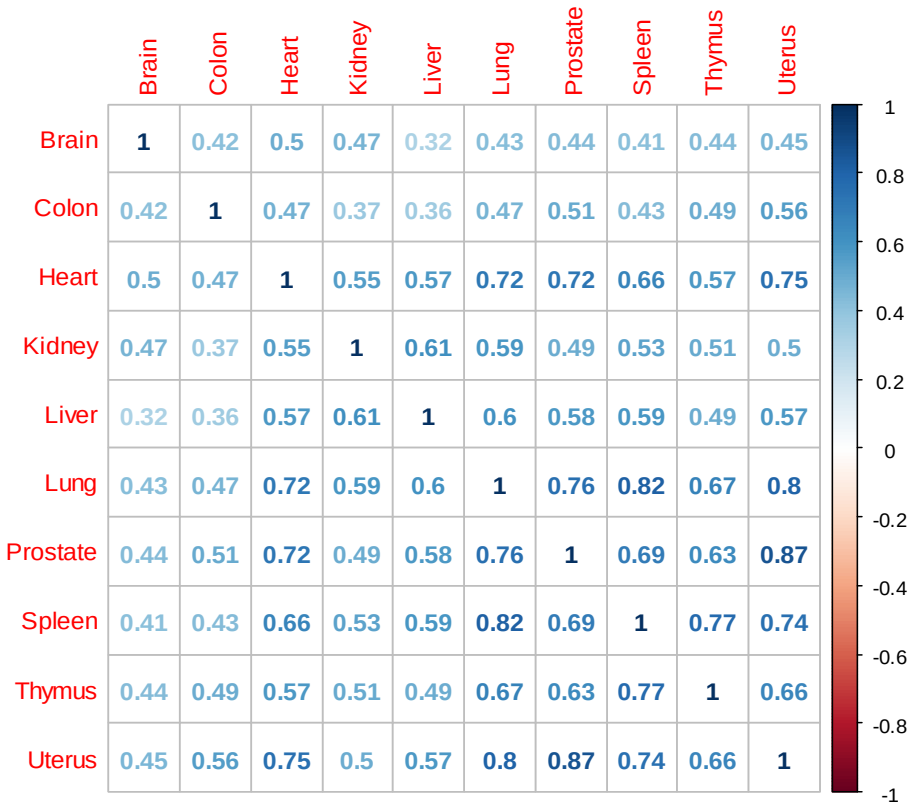

Figure 4.20: Correlation plot of residuals in human 20kbp models.

Table 4.24: Results of fitting an MMLR model for all tissues expression in the human 2kbp promoter region.

	Brain	**Heart**	**Kidney**	**Liver**	**Lung**
Alu	8.8E-05	7.6E-05	1.2E-04	6.9E-05	6.2E-05
ERVL-MaLR	-1.4E-05	7.3E-05	-4.9E-05	2.8E-05	5.8E-05
hAT.Charlie	1.3E-05	3.2E-04	4.7E-05	1.6E-04	2.6E-04
L1	-6.7E-05	-1.8E-05	-7.7E-05	-2.0E-06	-4.0E-06
Low_complexity	3.6E-04	-5.2E-04	-3.5E-04	-6.9E-04	-6.4E-04
MIR	-3.3E-05	-2.0E-04	-1.4E-04	-2.8E-04	-1.6E-04
Simple_repeat	-4.5E-05	-4.7E-04	-2.3E-04	-4.8E-04	-4.8E-04
TcMar	1.8E-05	1.7E-03	7.7E-04	6.3E-04	1.6E-03
	Prostate	**Spleen**	**Colon**	**Uterus**	**Thymus**
Alu	3.9E-05	7.1E-05	6.8E-05	4.1E-05	8.2E-05
ERVL-MaLR	1.1E-04	1.8E-05	2.1E-05	1.4E-04	-4.3E-05
hAT.Charlie	3.9E-04	2.6E-04	1.9E-04	4.2E-04	2.2E-04
L1	1.3E-05	-1.3E-05	-2.3E-05	-1.2E-05	-3.9E-05
Low_complexity	-3.4E-04	-5.6E-04	-3.9E-04	-5.9E-04	-3.9E-04
MIR	-3.0E-04	-2.9E-04	-2.8E-04	-2.9E-04	-3.7E-04
Simple_repeat	-5.3E-04	-4.1E-04	-4.2E-04	-5.5E-04	-4.1E-04
TcMar	1.9E-03	1.4E-03	6.2E-04	2.0E-03	7.3E-04

Table 4.25: Multivariate test statistics results for the human 20kbp promoter region.

RepFamily	**Pillai**	**Wilks**	**Roy**	**Hotelling-Lawley**
Alu	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
ERVL-MaLR	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
hAT.Charlie	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
L1	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Low_complexity	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
MIR	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
Simple_repeat	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
TcMar	0.00016	0.00016	0.00016	0.00016

4.5.4 Human BodyMap results

4.5.4.1 Human BodyMap regression model results in the 2kbp promoter region

Table 4.26 shows the of estimated regression coefficients with their corresponding p-values, *Bonferroni* correction p-values, and VIF values of the selected repeat families on the average gene expressions for the Human BodyMap dataset. Similar to our other human tissues results, we found a significant positive association between Alu elements and higher average gene expression. In contrast, ERVL, ERVL-MaLR, L1, and simple repeats results showed a significant negative association. Model assumptions were checked visually using the residual plot. Moreover, formal tests, including VIF, *Levene's* test (p=0.723), and *Durbin-Watson* demonstrated no violations.

Table 4.26: Results of fitting a multiple linear regression model to the average gene expression in the human BodyMap 2kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	3.0E-04	0.0859	2.2×10^{-16}	2.4×10^{-15}	1.0522
ERVL	-6.0E-04	-0.0444	1.8×10^{-8}	2.2×10^{-7}	1.0059
ERVL-MaLR	-3.0E-04	-0.0566	8.3×10^{-13}	9.9×10^{-12}	1.0110
Gypsy	-7.0E-04	-0.0172	2.9×10^{-2}	3.5×10^{-1}	1.0018
hAT.Blackjack	-1.3E-03	-0.0237	2.7×10^{-3}	3.3×10^{-2}	1.0007
hAT.Charlie	2.0E-04	0.0223	1.2×10^{-2}	2.4×10^{-15}	1.0235
L1	-4.0E-04	-0.0885	2.2×10^{-16}	1.5×10^{-2}	1.0349
L2	1.0E-04	0.02578	1.2×10^{-3}	2.5×10^{-1}	1.0226
Low_complexity	-2.0E-04	-0.0142	8.0×10^{-2}	4.7×10^{-6}	1.0686
MIR	-1.0E-04	-0.0186	2.1×10^{-2}	2.6×10^{-1}	1.0475
Simple_repeat	-1.0E-03	-0.0583	2.6×10^{-13}	3.1×10^{-12}	1.0249
tRNA	3.7E-03	0.0194	1.4×10^{-2}	1.6×10^{-1}	1.0017

*Unadjusted p-value cut-off 0.004 †Adjusted p-values (*Bonferroni*) cut-off 0.05

P values less than 0.0045 were deemed significant, after the *Bonferroni* adjustments.

4.5.4.2 Human BodyMap regression model results in 20kbp promoter region

Table 4.27 shows the of estimated regression coefficients with their corresponding p-values, *Bonferroni* correction p-values, and VIF values of the selected repeat families on the average gene expressions for the Human BodyMap dataset. Similar to our previous human tissue results, we found a significant positive association between Alu elements and higher gene expression. In contrast, L1, MIR, and simple repeats results showed a significant negative association. Model assumptions were checked visually using the residual plot. Moreover, formal tests, including VIF, *Levene's* test, and *Durbin-Watson* demonstrated no violations.

Table 4.27: Results of fitting a multiple linear regression model to the average gene expression in the human 20kbp promoter region.

RepFamily	Coefficients	Std. Coefficients	P-Value	Bonferroni	VIF
Alu	7.0E-05	0.16411	2.2×10^{-16}	1.8×10^{-15}	1.1717
ERV1	-2.0E-05	-0.02747	5.2×10^{-4}	4.7×10^{-3}	1.0554
hAT.Charlie	1.0E-04	0.03215	3.7×10^{-5}	3.3×10^{-4}	1.0224
L1	-4.0E-05	-0.09736	2.2×10^{-16}	1.8×10^{-15}	1.2072
Low_complexity	-1.9E-04	-0.02636	9.8×10^{-4}	8.8×10^{-3}	1.0767
MIR	-1.1E-04	-0.04784	9.1×10^{-9}	8.2×10^{-8}	1.1673
scRNA	2.0E-03	0.02589	8.0×10^{-4}	7.2×10^{-3}	1.0056
Simple_repeat	-3.2E-04	-0.06871	2.2×10^{-16}	1.8×10^{-15}	1.0454
TcMar.Mariner	-1.7E-04	-0.01568	4.2×10^{-2}	3.8×10^{-1}	1.0011

*Unadjusted p-value cut-off 0.005 †Adjusted p-values (*Bonferroni*) cut-off 0.05

P values less than 0.0045 were deemed significant, after the *Bonferroni* adjustments.

4.5.4.3 Human BodyMap regression model results in the 2kbp promoter region

Table 4.28 and Figure 4.21 show the results of the error using various λ and α values. We found that that optimal α and λ values that minimize error using the one standard error methodology are 0.1 and 0.31605442, respectively. Figure 4.22 shows the cross-validation curve for fitted human BodyMap 2kbp model using *glmnet* on the average gene expression data.

Table 4.28: Errors results using different values of λ and α simultaneously in both minimum error and one-standard-error cases in human BodyMap 2kbp model.

Alpha	*lambda.1se*	*error.1se*	*lambda.min*	*error.min*
0.1	**0.31605442**	**2.950505**	0.21784405	2.934151
0.15	0.23124622	2.9548	0.14522937	2.933688
0.2	0.17343467	2.953801	0.10892203	2.933519
0.25	0.13874773	2.953198	0.08713762	2.933442
0.3	0.11562311	2.952796	0.07261468	2.9334
0.35	0.09910552	2.952508	0.06224116	2.933376
0.4	0.08671733	2.952292	0.05446101	2.933361
0.45	0.07708207	2.952124	0.04840979	2.933352
0.5	0.06937387	2.95199	0.04356881	2.933345
0.55	0.06306715	2.951881	0.03960801	2.93334
0.6	0.05781156	2.95179	0.03630734	2.933337
0.65	0.05336451	2.951712	0.03351447	2.933335
0.7	0.04955276	2.951646	0.03112058	2.933333
0.75	0.04624924	2.951589	0.02904587	2.933332
0.8	0.04335867	2.951539	0.02723051	2.933331
0.85	0.04080816	2.951495	0.02562871	2.933331
0.9	0.03854104	2.951455	0.02420489	2.93333
0.95	0.03651256	2.95142	0.02293095	2.93333
1.00	0.03468693	2.951388	0.02178441	2.93333

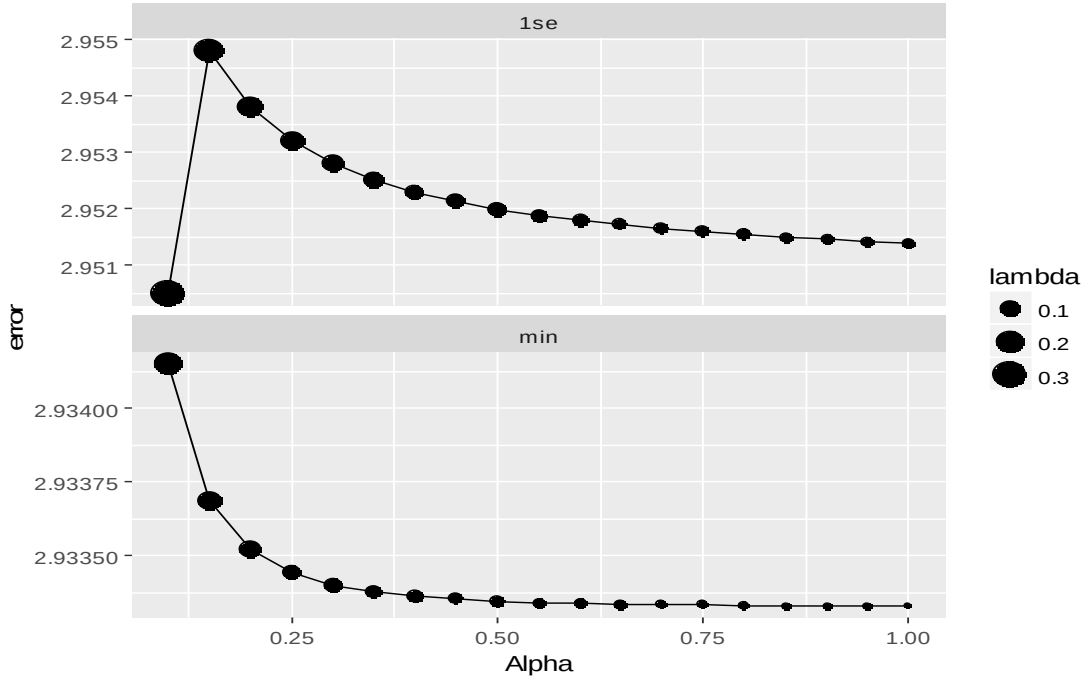

Figure 4.21: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error.

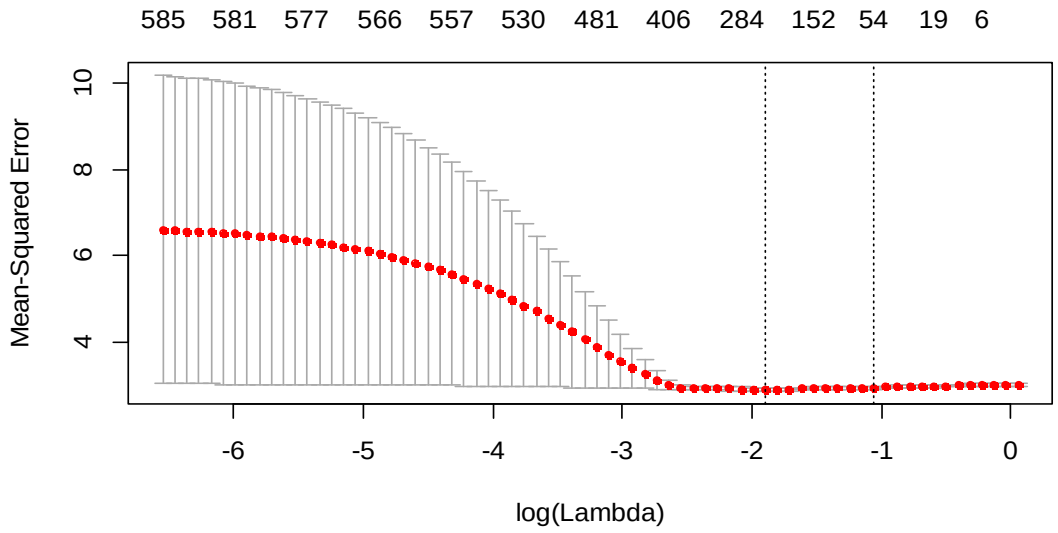

Figure 4.22: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that within one standard error of the minimum.

Table 4.29 shows the estimated model coefficients. Our results demonstrated that repeats belong to the Alu and DNA families are significantly associated with higher gene expressions. In contrast, repeats belong to the L1, Low_complexity, and simple_repeat families are significantly related to gene down-regulation.

Table 4.29: Results of fitting an elastic-net regression model to the average gene expression in the human BodyMap 2kbp promoter region.

RepName	Coefficients	RepName	Coefficients
(CA)n	-1.1E-03	L1MEf	1.8E-03
(GA)n	-4.4E-04	L1PA3	-1.0E-06
(TA)n	-2.8E-04	L1PA7	-6.9E-05
(TC)n	-3.4E-04	L2a	7.0E-05
(TG)n	-6.7E-04	LTR12D	-6.8E-05
(TGGGGG)n	-3.6E-04	LTR15	-2.4E-03
(TTTC)n	-2.1E-03	LTR16C	-3.4E-04
AluSc	2.9E-04	LTR42	-7.6E-04
AluSg	3.0E-04	LTR45C	-1.7E-03
AluSp	7.8E-04	LTR73	4.5E-03
AluSq2	1.3E-04	LTR8A	3.9E-04
AluSx	1.8E-04	MER21A	-2.8E-04
AluSx1	2.1E-04	MER4.int	-7.5E-04
AluSx3	2.8E-04	MER41D	-1.8E-03
AluSz	1.4E-04	MER5B	-5.4E-04
AluY	3.7E-04	MLT1A0	-3.4E-04
AT_rich	-1.3E-04	MLT1D	-7.9E-05
CT.rich	-5.2E-04	MLT1E3	2.8E-03
FLAM_A	2.2E-03	MLT2C1	-5.9E-05
FLAM_C	1.2E-03	MSTA	-2.3E-04
GA.rich	-1.5E-03	T-rich	-1.0E-06
HY4	5.2E-03	THE1C	-4.0E-06
L1M5	-4.8E-04	Tigger4	1.4E-03
L1MA10	3.4E-04	Tigger7	-4.1E-05
L1MB5	-2.9E-04	U2	2.1E-04
L1MDb	-1.6E-03	X7A_LINE	1.0E-03
L1ME2z	8.1E-05	X7B_LINE	2.2E-03

4.5.4.4 Human BodyMap regression model results in the 20kbp promoter region

Table 4.30 and Figure 4.23 show the results of the error using various λ and α values. We found that the optimal α and λ values that minimize error using the one standard error methodology are 0.2 and 0.16119486, respectively. Figure 4.24 shows the cross-validation curve for the fitted human BodyMap 20kbp model using *glmnet* on the average gene expression data.

Table 4.30: Errors results using different values of λ and α simultaneously in both minimum error and one-standard-error cases in human BodyMap 20kbp model.

Alpha	*lambda.1se*	*error.1se*	*lambda.min*	*error.min*
0.1	0.32238973	2.882285	0.13955482	2.850401
0.15	0.21492648	2.881142	0.09303654	2.851016
0.2	**0.16119486**	**2.880603**	0.06977741	2.851383
0.25	0.14152893	2.886505	0.05582193	2.851626
0.3	0.11794078	2.886273	0.04651827	2.851796
0.35	0.10109209	2.886111	0.0398728	2.851922
0.4	0.08845558	2.885992	0.0348887	2.85202
0.45	0.07862718	2.885902	0.03101218	2.852097
0.5	0.07076447	2.885831	0.02791096	2.85216
0.55	0.06433133	2.885774	0.0253736	2.852212
0.6	0.05897039	2.885727	0.02325914	2.852256
0.65	0.0544342	2.885688	0.02146997	2.852293
0.7	0.05054605	2.885654	0.0199364	2.852325
0.75	0.04717631	2.885625	0.01860731	2.852355
0.8	0.04422779	2.8856	0.01744435	2.85238
0.85	0.04162616	2.885578	0.01641821	2.852401
0.9	0.03931359	2.885559	0.01550609	2.852421
0.95	0.03724446	2.885541	0.01468998	2.852439
1	0.03538223	2.885526	0.01395548	2.852455

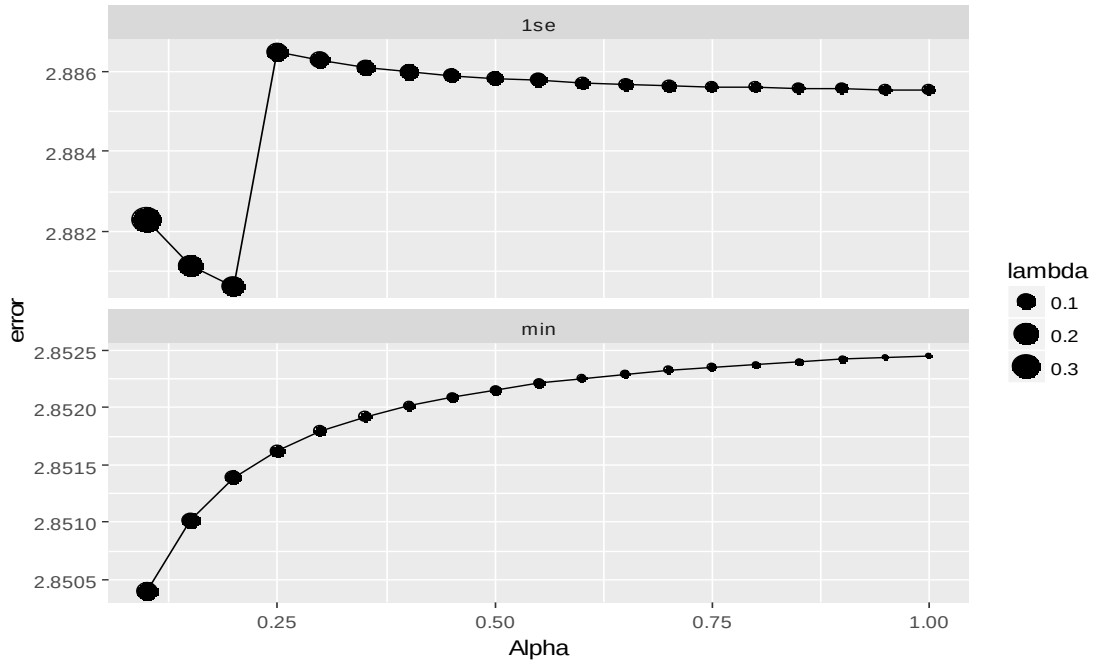

Figure 4.23: Plot of α versus error, the lower error is the better. The upper pane represents the error using the one standard error (lambda.1se) and the lower pane represents the error by selecting λ that minimize the error.

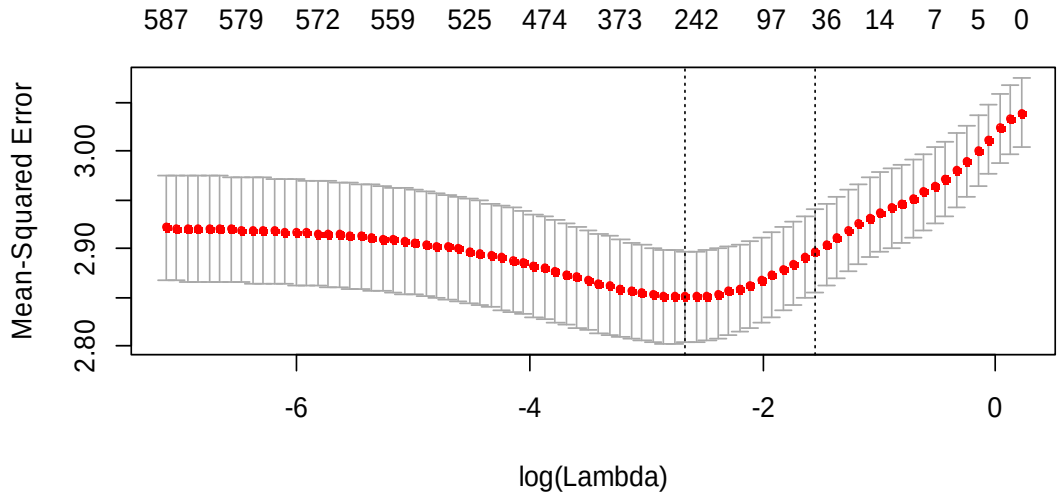

Figure 4.24: Cross-validation curve for the *glmnet* fitted on the gene expression data. The top row of numbers indicates how many variables (repName) are in the model for a given value of log (λ). The red dots represent the cross-validation error at that point and the vertical dot lines are the confidence interval for the error. The leftmost line indicates the value of λ where error is minimized and the rightmost one is the next largest value of λ error that within one standard error of the minimum.

Table 4.31 shows the estimated model coefficients. Our results revealed that repeats belonging to the Alu and DNA families are significantly associated with gene upregulation, while repeats belonging to the LI and simple_repeat families are significantly related to gene downregulation.

Table 4.31: Results of fitting an elastic-net regression model to the average gene expression in the human BodyMap 20kbp promoter region.

RepName	Coefficients	RepName	Coefficients
(TA)n	-4.3E-05	FLAM_C	4.8E-04
(TAGA)n	-5.0E-06	GArich	-2.4E-04
(TG)n	-2.2E-04	L1M1	-8.0E-06
(TTCC)n	-6.1E-05	L1M5	-4.3E-05
(TTTC)n	-1.5E-03	L1M6	-1.1E-04
AluJb	2.1E-04	L1MA2	-1.9E-05
AluJr4	1.2E-04	L1MA8	-7.9E-05
AluSc	1.0E-04	L1MA9	-5.9E-05
AluSc8	4.8E-05	L1MC	-1.8E-04
AluSg	2.3E-04	L1MC3	-7.5E-05
AluSp	2.0E-04	L1PA7	-1.3E-05
AluSq2	5.5E-05	LTR16C	-2.5E-04
AluSx	7.6E-05	MER1B	9.6E-05
AluSx1	4.0E-05	MER41D	-7.0E-04
AluSx3	1.4E-04	MIRb	-2.4E-05
AluSz	3.0E-05	MLT1E3	3.1E-04
AluSz6	8.8E-05	MLT1F1	2.8E-05
AluY	8.9E-05	SVA_D	7.2E-05
Charlie9	2.0E-04	THE1B	-3.2E-05
FAM	4.6E-04	Tigger3	1.9E-04
FLAM_A	9.0E-06	U2	8.5E-05

4.6 Conclusion

Our results from the various regression model approaches showed that Alu and LINE-1 elements, which comprise a significant portion of human and mouse genomes, are significantly associated with gene expression. Alu elements in both human and mouse are significantly associated with higher average expression in the promoter region. Furthermore, we found that the B2 in both mouse 2kbp and 20kbp and hAT.Charlie elements in the human 20kbp, are also significantly associated with up-regulated gene expression in the 2kpb promoter. In addition to Alu and B2 in 2kbp, we found that the ERV1 have a significant association with higher average expression in the 20kbp promoter in mouse tissues. We also found that L1 and Simple_repeat elements are significantly associated with lower average expression in both human and mouse. Furthermore, in the human, we found that the MIR is also with lower average expression. The effects of Alu elements in both human and mouse are stronger at 2kbp than at 20kbp. In contrast, the L1 effect at 20kbp is stronger than at 2kbp.

We confirmed our results by applying our models to a different gene expression dataset (Ensembl Human BodyMap 2.0). Human BodyMap 2.0 results yielded results similar to our initial results. For example, it showed that Alu elements are associated with higher gene expression, while L1, Low_complexity, MIR, and simple_repeat elements downregulate the gene expression in both 2kbp and 20kbp regions.

Chapter 5 - Discussion and Conclusion

5.1 Discussion

In this work, we have studied the distribution of repetitive DNA elements in ten mode organisms and found much evidence of non-randomness concerning the location, frequency, and strand-preferences of different repeats. Often found near genes are repeats such as the Alu family repeats in human and mouse, as well as GC-rich simple and low complexity repeats in the most organisms. Other repeats such as LINEs in mammals are more frequently found away from the genes. Also, some of the repeats show strong strand-bias compared to nearby genes, which indicates that these retrotransposons might be linked to the evolution of these genes. We also identified many LTRs that are specifically enriched in promoter regions, some with a strong bias towards the same strand as the nearby gene. This raises the possibility that the LTRs, may play a regulatory role. Compared to LINEs and SINEs, LTRs have a higher degree of diversity, which supports the possibility of their performing regulatory functions. While the composition of different repeat classes and their coverage in mammalian genomes are similar, vast differences can be seen among the various vertebrate genomes. In each organism, there are examples of extremely prevalent repeats successfully fixed in the genome. The most frequently observed transposable elements in mammals is SINE followed by LINE. In contrast, DNA transposons, LINE, and low complexity repeats are the most commonly observed repeat classes in the zebrafish, chicken, and C. elegans genomes, respectively. These repeats may have a substantial influence on the genetic landscape of the genomes.

We have shown that repetitive DNA elements vary in their coverage among organism, from 7.3% in the Fugu genome to 52% in zebrafish. With the exception of C.

elegans and fruit fly, the frequency of the TEs follows a log-normal distribution, characterized by a few highly prevalent repeats in each organism. Surprisingly, we found that most intronic repeats, excluding DNA transposons, have a strong tendency to be on the opposite DNA strand as the host gene. One possible explanation is that intronic RNAs that resulted from splicing may contribute to retrotransposition to the original intronic loci.

Our findings from exploratory data analysis of repetitive DNA elements strongly suggest that there is a potential impact of these repeats on the gene expression. Although most transposable elements are primarily involved in reduced gene expression, our model's results showed that Alu elements in both human and mouse are significantly associated with higher average expression in the promoter region. Furthermore, we found that the B2 in both mouse 2kbp and 20kbp and hAT.Charlie elements in the human 20kbp, are also significantly associated with up-regulated gene expression in the 2kpb promoter. In addition to Alu and B2 in 2kbp, we found that the ERV1 have a significant association with higher average expression in the 20kbp promoter in mouse tissues. We also found that L1 and Simple_repeat elements are significantly associated with lower average expression in both human and mouse. Furthermore, in the human, we found that the MIR is also with lower average expression. The effects of Alu elements in both human and mouse are stronger at 2kbp than at 20kbp. In contrast, the L1 effect at 20kbp is stronger than at 2kbp. Based on previous studies, about 4% of protein-coding sequences include TEs, and Alu elements insertions comprise one-third of them [89]. Thus, Alu elements may play a significant role in modifying gene expression. The effect of Alu elements is stronger at 2kbp comparing with 20kbp. In contrast, the L1 effect in 20kbp is stronger than 2kbp.

5.2 Conclusion

Together with other recent studies, our results indicate that comparative studies of TEs in multiple organisms can lead to insights into their evolution and expansion, thus elucidating their potential function. The non-random of distribution of repeats across multiple organisms adds to the existing evidence that some repetitive DNA elements are drivers of genome evolution, rather than just "junk" DNA.

5.3 Potential weakness of this study and future work

Due to the lack of biological replications, we were only able to use the overall gene expression to determine the association between repetitive DNA elements and gene expression. All human and mouse tissues that were used to quantify the gene levels were normal tissues.

Possible future work may consider comparing the impact of repetitive DNA elements on gene expression between normal and cancer tissues. We may also use different distances upstream from the promoter region of genes to explore their function.

REFERENCES

1. López-Flores I, & Garrido-Ramos, M. : **The repetitive DNA content of eukaryotic genomes.** *Genome Dyn* 2012, **7**:1–28.
2. Ponomarenko M, Orlova G, Kolchanov N: **Unique DNA** In: *Brenner's Encyclopedia of Genetics (Second Edition).* vol. 7: Elsevier; 2013: 259-262.
3. Haubold B, Wiehe T: **How repetitive are genomes?** *BMC Bioinformatics* 2006, **7**:541-541.
4. Waring M, Britten RJ: **Nucleotide Sequence Repetition: A Rapidly Reassociating Fraction of Mouse DNA**. *Science* 1966, **154**(3750):791-794.
5. Britten RJ, Kohne DE: **Repeated Sequences in DNA**. *Science* 1968, **161**(3841):529-540.
6. Liang K-C, Tseng JT, Tsai S-J, Sun HS: **Characterization and distribution of repetitive elements in association with genes in the human genome**. *Computational Biology and Chemistry* 2015, **57**:29-38.
7. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nat Rev Genet* 2012, **13**(1):36-46.
8. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD: **Repetitive elements may comprise over two-thirds of the human genome**. *PLoS Genet* 2011, **7**(12):e1002384.
9. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME *et al*: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
10. Venkatesh B, Gilligan P, Brenner S: **Fugu: a compact vertebrate reference genome**. *FEBS Letters* 2000, **476**(1–2):3-7.
11. Tyekucheva S, Yolken RH, McCombie WR, Parla J, Kramer M, Wheelan SJ, Sabunciyan S: **Establishing the baseline level of repetitive element expression in the human cortex**. *BMC Genomics* 2011, **12**:495-495.
12. Padeken J, Zeller P, Gasser SM: **Repeat DNA in genome organization and stability**. *Current Opinion in Genetics & Development* 2015, **31**:12-19.
13. Lerat E: **Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs**. *Heredity* 2010, **104**(6):520-533.
14. Saha S, Bridges S, Magbanua ZV, Peterson DG: **Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences**. *Tropical Plant Biology* 2008, **1**(1):85-96.
15. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y: **Simple Sequence Repeats in Escherichia coli: Abundance, Distribution, Composition, and Polymorphism**. *Genome Research* 2000, **10**(1):62-71.
16. Ellegren H: **Microsatellites: simple sequences with complex evolution**. *Nature Reviews Genetics* 2004, **5**(6):435-445.
17. Ramel C: **Mini- and microsatellites**. *Environmental Health Perspectives* 1997, **105**(Suppl 4):781-789.
18. Moran JV, Morrish TA: **Chromosomes: Noncoding DNA (including Satellite DNA)**. In: *eLS.* John Wiley & Sons, Ltd; 2001.

19. Jeffreys AJ, Wilson V, Thein SL: **Hypervariable 'minisatellite' regions in human DNA**. *Nature* 1985, **314**(6006):67-73.
20. Gelfand Y, Rodriguez A, Benson G: **TRDB--the Tandem Repeats Database**. *Nucleic Acids Res* 2007, **35**(Database issue):D80-87.
21. Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM: **Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex**. *G3: Genes|Genomes|Genetics* 2017, **7**(2):693-704.
22. Garrido-Ramos M: **Satellite DNA: An Evolving Topic**. *Genes* (2017), **8(9)**(230).
23. Miller WJ: **Mobile genetic elements: Protocols and genomic applications**, vol. 260: Springer Science & Business Media; 2004.
24. Ravindran S: **Barbara McClintock and the discovery of jumping genes**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(50):20198-20199.
25. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements**. *Trends in Genetics* 2003, **19**(2):68-72.
26. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 Maize Genome: Complexity, Diversity, and Dynamics**. *Science* 2009, **326**(5956):1112-1115.
27. Han JS, Szak ST, Boeke JD: **Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes**. *Nature* 2004, **429**(6989):268-274.
28. Vassetzky NS, Kramerov DA: **SINEBase: a database and tool for SINE analysis**. *Nucleic Acids Research* 2013, **41**(Database issue):D83-D89.
29. Tajaddod M, Tanzer A, Licht K, Wolfinger MT, Badelt S, Huber F, Pusch O, Schopoff S, Janisiw M, Hofacker I *et al*: **Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity**. *Genome Biology* 2016, **17**(1):220.
30. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution**. *Nat Rev Genet* 2009, **10**(10):691-703.
31. Griffiths AJF: **An introduction to genetic analysis**: New York : W.H. Freeman, c2015. 11th ed.; 2015.
32. Capy P: **A plastic genome**. *Nature* 1998, **396**(6711):522-523.
33. Deininger P: **Alu elements: know the SINEs**. *Genome Biology* 2011, **12**(12):236.
34. Häsler J, Strub K: **Alu elements as regulators of gene expression**. *Nucleic Acids Research* 2006, **34**(19):5491-5497.
35. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
36. Hancks DC, Kazazian HH: **Roles for retrotransposon insertions in human disease**. *Mobile DNA* 2016, **7**(1):9.
37. Muñoz-López M, García-Pérez JL: **DNA transposons: nature and applications in genomics**. *Current genomics* 2010, **11**(2):115-128.
38. Jurka Jea: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogenetic and genome research* 2005, **110** 462-467.

39. Kazazian HH, Jr **Mobile Elements Drivers of Genome Evolution**. *Science* 2004, **303**:1626-1632.
40. Belancio VP, Hedges DJ, Deininger P: **Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health**. *Genome research* 2008, **18**(3):343-358.
41. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution**. *Nat Rev Genet* 2009, **10**(10):691-703.
42. Konkel MK, Batzer MA: **A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome**. In: *Seminars in cancer biology: 2010*. Elsevier: 211-221.
43. Usdin K: **The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases**. *Genome Research* 2008, **18**(7):1011-1019.
44. Mita P, Boeke JD: **How retrotransposons shape genome regulation**. *Current Opinion in Genetics & Development* 2016, **37**:90-100.
45. Biscotti MA, Olmo E, Heslop-Harrison JS: **Repetitive DNA in eukaryotic genomes**. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 2015, **23**(3):415-420.
46. Kent WJea: **The Human Genome Browser at UCSC** *Genome research* 2002, **12**:996-1006.
47. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**(8):e1003118.
48. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.
49. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L *et al*: **The zebrafish reference genome sequence and its relationship to the human genome**. *Nature* 2013, **496**(7446):498-503.
50. Begon M HJ, Townsend CR: **individuals, populations, and communities**. In: *Ecology*. Oxford ; Cambridge, Mass.: Blackwell Science; 1996.
51. **Power laws, Pareto distributions and Zipf's law**. *Contemporary Physics* 2005, **46**(5):323-351.
52. Sutton J: **Gibrat's Legacy**. *Journal of Economic Literature* 1997, **35**(1):40-59.
53. Klimopoulos A, Sellis D, Almirantis Y: **Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics**. *Gene* 2012, **499**(1):88-98.
54. Miyagishi M, Taira K: **U6 promoter-driven siRNAs with four uridine 3′ overhangs efficiently suppress targeted gene expression in mammalian cells**. *Nat Biotech* 2002, **20**(5):497-500.
55. Kazazian HH, Jr.: **Mobile elements: drivers of genome evolution**. *Science* 2004, **303**(5664):1626-1632.
56. Lynch VJ, Leclerc RD, May G, Wagner GP: **Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals**. *Nature genetics* 2011, **43**(11):1154-1159.

57. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB: **Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos**. *Developmental cell* 2004, **7**(4):597-606.
58. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells**. *Nature genetics* 2010, **42**(7):631-634.
59. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**(5258):561-563.
60. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nature Biotechnology* 2008, **26**(10):1135-1145.
61. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.
62. Chu Y, Corey DR: **RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation**. *Nucleic Acid Therapeutics* 2012, **22**(4):271-274.
63. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS: **The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments**. *Nature protocols* 2012, **7**(8):1534-1550.
64. Han Y, Gao S, Muegge K, Zhang W, Zhou B: **Advanced Applications of RNA Sequencing and Challenges**. *Bioinformatics and Biology Insights* 2015, **9**(Suppl 1):29-46.
65. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X *et al*: **A survey of best practices for RNA-seq data analysis**. *Genome Biology* 2016, **17**:13.
66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature Protocols* 2012, **7**(3):562-578.
67. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*: **Ensembl 2014**. *Nucleic Acids Research* 2014, **42**(D1):D749-D755.
68. Chen C, Khaleel SS, Huang H, Wu CH: **Software for pre-processing Illumina next-generation sequencing short read sequences**. *Source Code for Biology and Medicine* 2014, **9**:8-8.
69. Andrews S, FastQC A: **A quality control tool for high throughput sequence data. 2010**. *Google Scholar* 2015.
70. Hannon G: **Fastx-toolkit**. In.: Cold Spring Harbor Laboratory. http://hannonlab. cshl. edu/fastx_toolkit//(27 Feb. 2014); 2010.
71. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357-359.
72. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.
73. Draper NR, Smith H: **Applied Regression Analysis**. In., 3 edn. New York, NY: John Wiley & Sons; 2014.
74. **R: A language and environment for statistical computing**

75. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for Computing and Annotating Genomic Ranges**. *PLOS Computational Biology* 2013, **9**(8):e1003118.
76. Durinck S, Spellman PT, Birney E, Huber W: **Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt**. *Nature protocols* 2009, **4**(8):1184-1191.
77. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S: **gplots: Various R programming tools for plotting data**. *R package version* 2009, **2**(4):1.
78. John F, Sanford W: **An R companion to applied regression**. In.: SAGE Inc., Thousand Oaks; 2011.
79. Eddelbuettel D: **CRAN task view: High-performance and parallel computing with R**. 2017.
80. Keith TZ: **Multiple regression and beyond: An introduction to multiple regression and structural equation modeling**: Routledge; 2014.
81. Darlington RB: **Multiple regression in psychological research and practice**. *Psychological Bulletin* 1968, **69**(3):161-182.
82. Osborne JW, Waters E: **1 Four Assumptions Of Multiple Regression That Researchers Should Always Test**. 2002.
83. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J: **Evaluation of the lasso and the elastic net in genome-wide association studies**. *Frontiers in Genetics* 2013, **4**:270.
84. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society Series B (Methodological)* 1996:267-288.
85. Friedman JH, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *2010* 2010, **33**(1):22.
86. Zou H, Hastie T: **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67**(2):301-320.
87. Izenman AJ: **Modern multivariate statistical techniques**. *Regression, classification and manifold learning* 2008.
88. Cramer EM, Nicewander WA: **Some symmetric, invariant measures of multivariate association**. *Psychometrika* 1979, **44**(1):43-54.

APPENDICES

Appendix A1: RepeatMasker files and chromosomes information for each organism

Human: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/

Mouse: http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/

Cow: http://hgdownload.soe.ucsc.edu/goldenPath/bosTau7/database/

Rat: http://hgdownload.soe.ucsc.edu/goldenPath/rn5/database/

Rhesus: http://hgdownload.soe.ucsc.edu/goldenPath/rheMac2/database/

Chicken: http://hgdownload.soe.ucsc.edu/goldenPath/galGal4/database/

Zebrafish: http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/database/

Fruit fly: http://hgdownload.soe.ucsc.edu/goldenPath/dm6/database/

C. elegans: http://hgdownload.soe.ucsc.edu/goldenPath/ce10/database/

Fugu: http://hgdownload.soe.ucsc.edu/goldenPath/fr3/database/

Repeat Masker library release and version information for each species

Human: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/README.txt

Mouse: ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/README.txt

Cow: ftp://hgdownload.cse.ucsc.edu/goldenPath/bosTau7/bigZips/README.txt

Rat: ftp://hgdownload.cse.ucsc.edu/goldenPath/rn5/bigZips/README.txt

Rhesus: ftp://hgdownload.cse.ucsc.edu/goldenPath/rheMac2/bigZips/README.txt

Chicken: ftp://hgdownload.cse.ucsc.edu/goldenPath/galGal4/bigZips/README.txt

Zebrafish: ftp://hgdownload.cse.ucsc.edu/goldenPath/danRer7/bigZips/README.txt

Fugu: ftp://hgdownload.cse.ucsc.edu/goldenPath/fr3/bigZips/README.txt

Fruit fly: ftp://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/README.txt

C. elegans: ftp://hgdownload.cse.ucsc.edu/goldenPath/ce10/bigZips/README.txt

Appendix A2: R script for enrichment/depletion and strand-preference calculations

Github repository for R code: https://github.com/mkmb2004

Appendix A3: Supplementary Figures

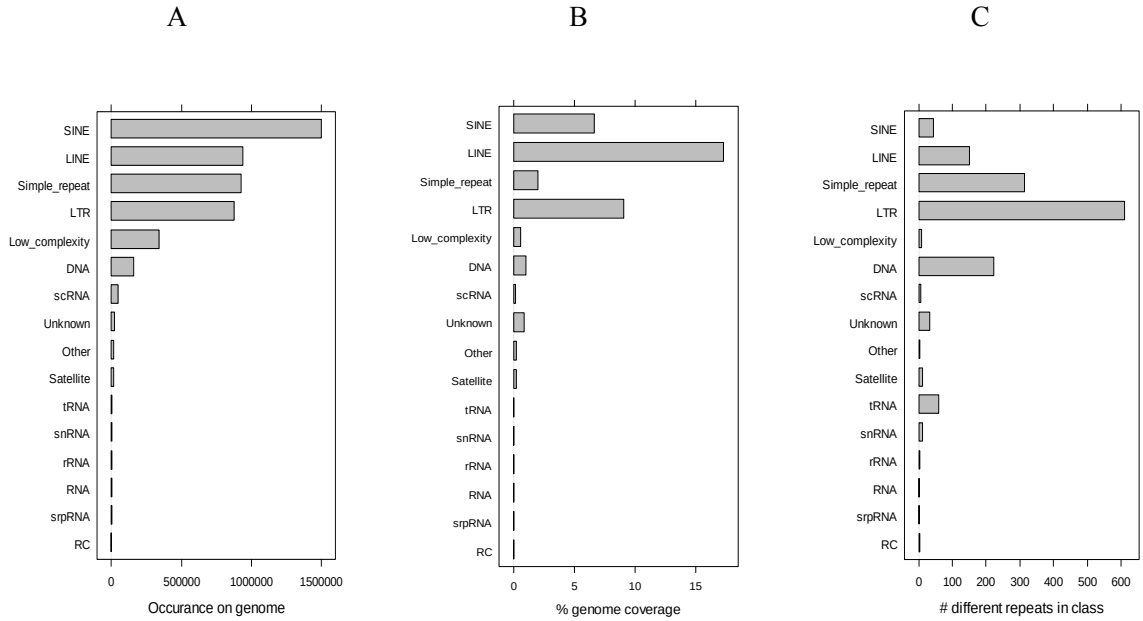

Figure S1. Repetitive DNA by class in the rat genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

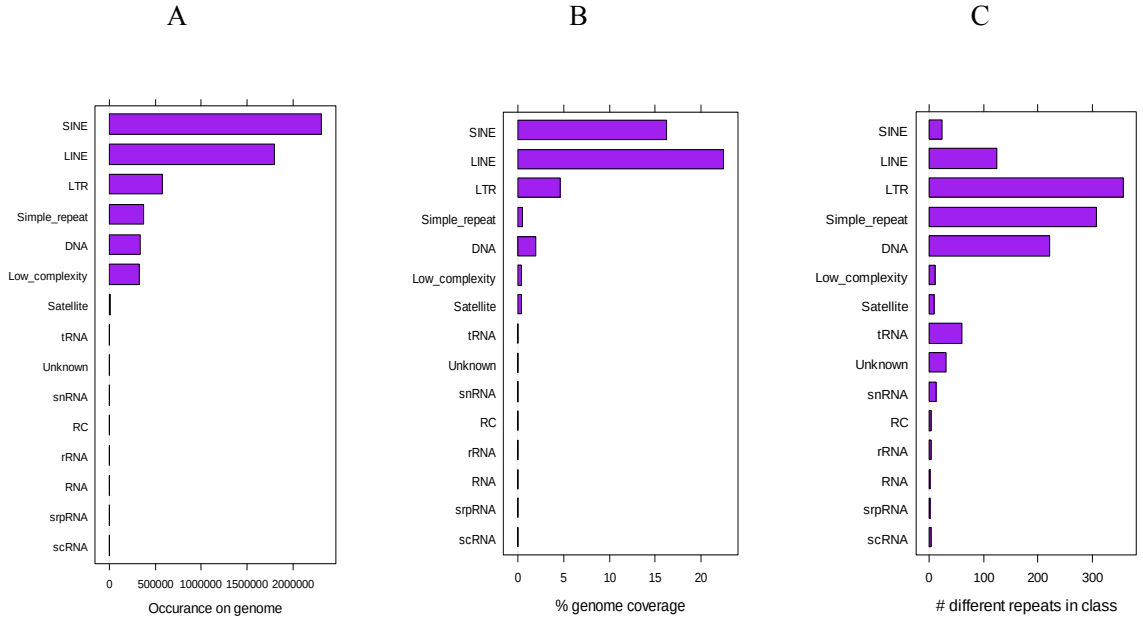

Figure S2. Repetitive DNA by class in the cow genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

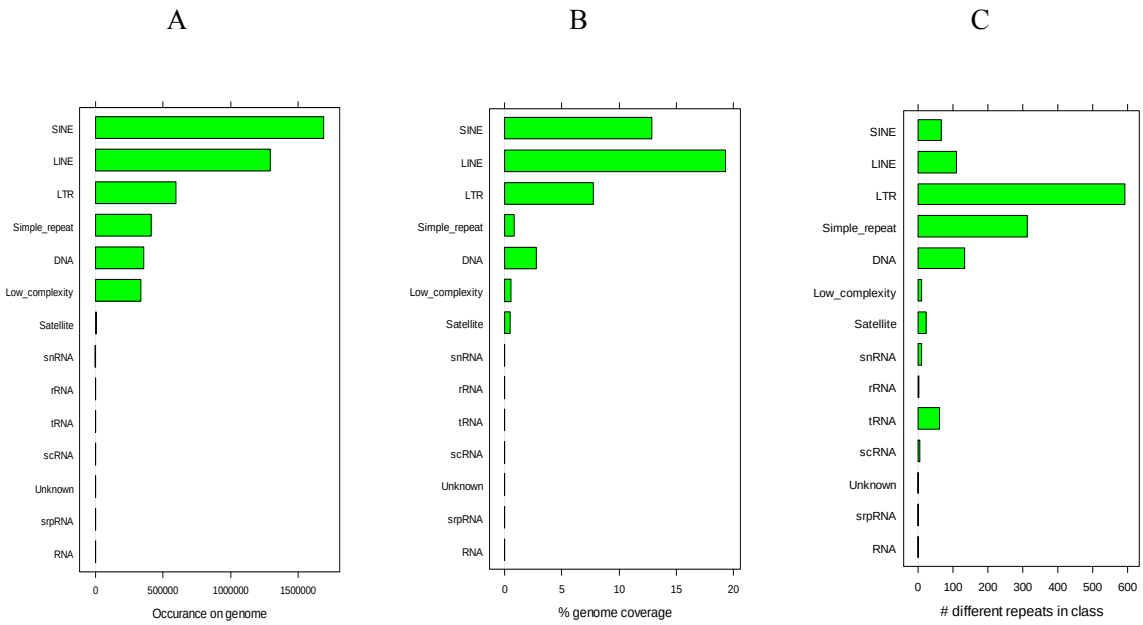

Figure S3. Repetitive DNA by class in the rhesus genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

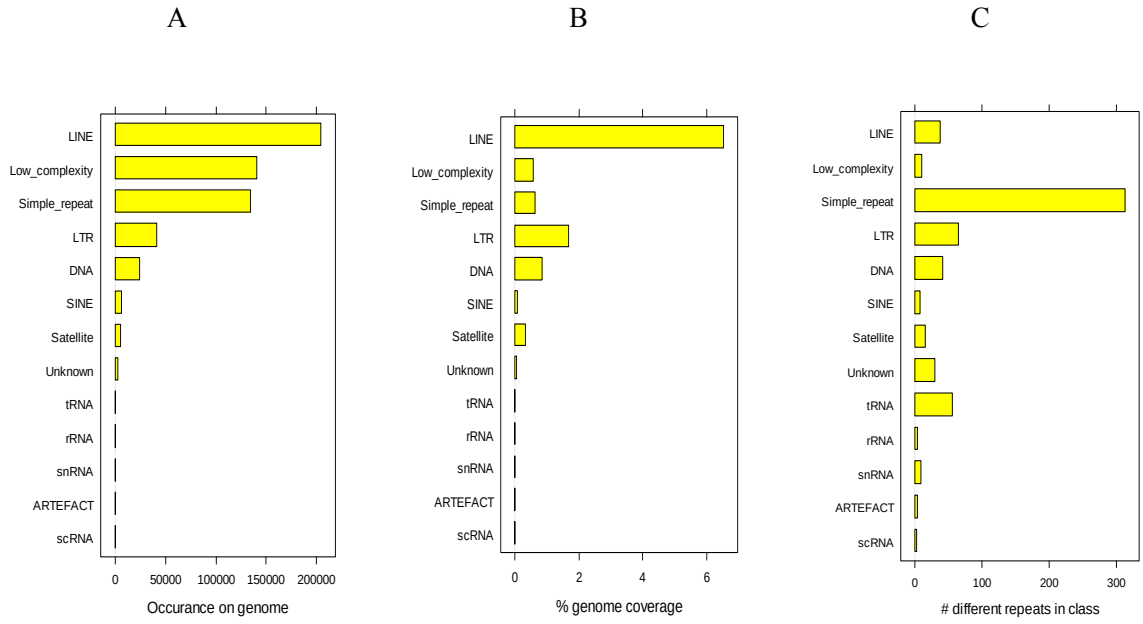

Figure S4. Repetitive DNA by class in the chicken genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

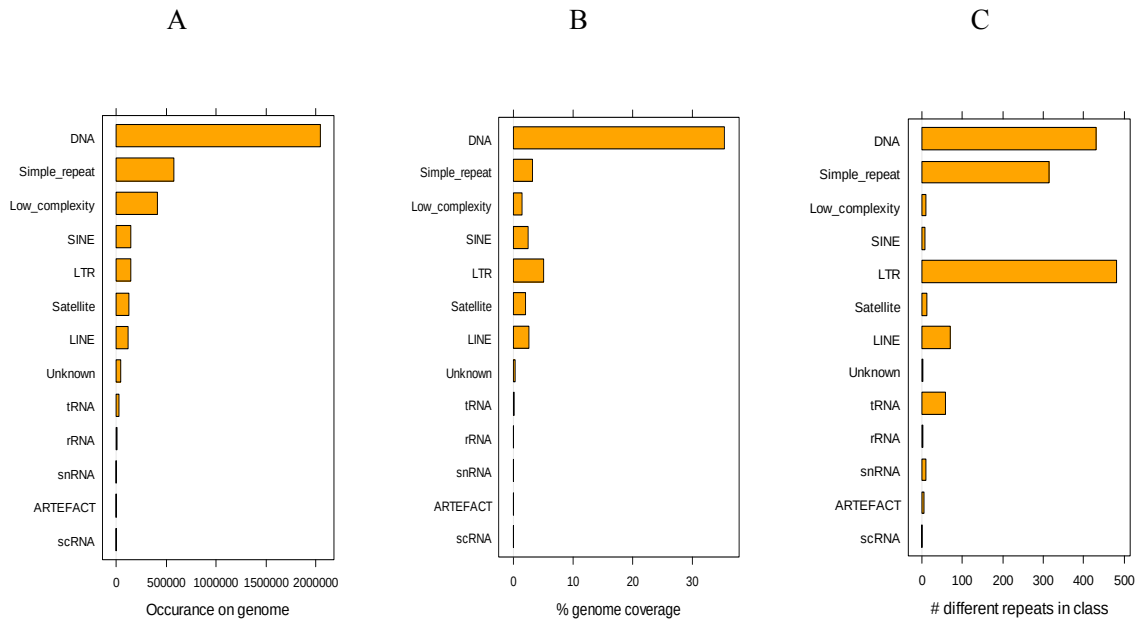

Figure S5. Repetitive DNA by class in the zebrafish genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

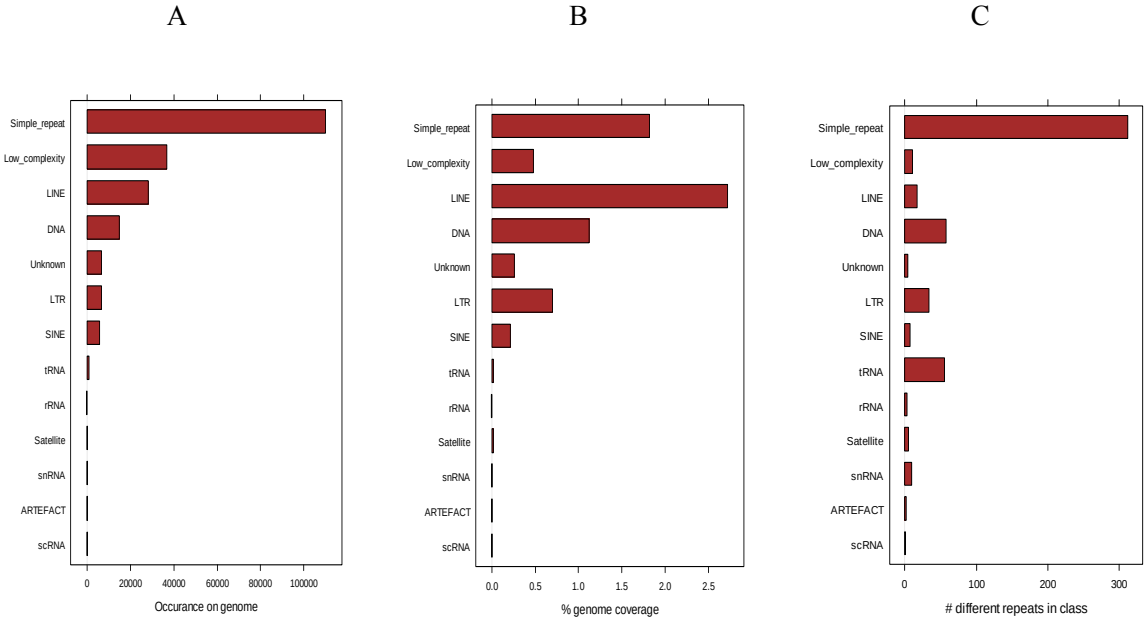

Figure S6. Repetitive DNA by class in the Fugu genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

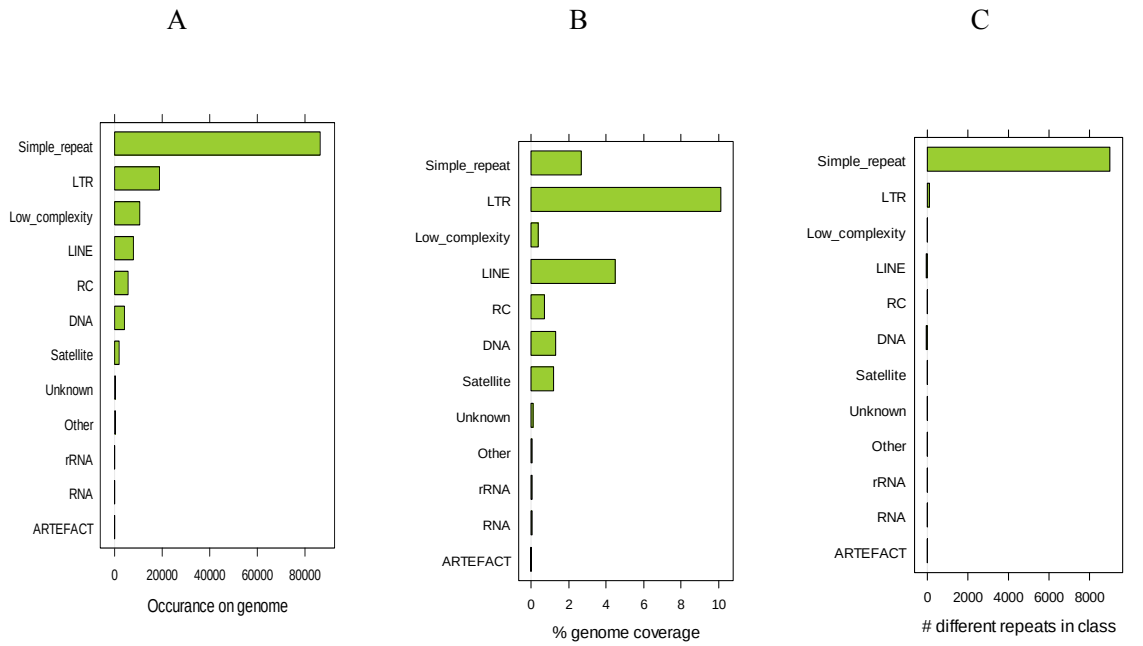

Figure S7. Repetitive DNA by class in the fruit fly genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

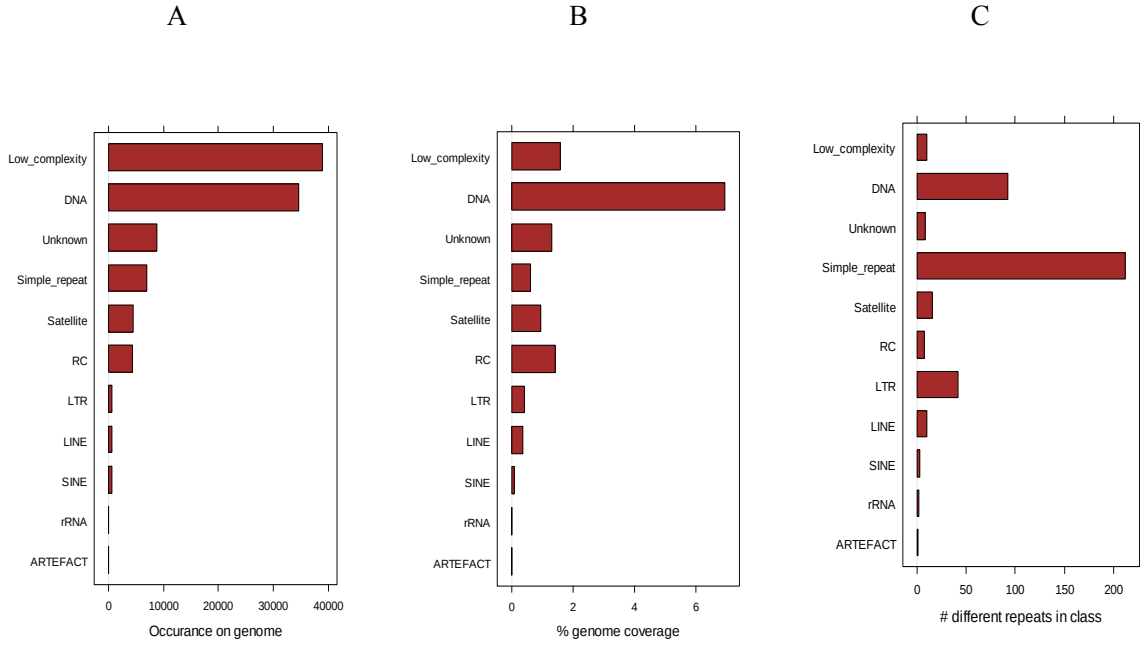

Figure S8. Repetitive DNA by class in the C. elegans genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each class.

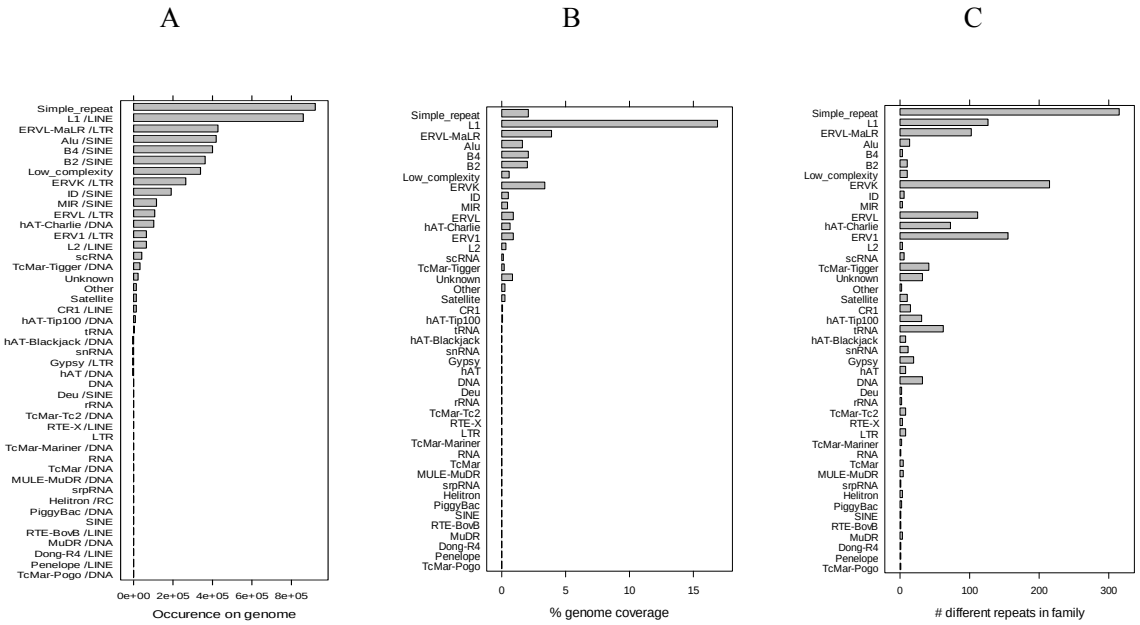

Figure S9. Repetitive DNA by family in the rat genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

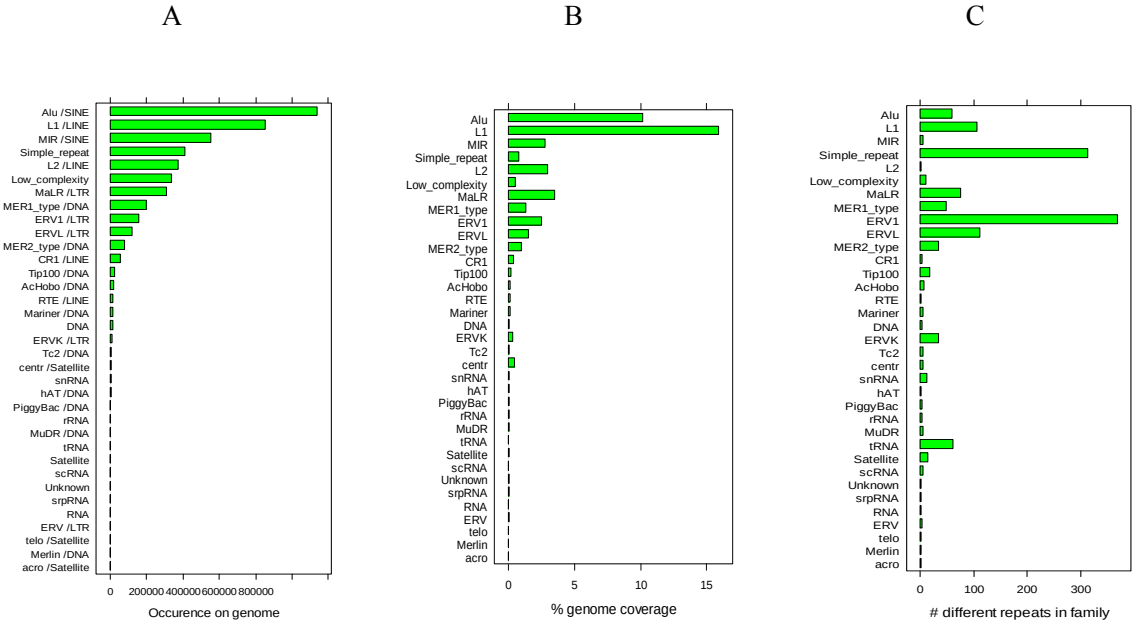

Figure S10. Repetitive DNA by family in the rhesus genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

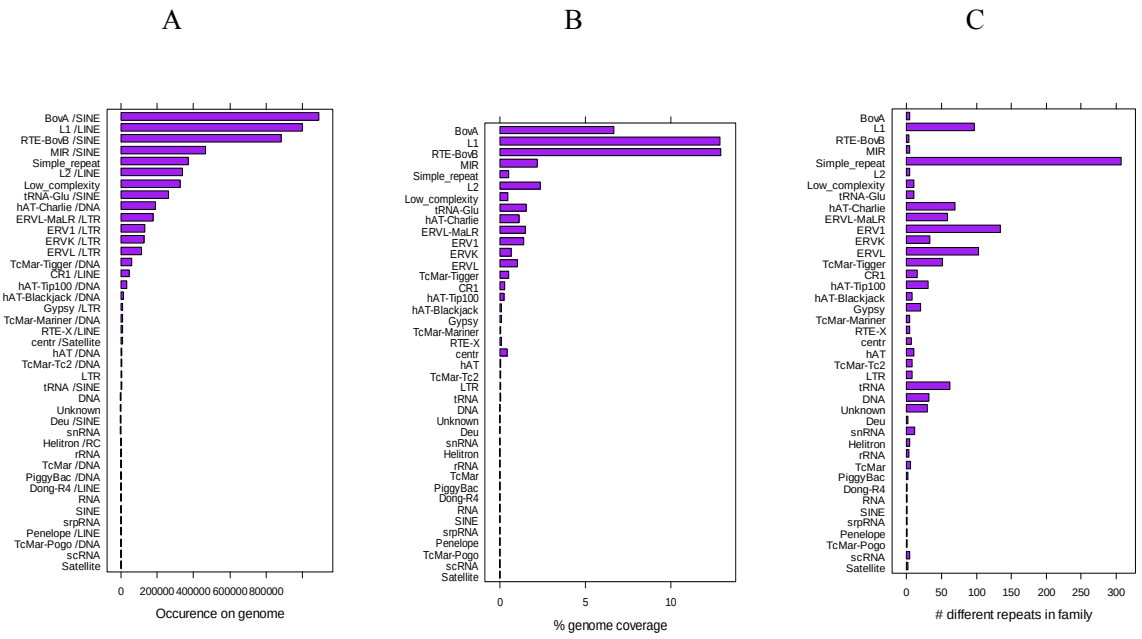

Figure S11. Repetitive DNA by family in the cow genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

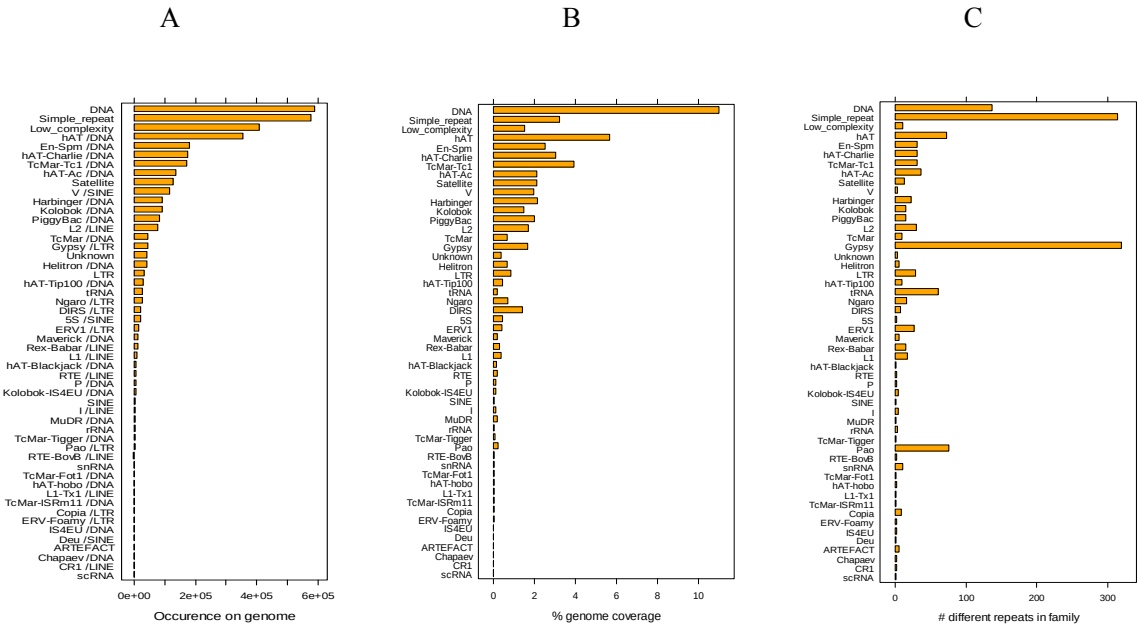

Figure S12. Repetitive DNA by family in the zebrafish genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

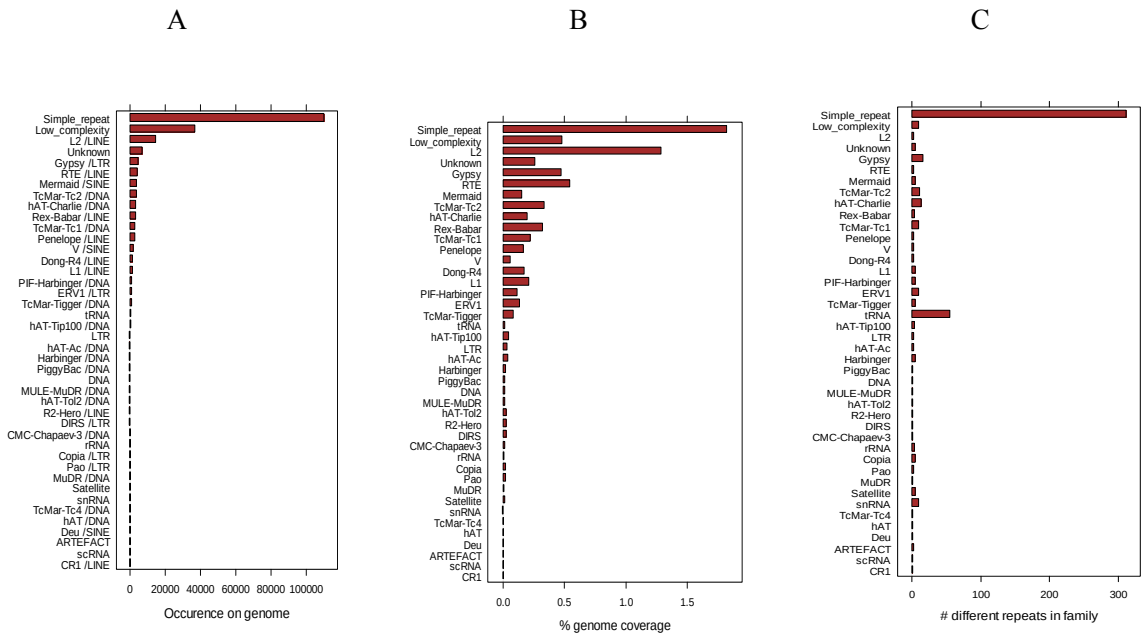

Figure S13. Repetitive DNA by family in the Fugu genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

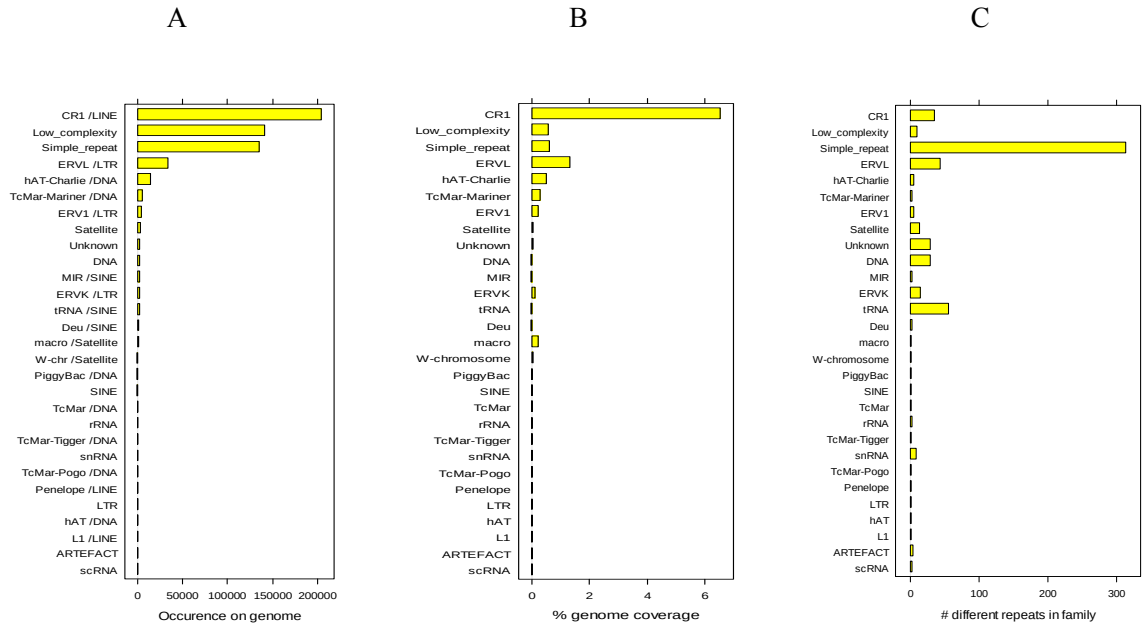

Figure S14. Repetitive DNA by family in the chicken genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

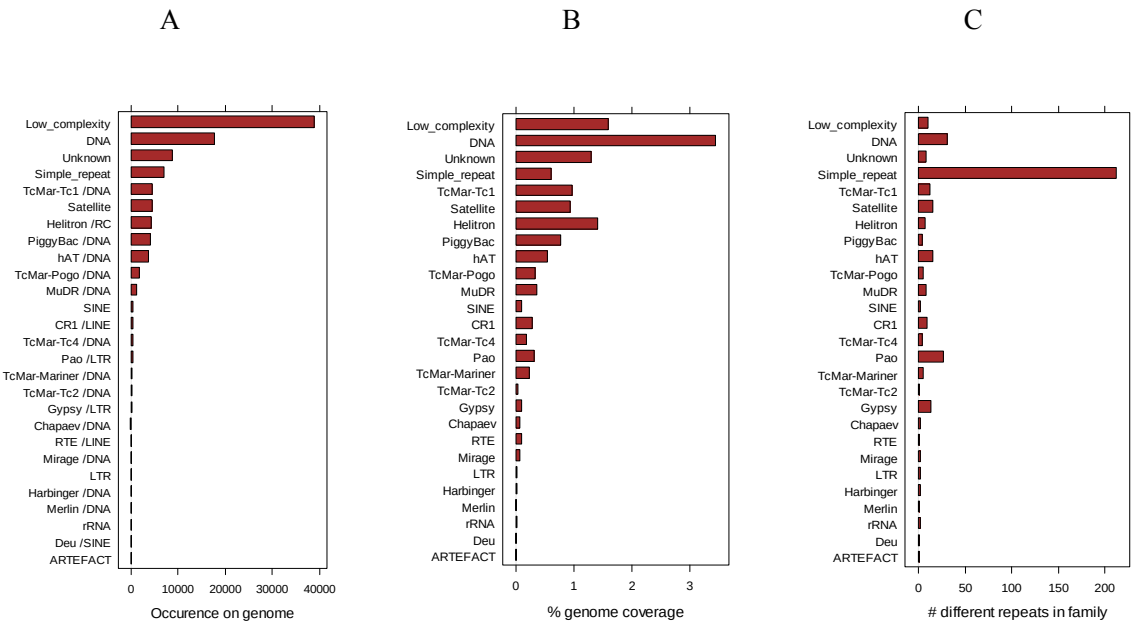

Figure S15. Repetitive DNA by family in the C. elegans genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

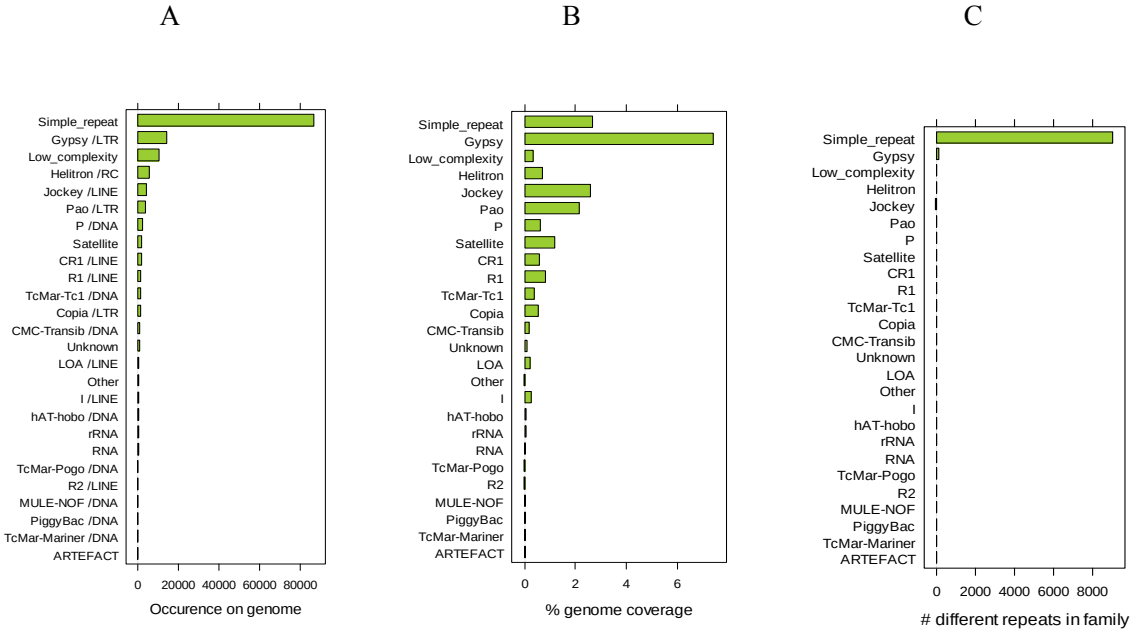

Figure S16. Repetitive DNA by family in the fruit fly genome (A) occurrence, (B) percentage of genomic coverage, and (C) number of different repeats in each family.

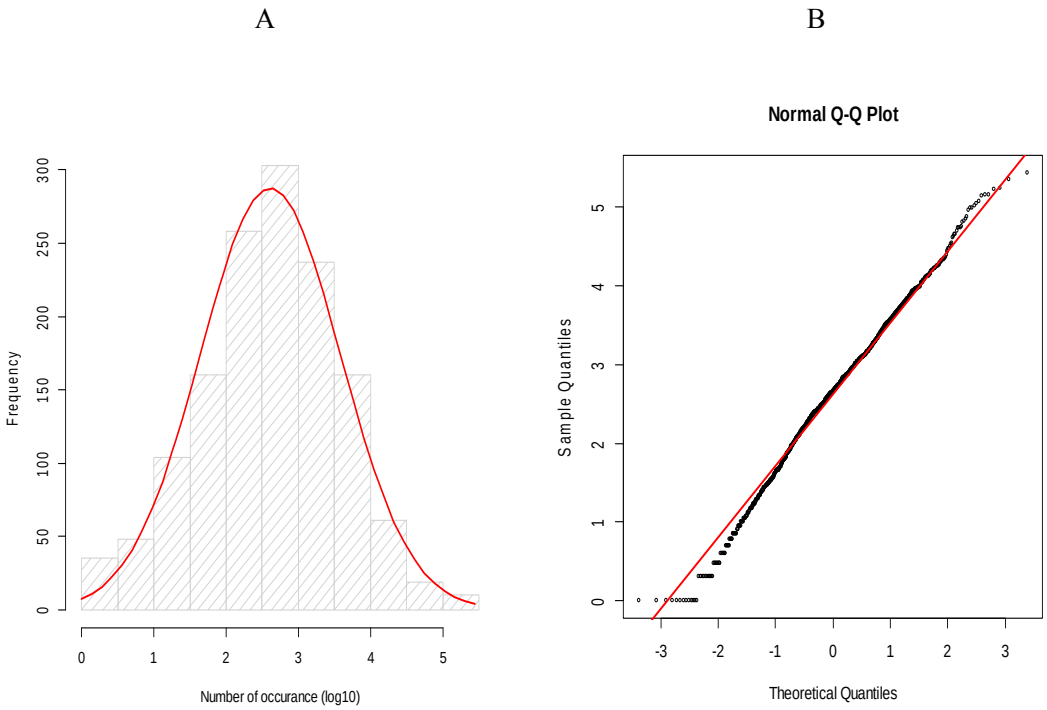

Figure S17. Histogram with normal distribution curve and Q-Q plot for the human repetitive DNA.

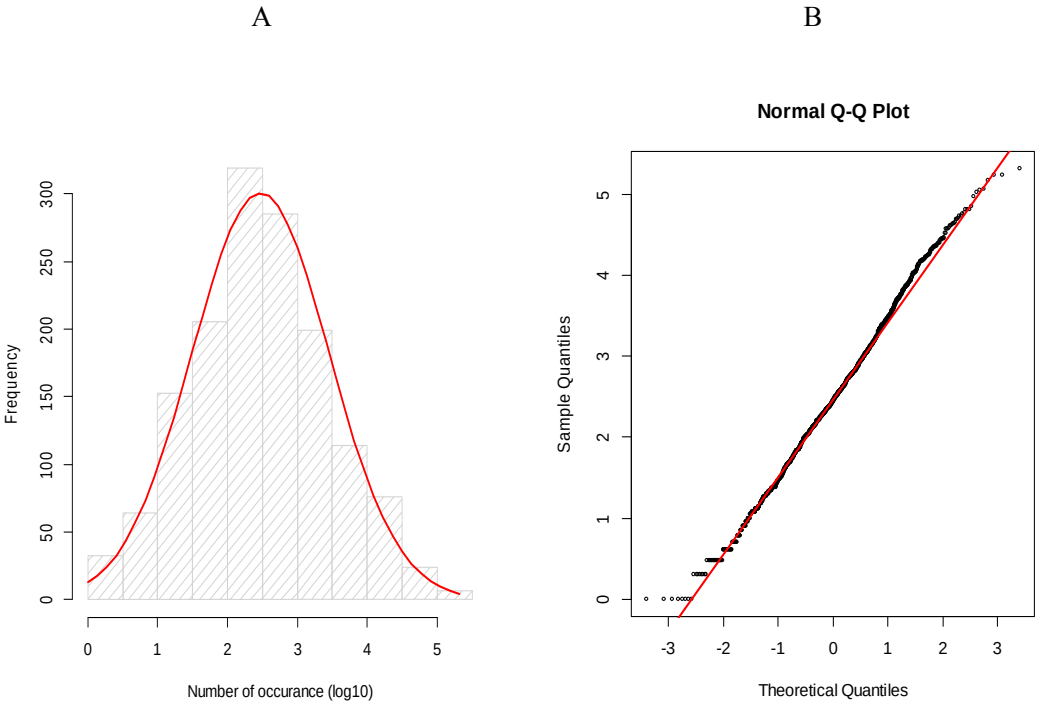

Figure S18. Histogram with normal distribution curve and Q-Q plot for the rat repetitive DNA.

A B

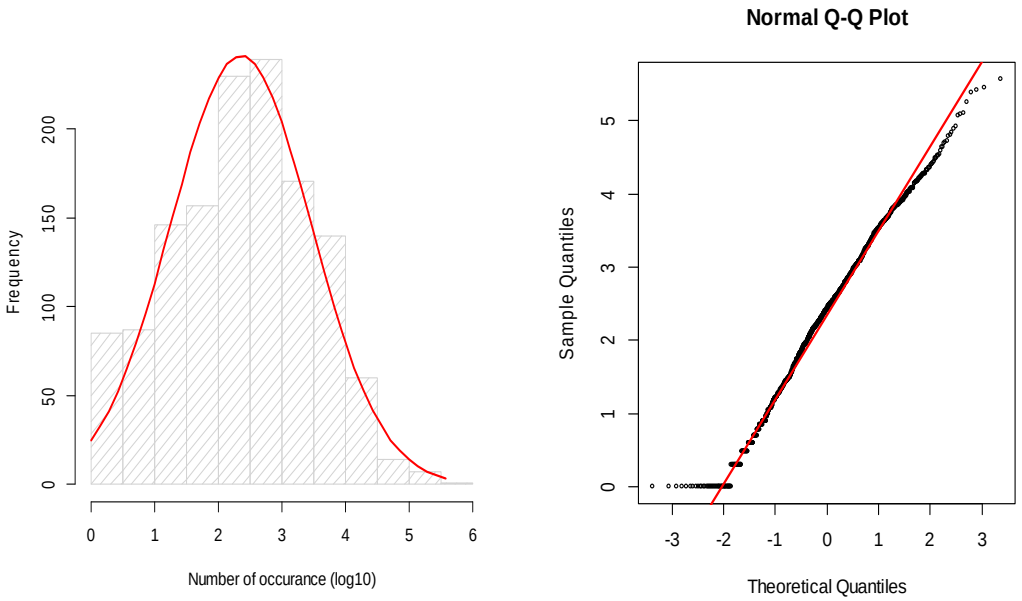

Figure S19. Histogram with normal distribution curve and Q-Q plot for the rhesus repetitive DNA.

A B

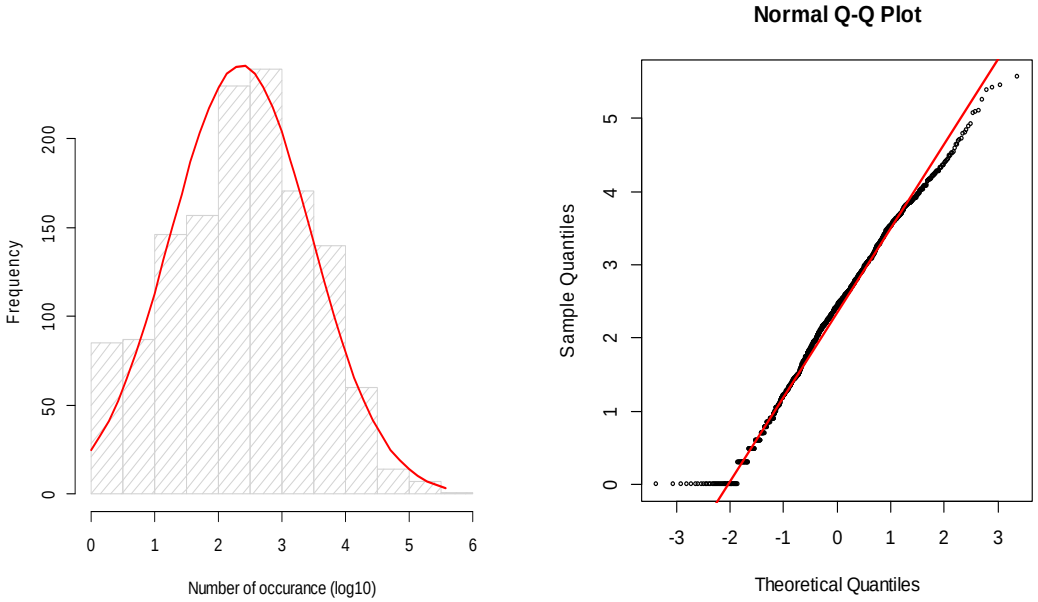

Figure S20. Histogram with normal distribution curve and Q-Q plot for the cow repetitive DNA.

A B

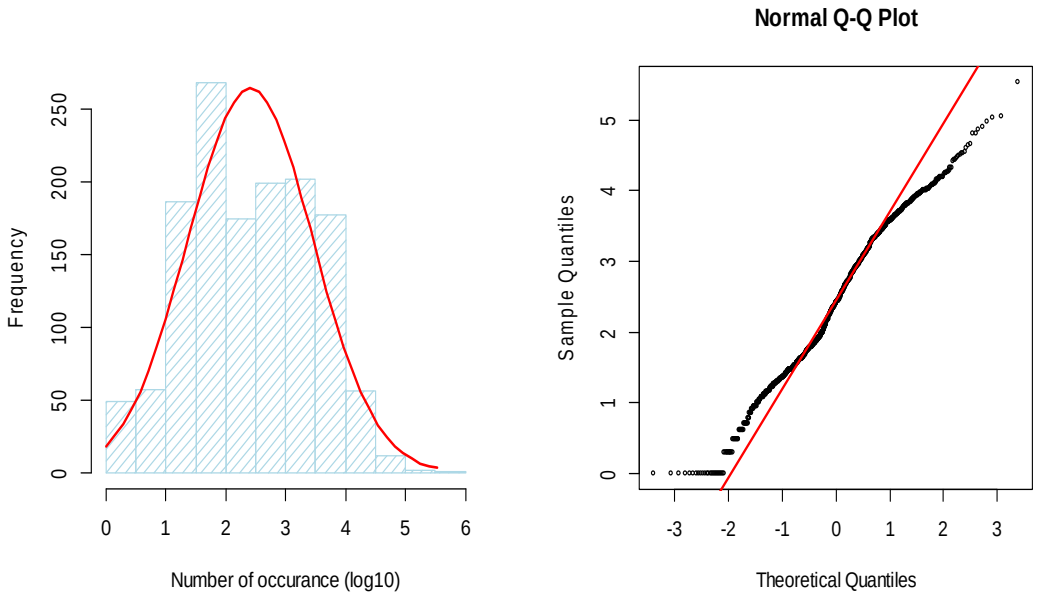

Figure S21. Histogram with normal distribution curve and Q-Q plot for the zebrafish repetitive DNA.

A B

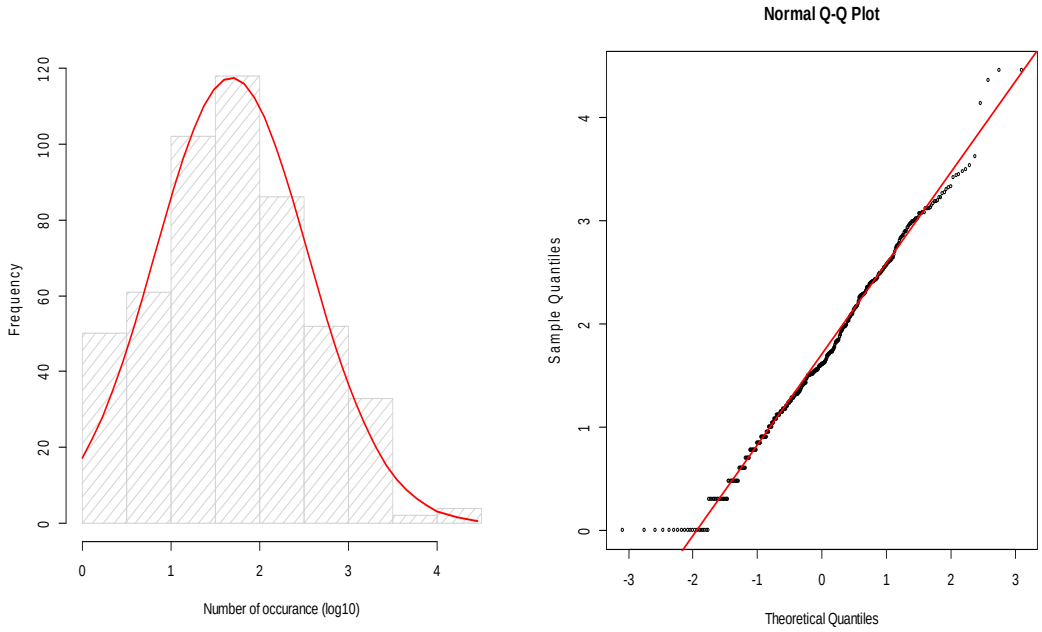

Figure S22. Histogram with normal distribution curve and Q-Q plot for the Fugu repetitive DNA.

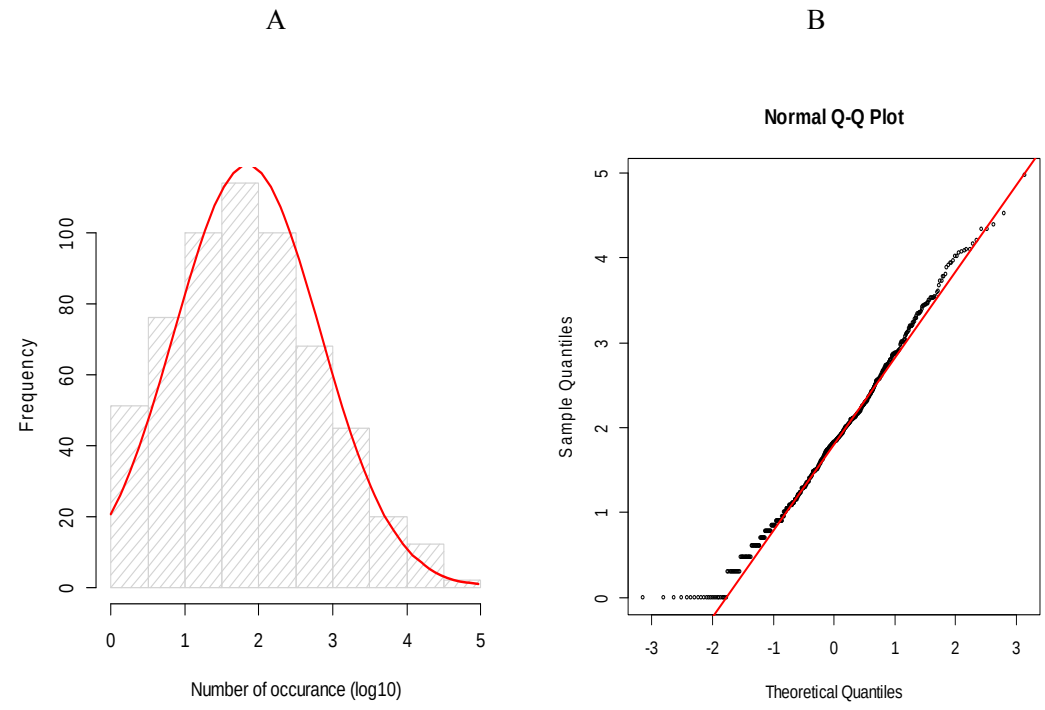

Figure S23. Histogram with normal distribution curve and Q-Q plot for the chicken repetitive DNA.

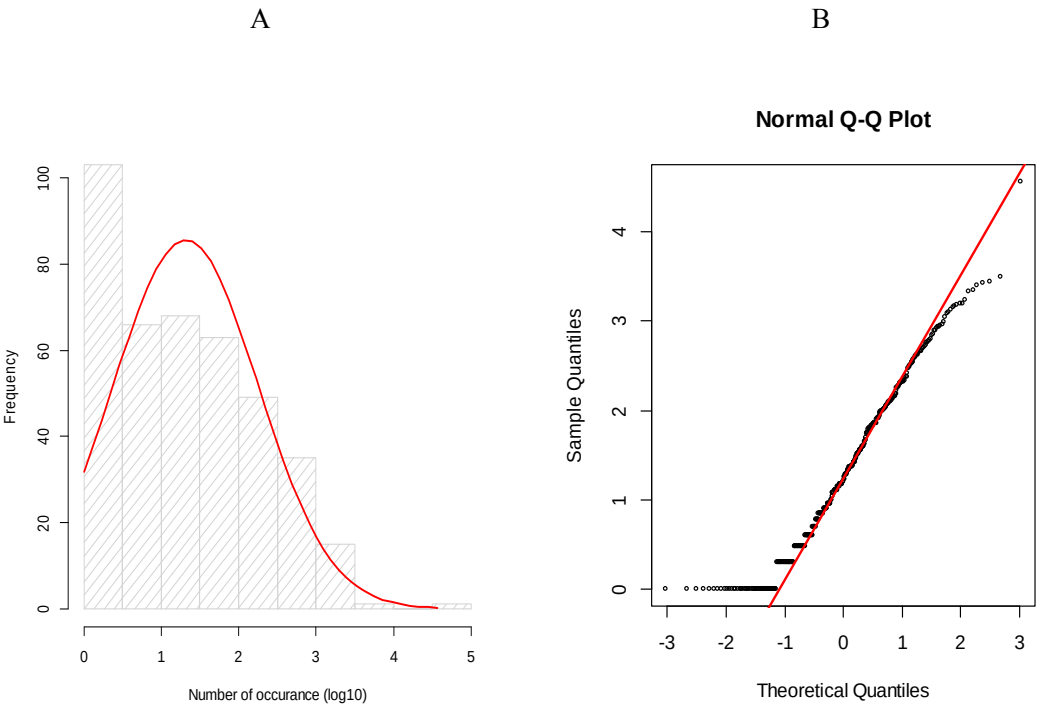

Figure S24. Histogram with normal distribution curve and Q-Q plot for the C. elegans repetitive DNA

A B

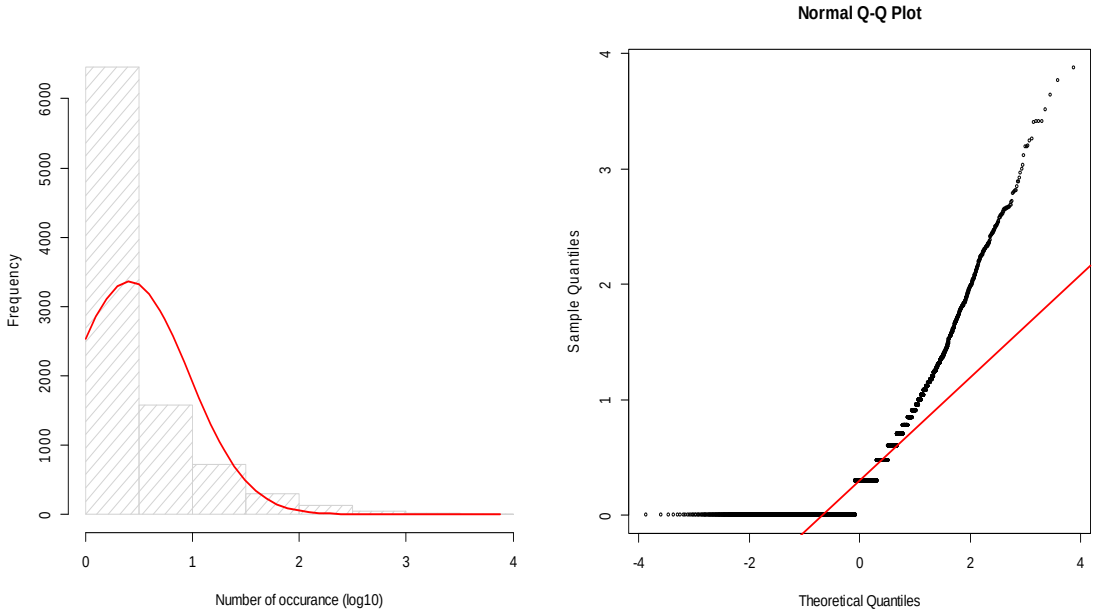

Figure S25. Histogram with normal distribution curve and Q-Q plot for the fruit fly repetitive DNA.

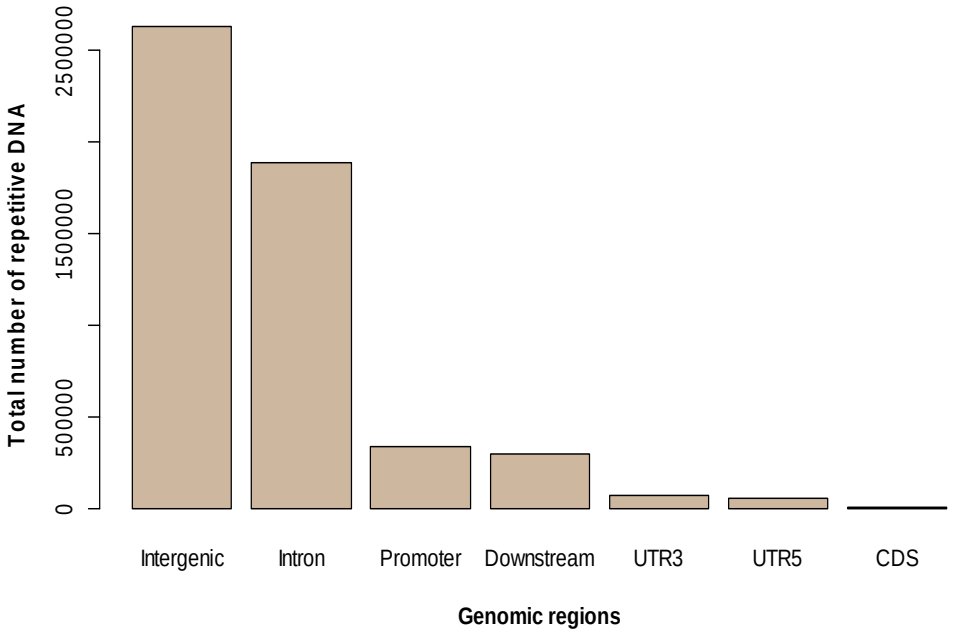

Figure 26. Total number of human repetitive DNA in different genomic regions.

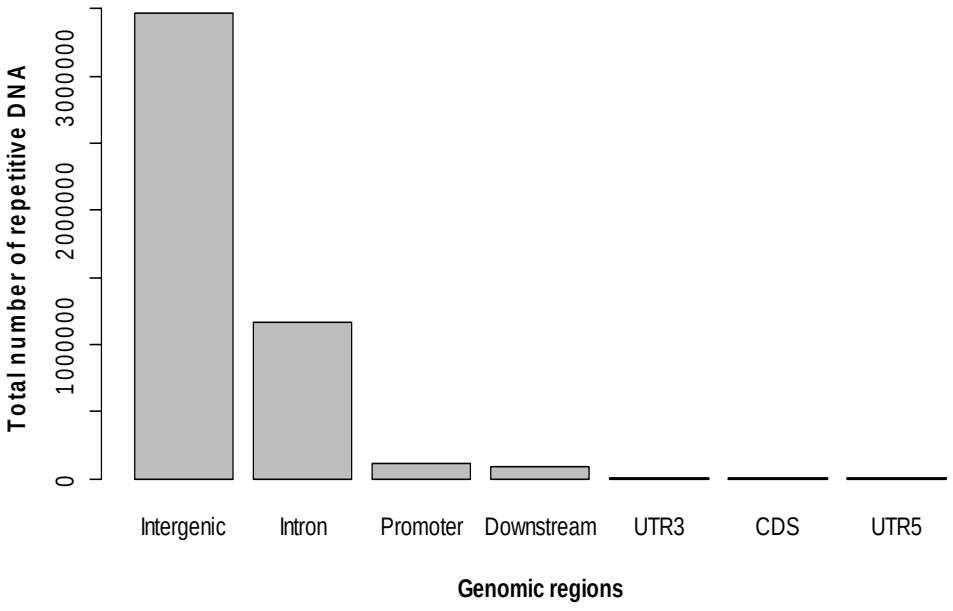

Figure 27. Total number of rat repetitive DNA in different genomic regions.

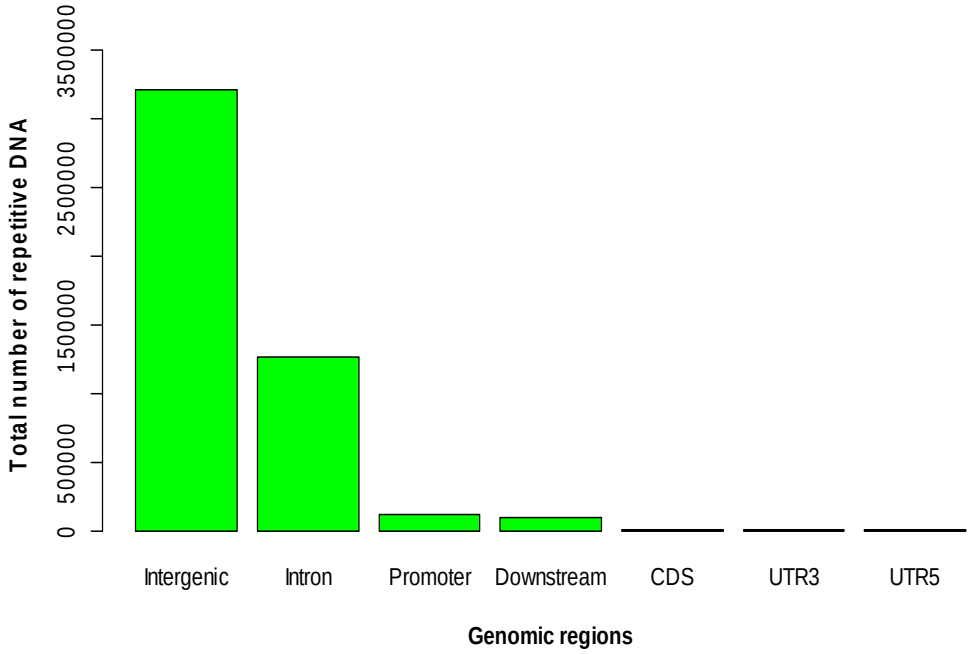

Figure 28. Total number of rhesus repetitive DNA in different genomic regions.

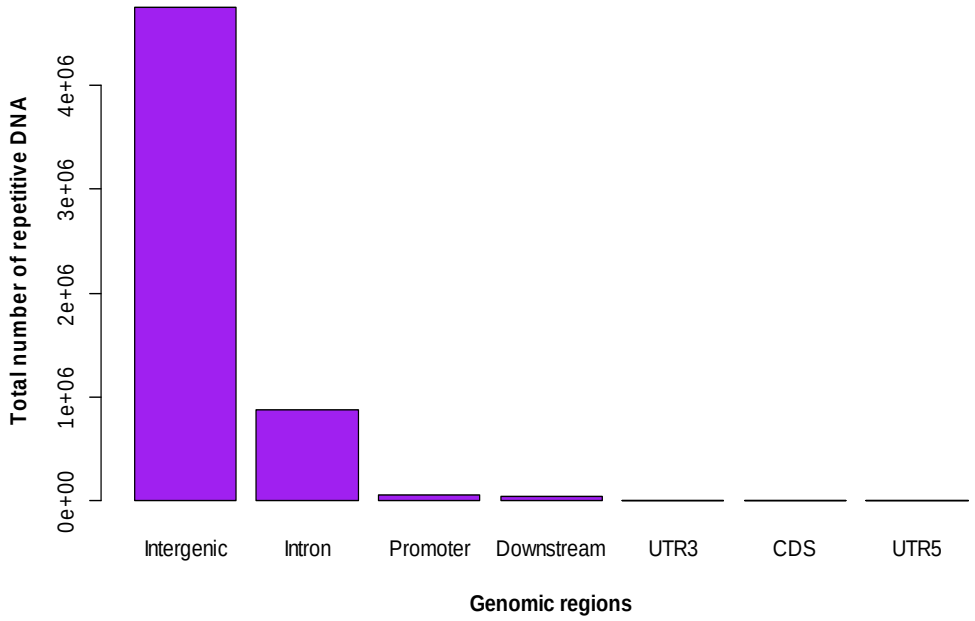

Figure 29. Total number of cow repetitive DNA in different genomic regions.

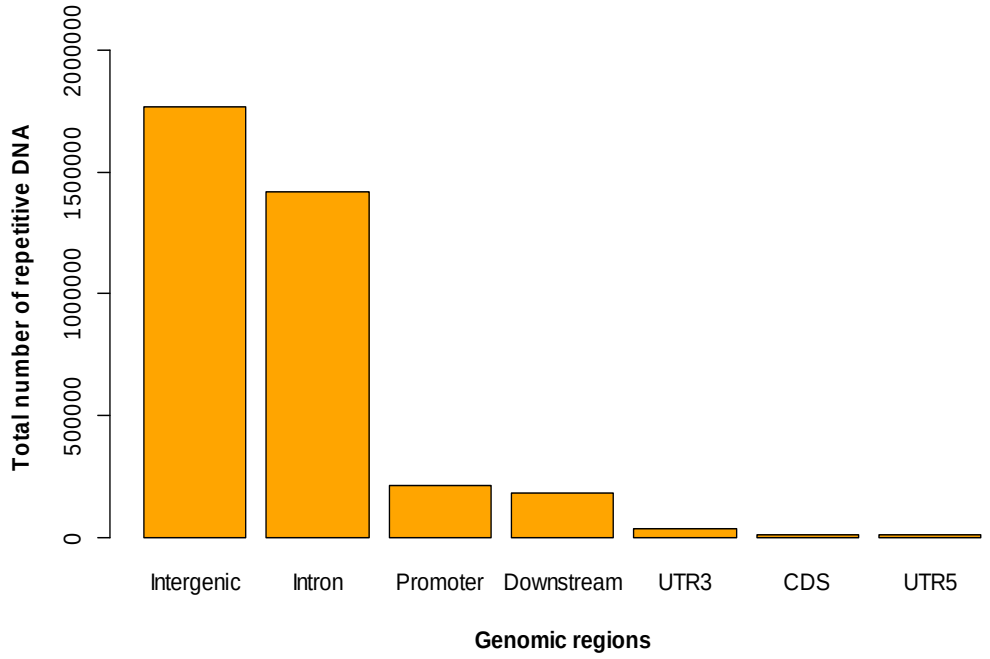

Figure 30. Total number of zebrafish repetitive DNA in different genomic regions.

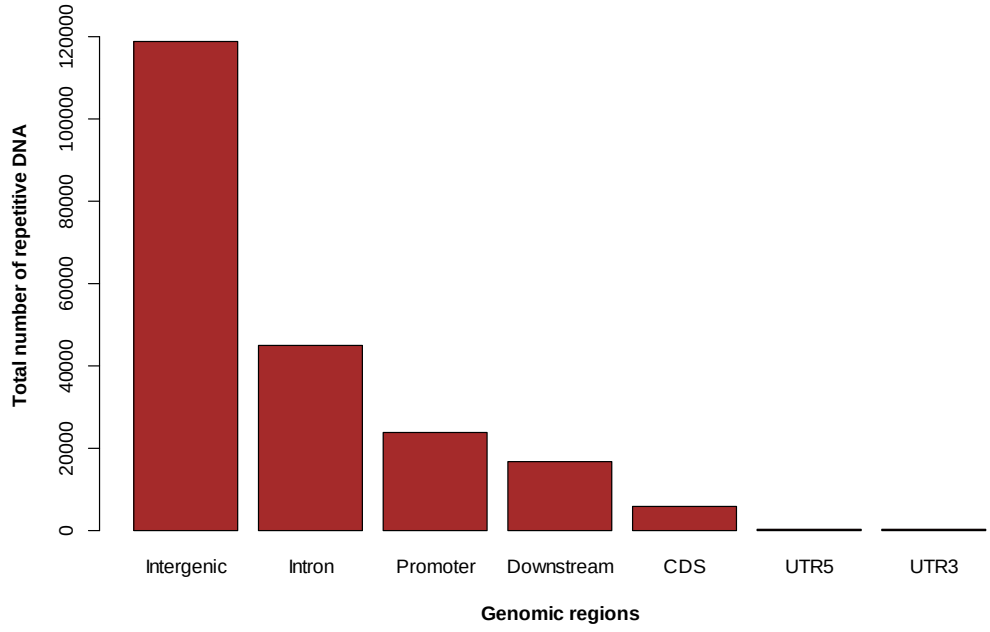

Figure 31. Total number of Fugu repetitive DNA in different genomic regions.

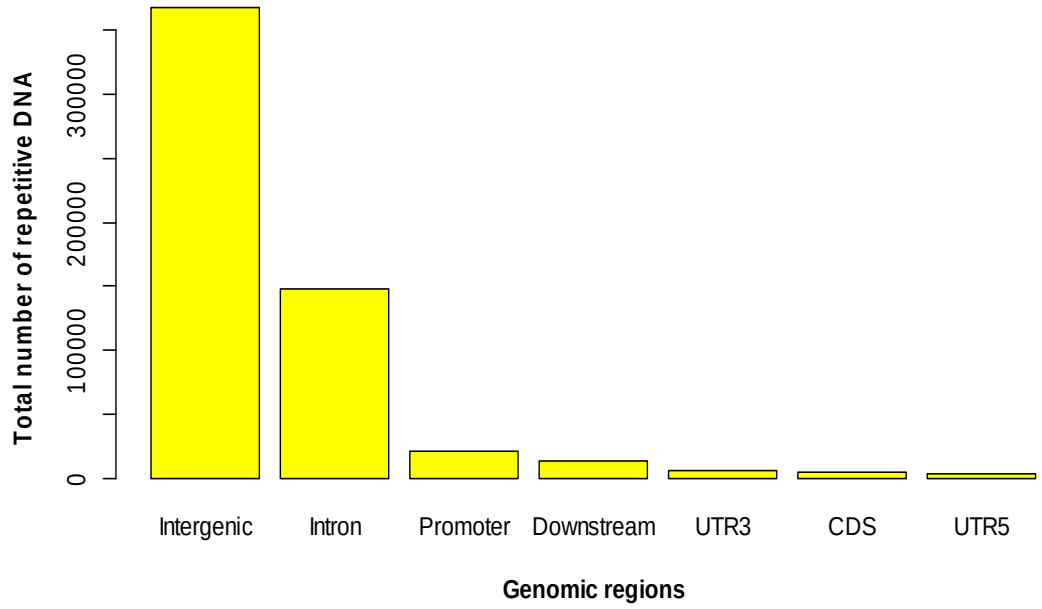

Figure 32. Total number of chicken repetitive DNA in different genomic regions.

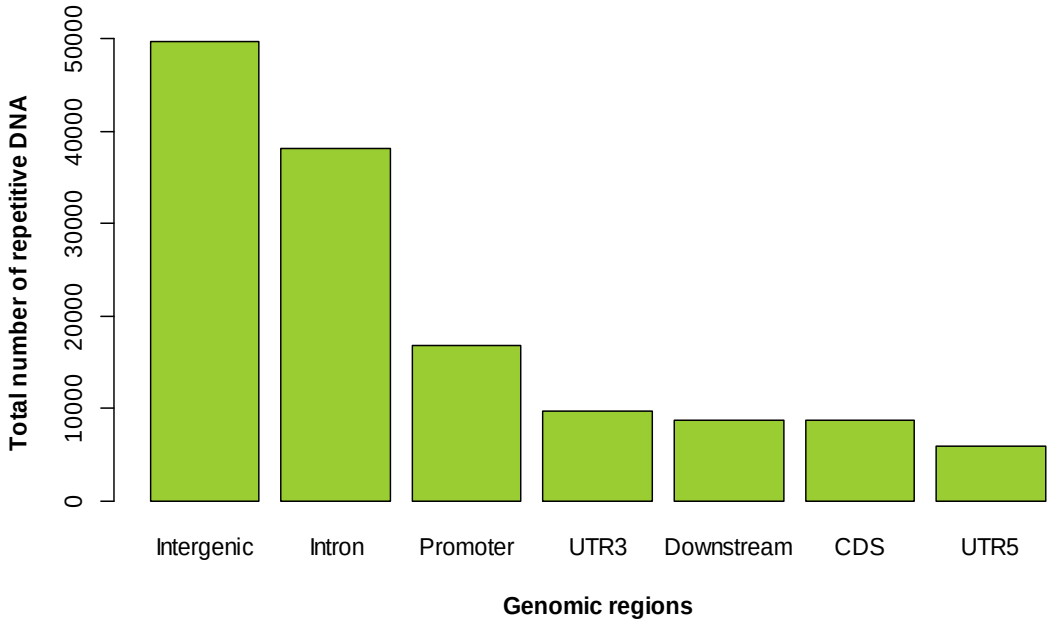

Figure 33. Total number of fruit fly repetitive DNA in different genomic regions.

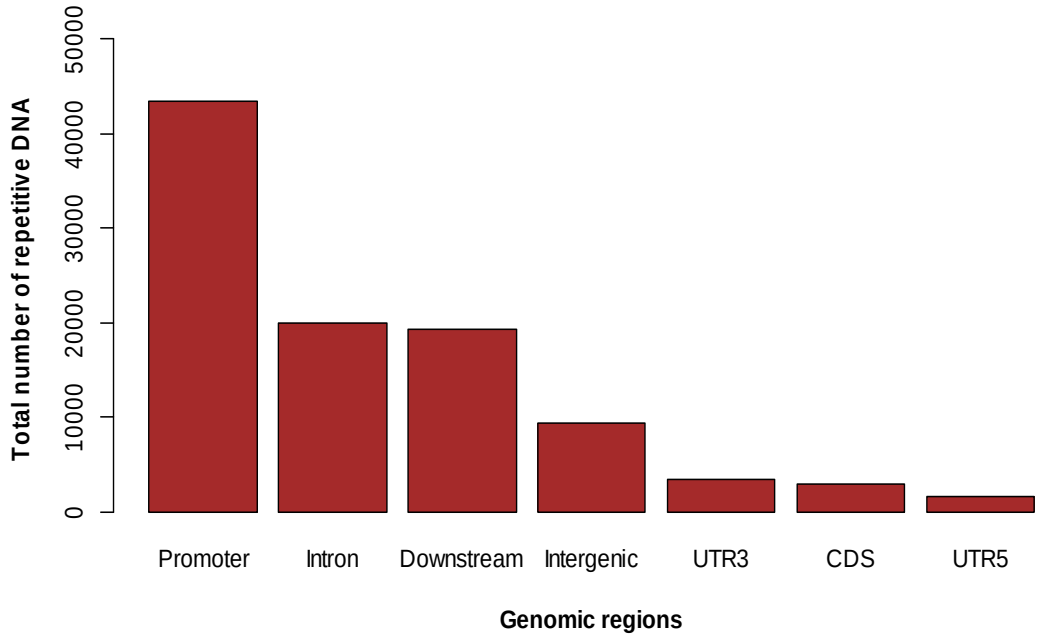

Figure 34. Total number of C. elegans repetitive DNA in different genomic regions.

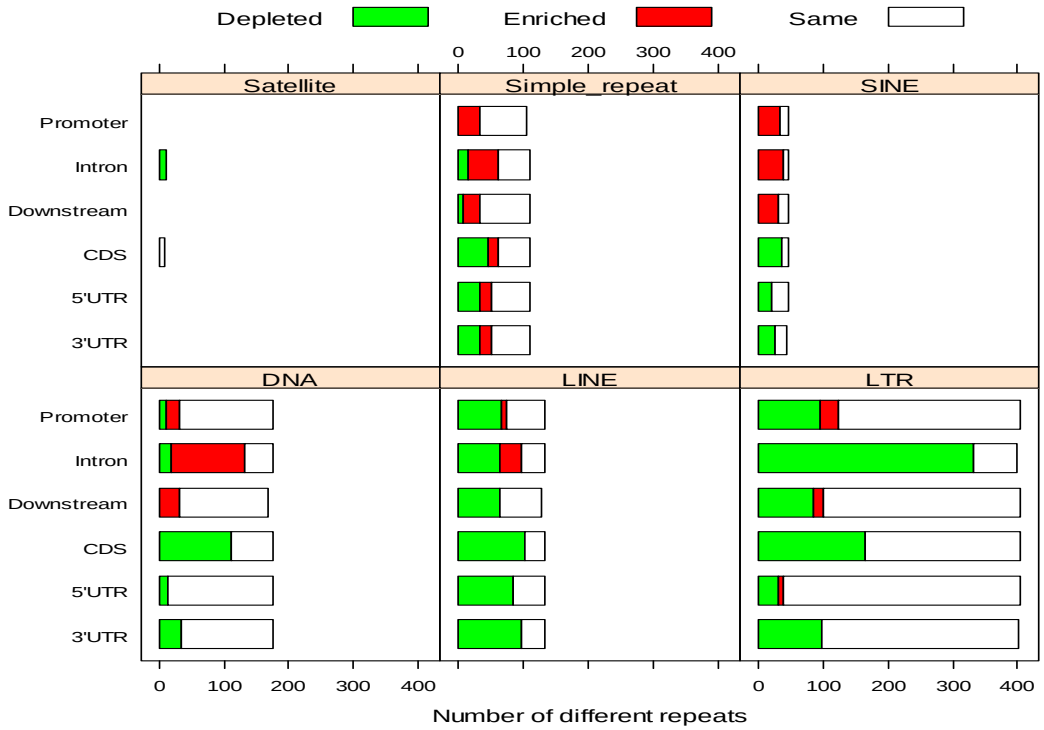

Figure 35. Distribution of repetitive DNA in different genomic regions for the human.

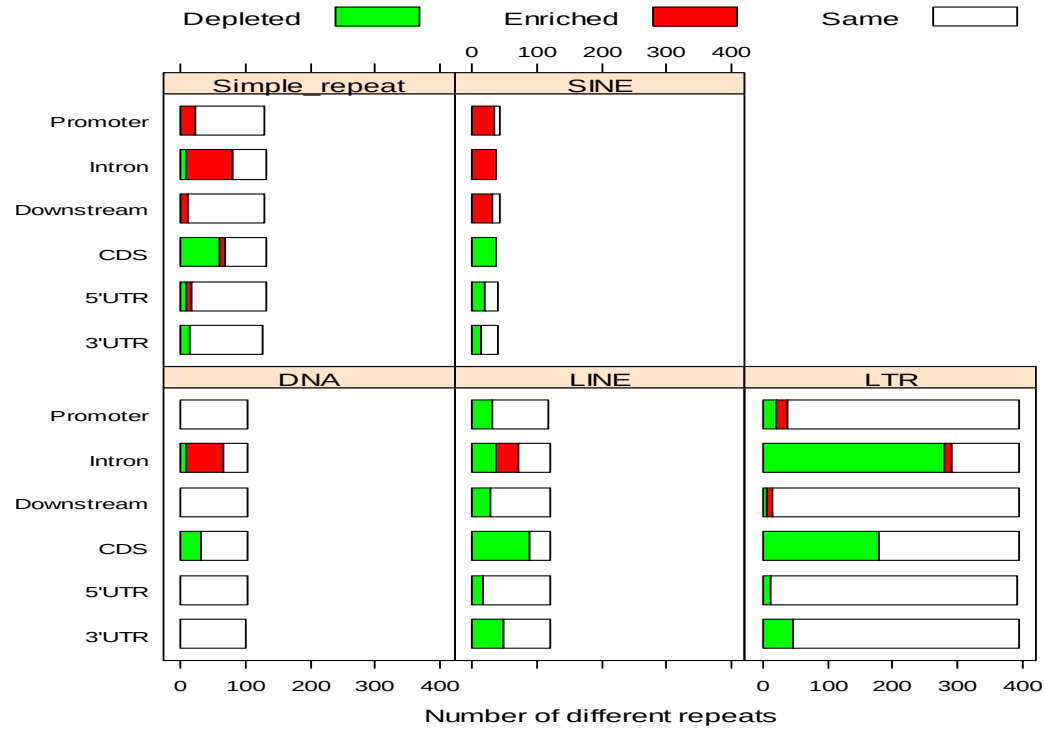

Figure 36. Distribution of repetitive DNA in different genomic regions for the rat.

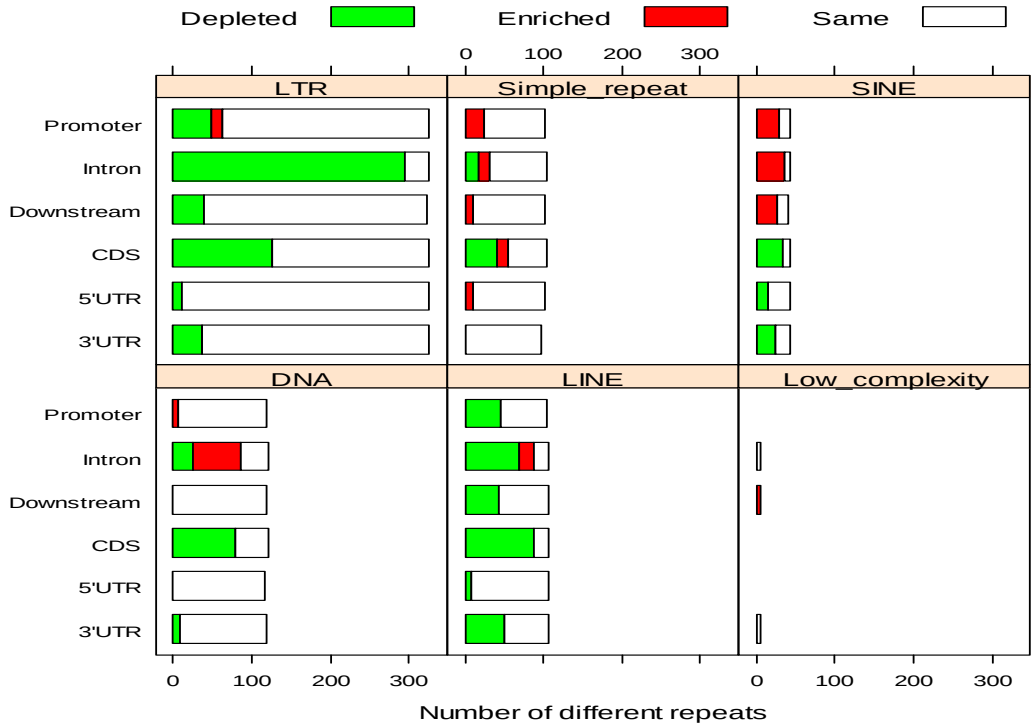

Figure 37. Distribution of repetitive DNA in different genomic regions for the rhesus.

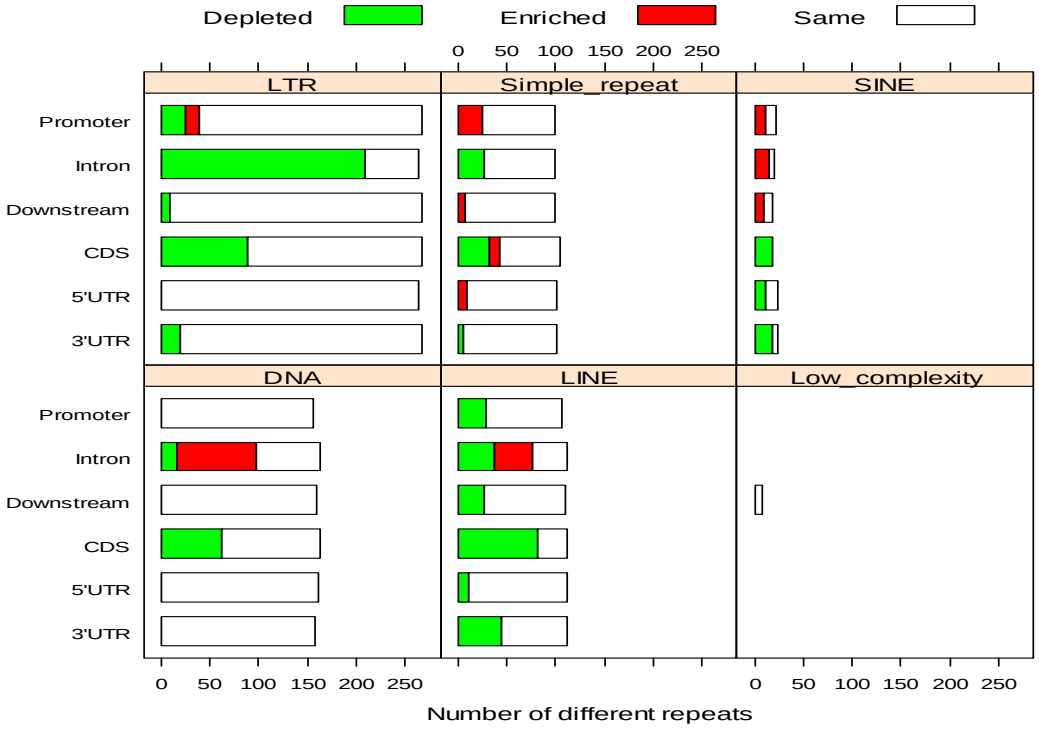

Figure 38. Distribution of repetitive DNA in different genomic regions for the cow.

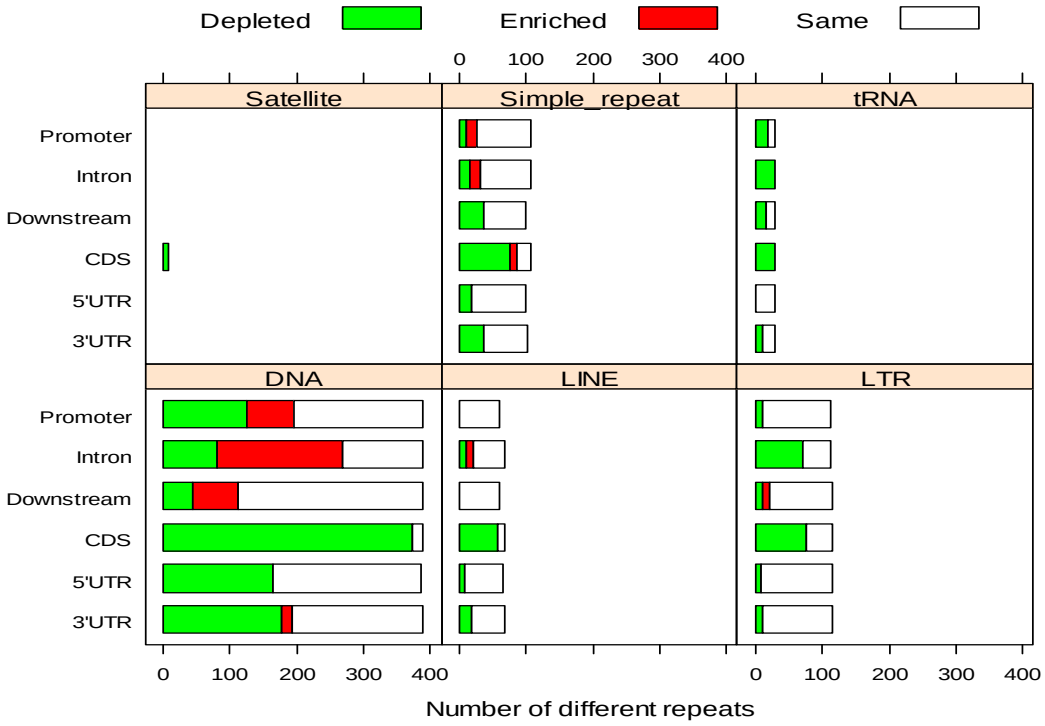

Figure 39. Distribution of repetitive DNA in different genomic regions for the zebrafish.

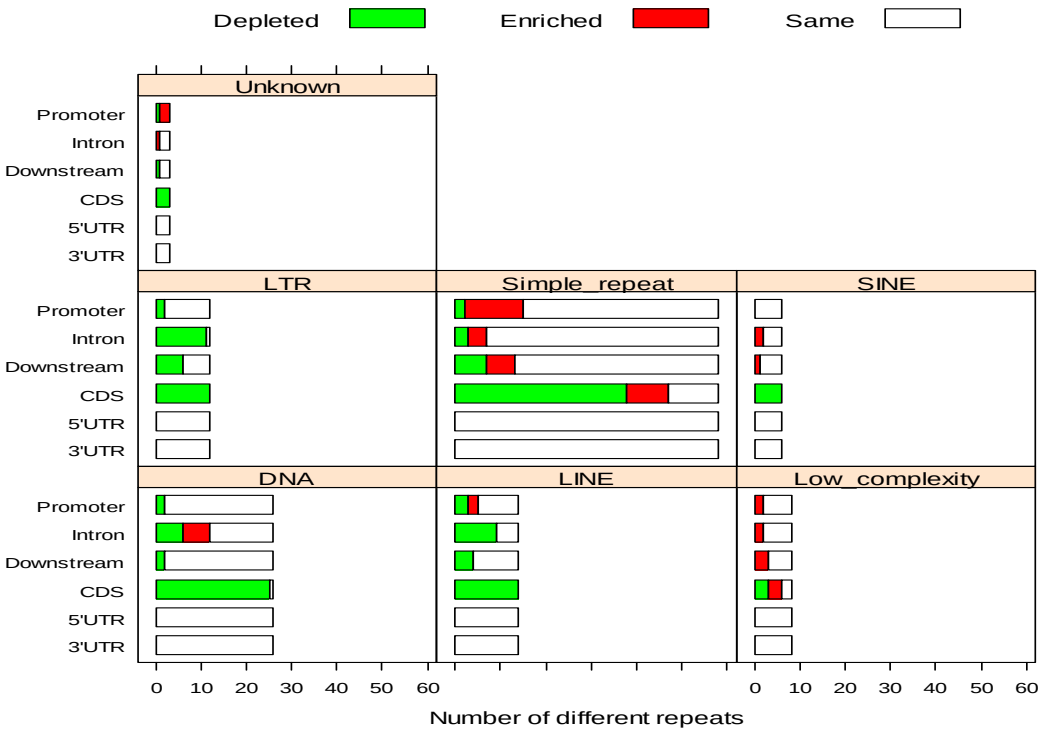

Figure 40. Distribution of repetitive DNA in different genomic regions for the Fugu.

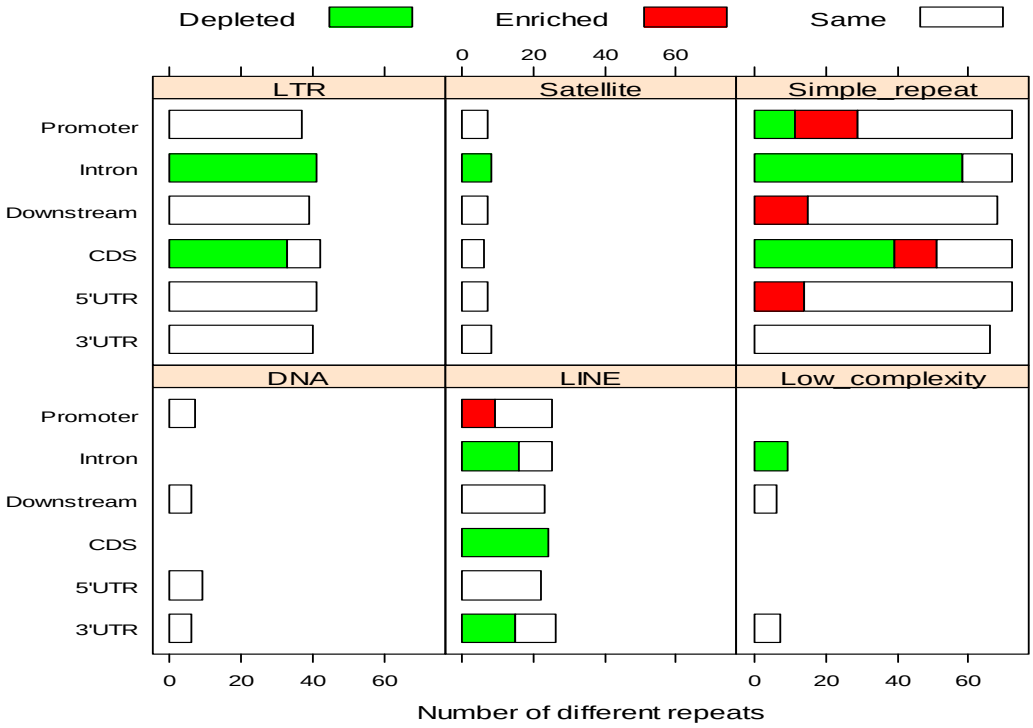

Figure 41. Distribution of repetitive DNA in different genomic regions for the chicken.

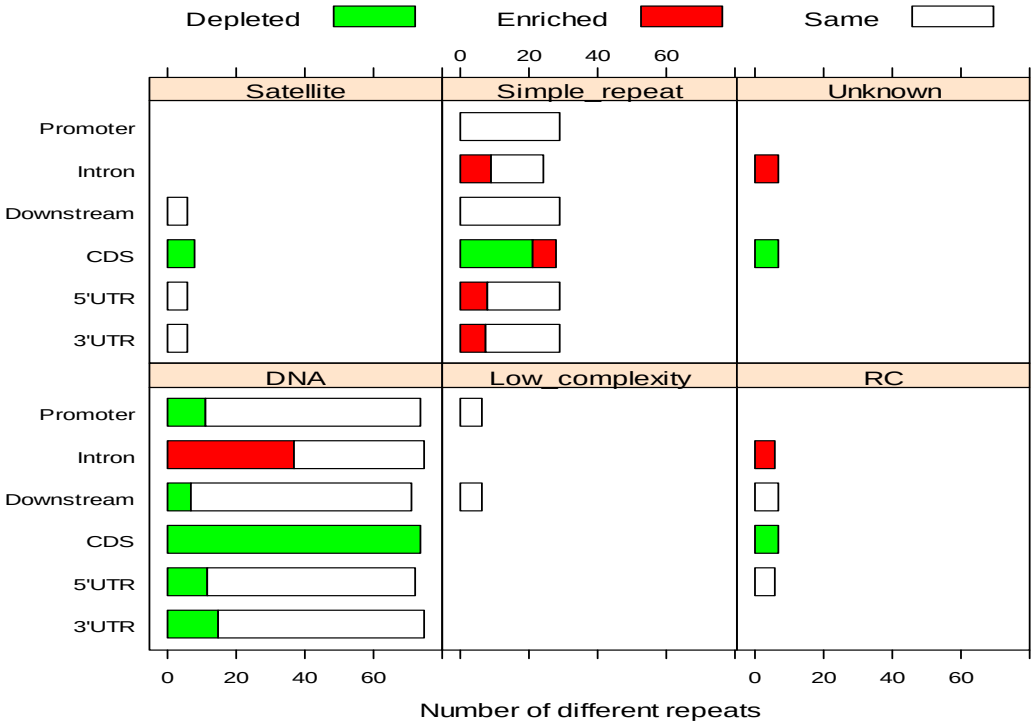

Figure 42. Distribution of repetitive DNA in different genomic regions for the C. elegans.

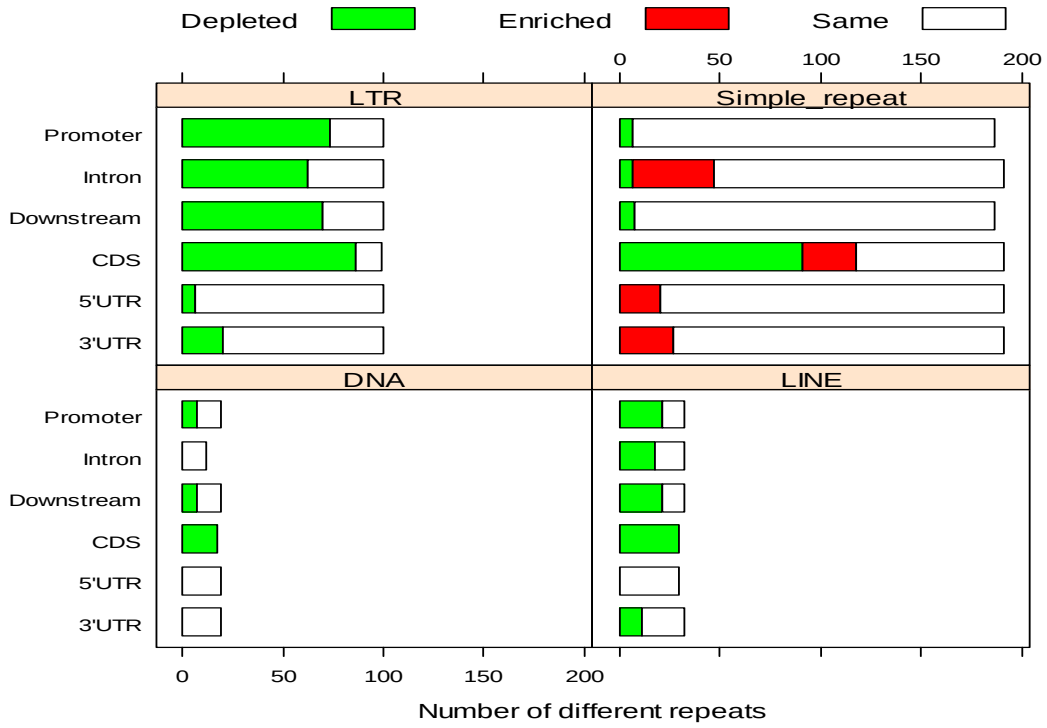

Figure 43. Distribution of repetitive DNA in different genomic regions for the fruit fly.

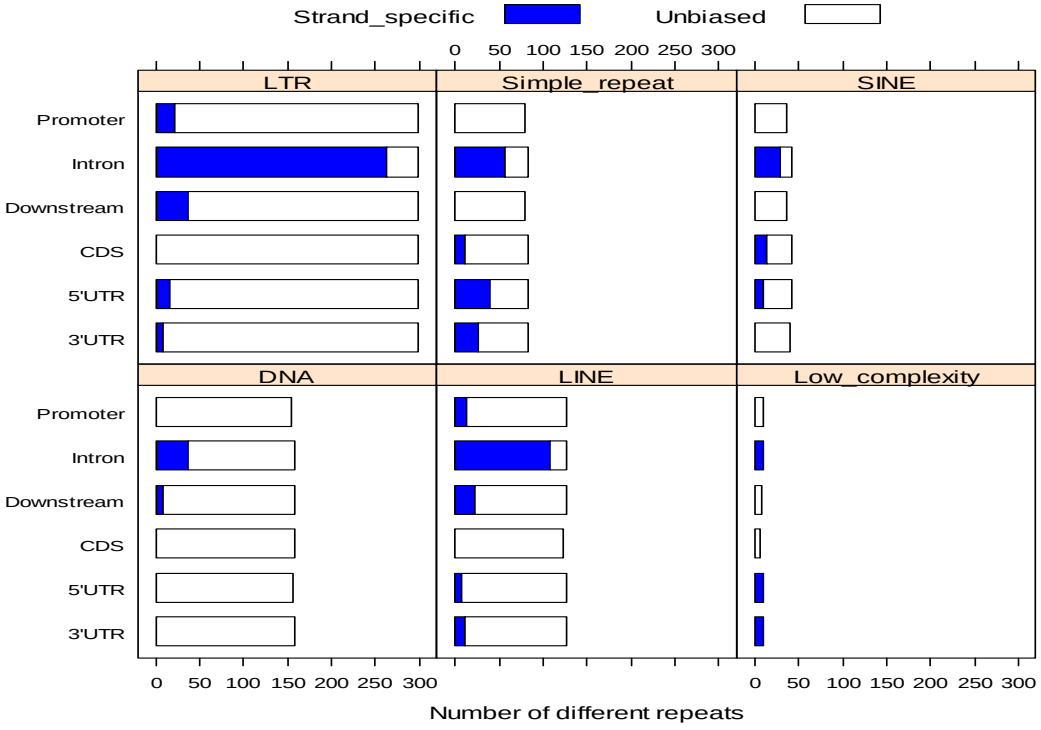

Figure 44. Strand-specificity of the repetitive DNA different genomic regions in the human genome.

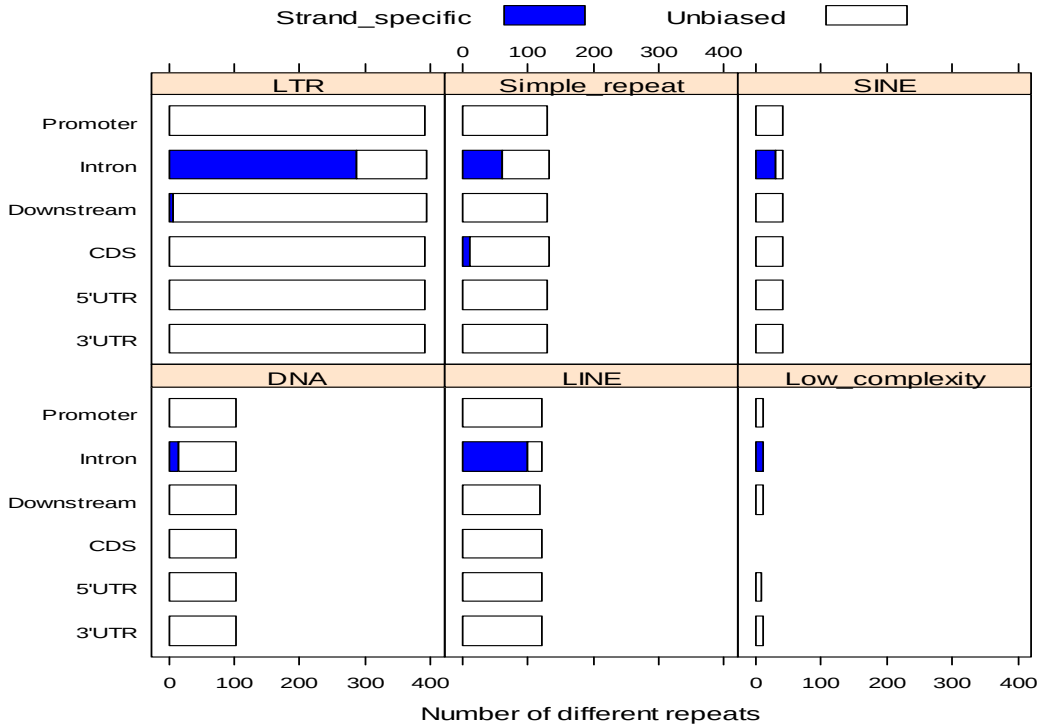

Figure 45. Strand-specificity of the repetitive DNA different genomic regions in the rat genome.

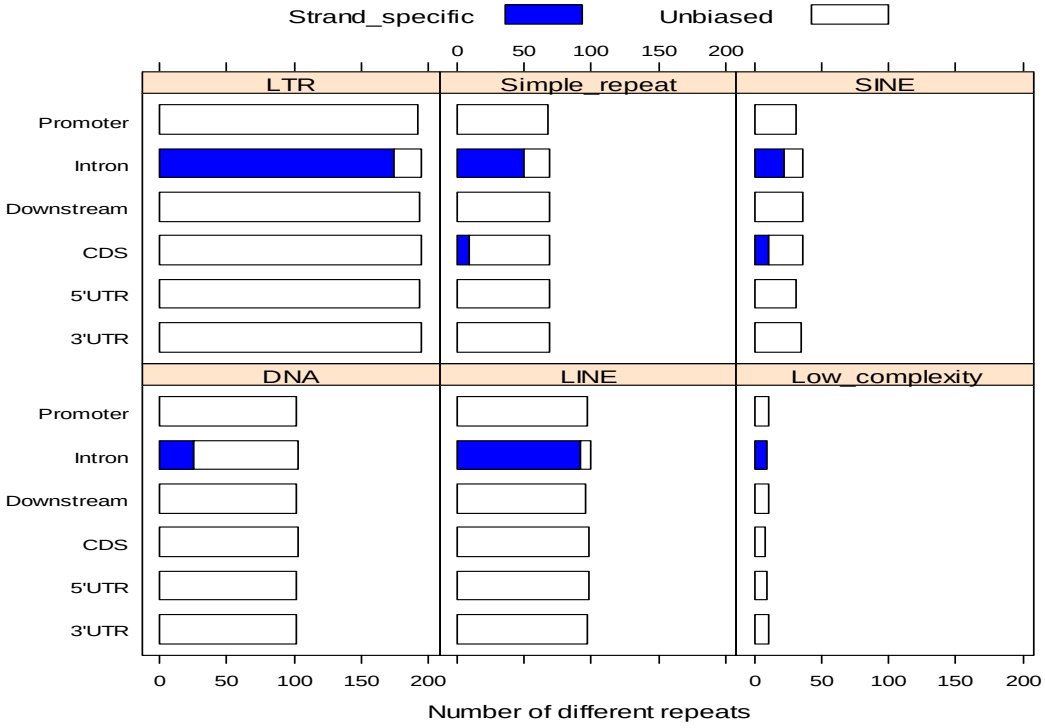

Figure 46. Strand-specificity of the repetitive DNA different genomic regions in the rhesus genome.

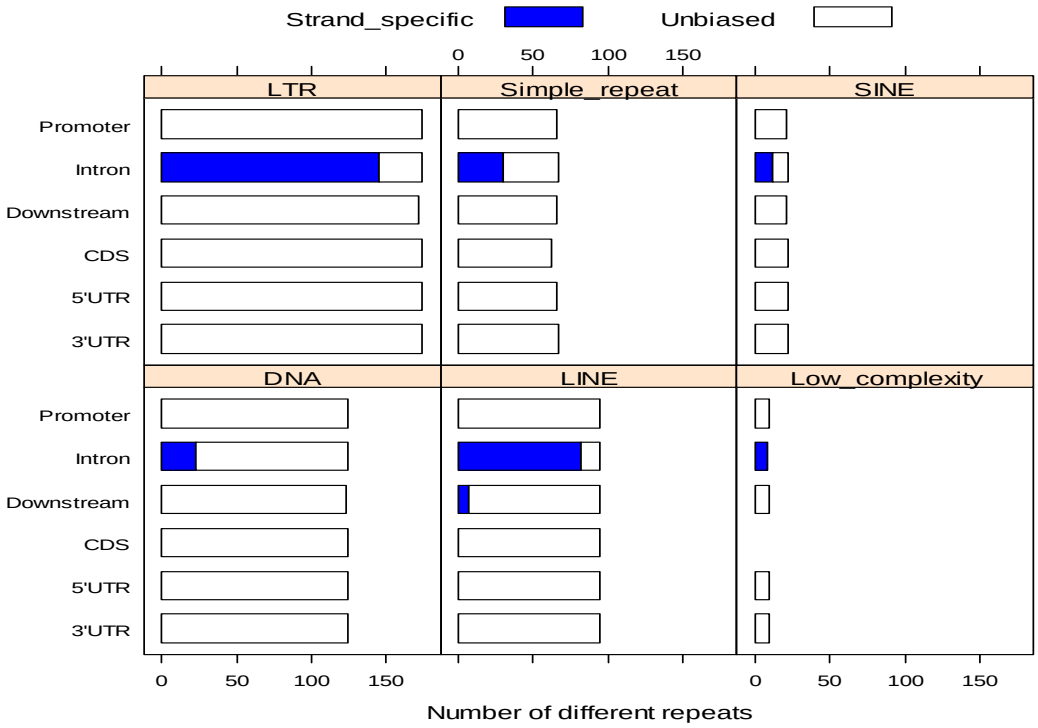

Figure 47. Strand-specificity of the repetitive DNA different genomic regions in the cow genome.

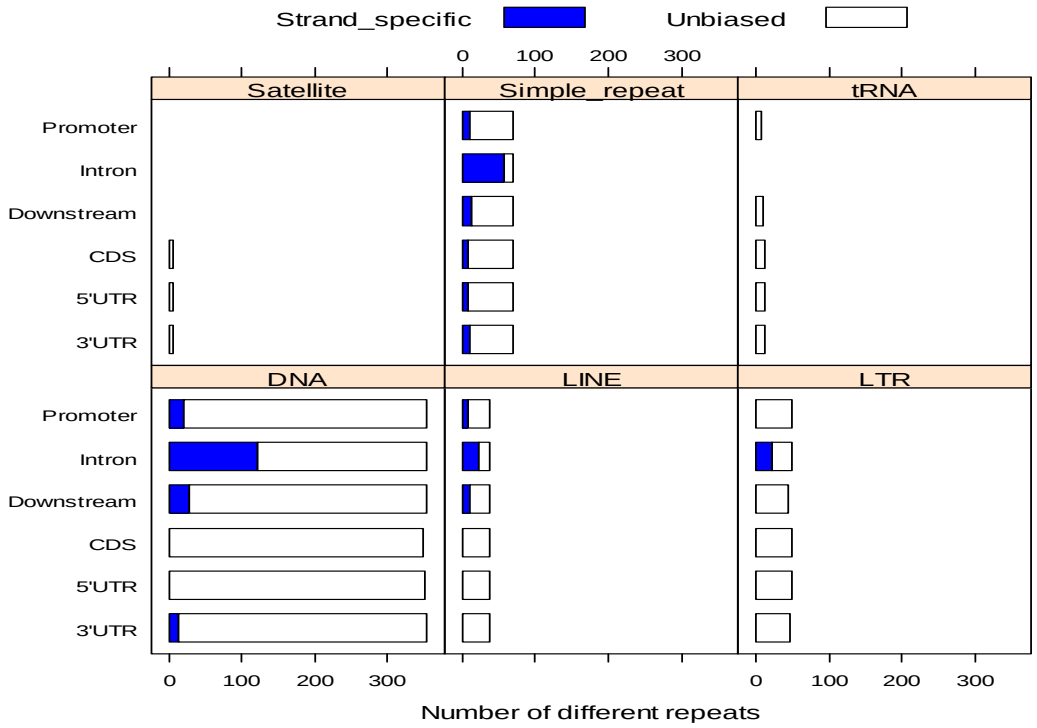

Figure 48. Strand-specificity of the repetitive DNA different genomic regions in the zebrafish genome.

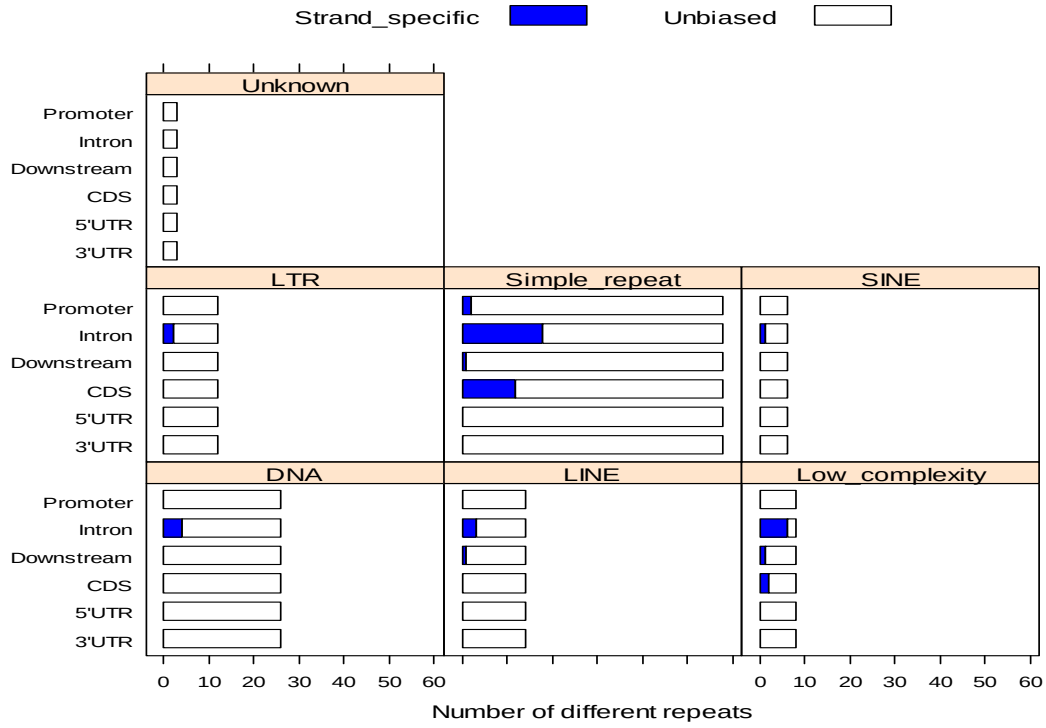

Figure 49. Strand-specificity of the repetitive DNA different genomic regions in the Fugu genome.

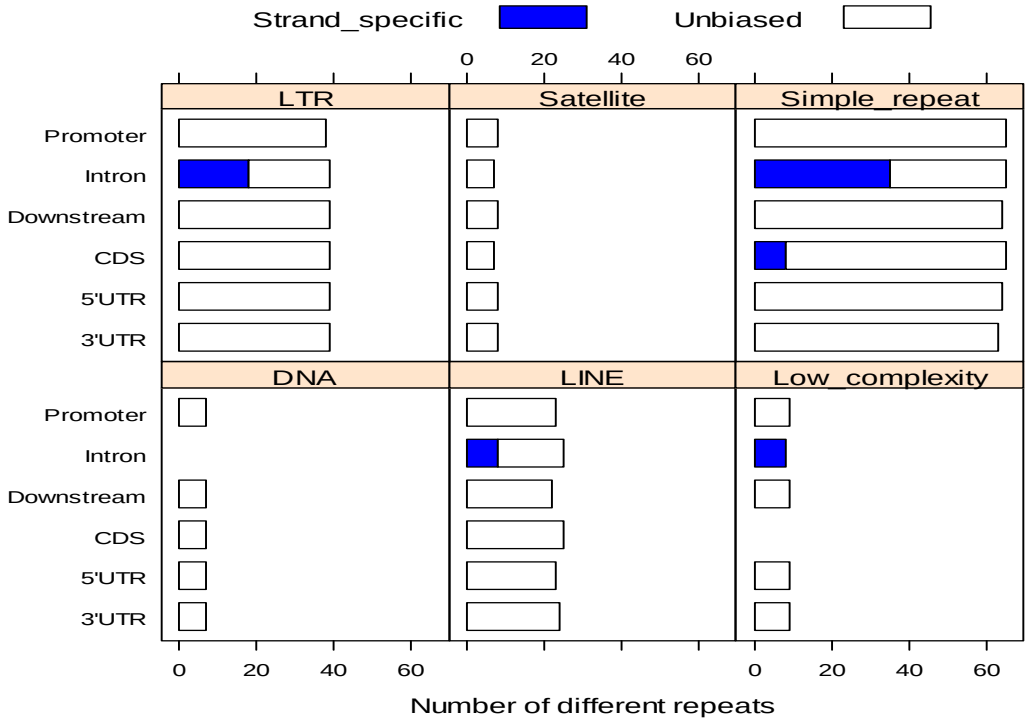

Figure 50. Strand-specificity of the repetitive DNA different genomic regions in the chicken genome.

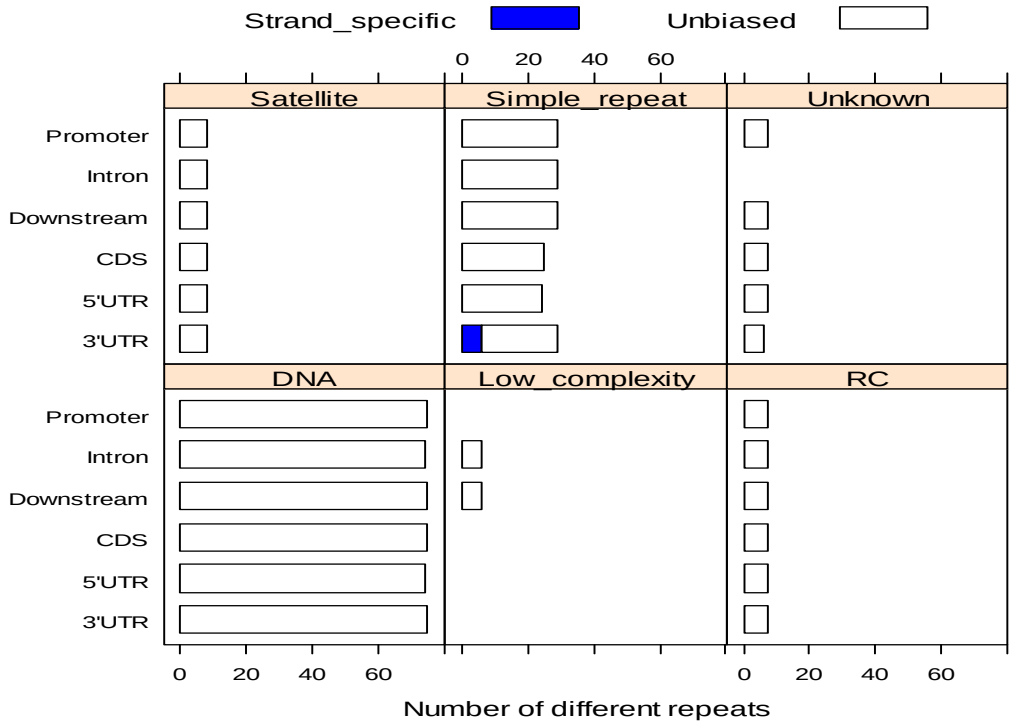

Figure 51. Strand-specificity of the repetitive DNA different genomic regions in the C. elegans genome.

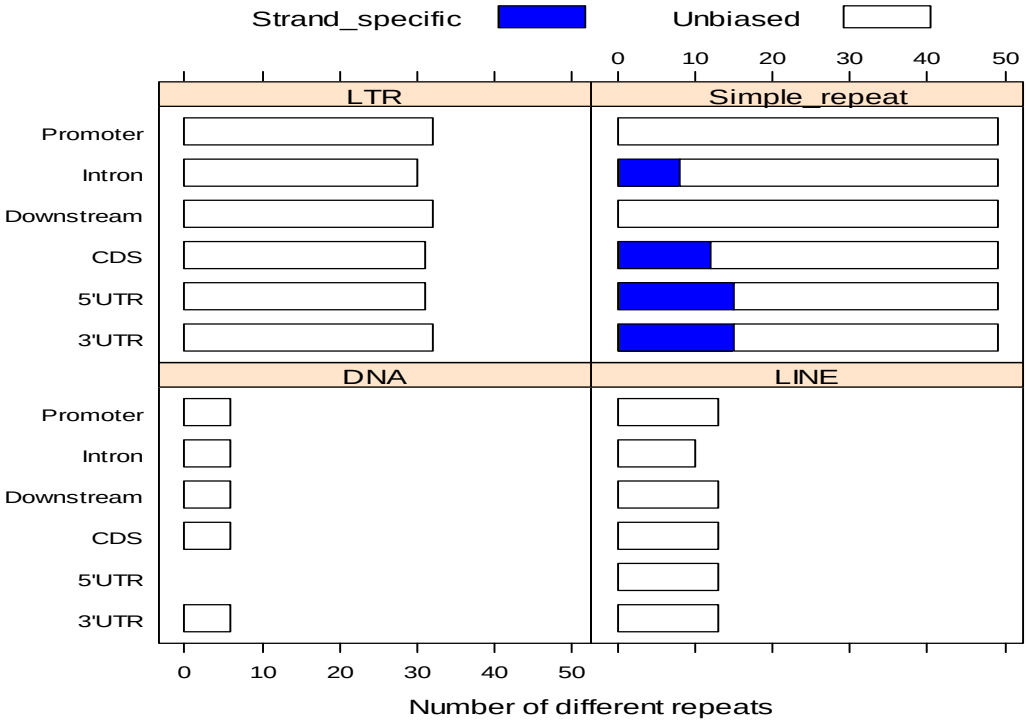

Figure 52. Strand-specificity of the repetitive DNA different genomic regions in the fruit fly genome.