Analyses of genomic changes associated with rodent evolution

Babarinde Isaac Adeyemi

Doctor of Philosophy

Department of Genetics

School of Life Science

SOKENDAI (The Graduate University for Advanced Studies)

2016

Acknowledgements

This study would not have been started, let alone completed, without the supports from my thesis adviser, Professor Saitou Naruya. His moral supports and encouragement cannot be overemphasized. He provided a wonderful environment in which I could develop from a young boy with zero evolutionary and computational knowledge to the point I am now. I could not have asked for a better supervisor. Past and current members of the Division of Population Genetics (Saitou laboratory), National Institute of Genetics have also contributed immensely to the progress of the study. The guidance, comments and suggestions during progress reports by Drs. Tim Jinam, Sumiyama Kenta, Kryukov Kirill, Takahashi Mahoko, Matsunami Masatoshi and Kanzawa Hideaki are well appreciated. I also thank Ngai Jeff, Hettiarachchi Nilmini, Mahmoudi Saber and Shokait Shayire for their supports. Mrs. Mizuguchi Masako, Iida Ai, Noaki Yoshimi and late (Mrs.) Kawamoto Tatsuko's support, patience and tolerance are well appreciated.

Some parts of the study were collaborative studies that involved scientists from across Japan. The large Japanese wood mouse tissue samples were harvested and supplied by Dr. Tomozawa Morihiko. Japanese giant flying squirrel sample was provided by Professor Oshida Tatuso of Obihiro University of Agriculture and Veterinary Medicine. The capybara tissue sample was donated by Izu Shaboten Park, Izu, Japan. Japanese giant flying squirrel genome sequencing project was in collaboration with Dr. Takeda Tadayuki of RIKEN, Yokohama, Dr. Ikeo Kazuho and Mr. Monma Norikazu of National Institute of Genetics. Capybara genome sequencing was done in Prof Inoue Ituro’s lab with valuable comments and guidance from Drs. Hosomichi Kazuyoshi and Nakaoka Hirofumi. Many of the computationally expensive analyses were done on NIG super computer.

My profound gratitude also goes to my progress committee members, Profs Shiroishi Toshihiko, Fujiyama Asao, Kitano Jun, Drs. Nozawa Masafumi and Koide Tsuyoshi. They not only gave wonderful and productive comments, they also assisted in the actual design and execution of some part of the experiments. I am indeed grateful. I also thank Prof Jianzhi Zhang for hosting me during the short study abroad program.

My stay and study in Japan would not have been possible if not for the generosity and the scholarship offered by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. During my study, I also received the support for Short-Stay Study Abroad Program offered by SOKENDAI (Graduate University for Advanced Studies, Japan) for a brief stay at University of Michigan, Ann Arbor.

The supports, encouragement and sacrifices from my family cannot be overemphasized. To my dad and mom, Emmanuel A. and Mary B. Babarinde, I say thank you. I would also like to appreciate my siblings, Grace A. Oyetoro, Elizabeth O. Ogunniyi, Samuel A. Babarinde, Deborah A. Omotunde, John A. Babarinde, Amos A. Babarinde and their families. I am so lucky to have you as brothers and sisters. In addition, my host family, Mr. and Mrs. Fujinami have made Japan a home-away-from-home for me.

Contents

List of Figures

List of Tables

List of Abbreviations

CGN: Conserved Genomic Neighborhood

CNS: Conserved Noncoding Sequence

DAF: Derived Allele Frequency

dN: Nonsynonymous substitution distance

dS: Synonymous substitution distance

GEO: Gene Expression Omnibus

GO: Gene Ontology

modCGN: Modified Conserved Genomic Neighborhood

NJ: Neighbor-Joining

RDD: Relative Distance Difference

RPKM: Reads Per Kilobase per Million mapped reads

Abstract

Rodents are the mammalian order generally characterized by continuously growing pair of incisors. They have the highest number of species and an unrivalled ecological success among mammals. Rodents are found in many habitats and have huge phenotypic diversity across body size, longevity, physiology, morphology, litter size and so on. Despite the phenotypic diversity among rodent species, limited number of genome sequences constrains extensive investigation of genomic changes associated with rodent phenotypic diversity.

Extensive comparative genomics requires the availability of genome sequences. Therefore, I set to sequence the whole genomes of two strategically located species in rodent phylogeny. The first species sequenced is capybara, the largest living rodent species. The second species is Japanese giant flying squirrel. These two species, in addition to mouse, are representatives of the three major rodent lineages. The main aim of this study is to investigate the genomic changes that have accompanied rodent evolution. Based on the long standing hypothesis that regulatory evolution might have contributed to morphological evolution, I decided to investigate the role of conserved noncoding sequence (CNS) evolution in rodent phenotypic diversity. In this study, I found that CNSs often have regulatory signals and that CNSs are probably involved in more conserved expression of the flanking protein-coding genes. In addition, I found that the the physical distance between CNS and the closest protein-coding gene tend to be evolutionarily conserved. Furthermore, I found that CNS-associated genes tend to be enriched in transcription, development and nervous system, but underrepresented in genes associated with defense and immunity. As expected from gene ontology enrichment, CNS-associated genes are highly expressed in embryonic brain and poorly expressed in testis. These observations suggest the importance of CNS evolution in phenotypic diversity as the major players of morphological evolution should be active during developmental stage.

The evolutionary dynamics of CNS across four mammalian orders and evolutionary ages revealed interesting features of rodents. I found that CNS loss is highest in rodents but lowest in primates. On the contrary, CNS gain is highest in primates but lowest in rodents. Carnivores and cetartiodactyls have intermediate values. These observations highlight the high turnover rate of rodent CNSs, and by extension, regulatory elements. Focusing on primate CNSs due to limited rodent data; I found that more ancestral CNSs tend to have stronger constraints than the recently evolved ones. These results suggest that rodent phenotypic diversity might be connected to the high regulatory turnover in rodents. Even among rodents, CNS evolution was found to be heterogeneous with the highest loss in mouse-related lineage and the lowest in squirrel-related lineage.

To further investigate these observations, I asked how quickly regulatory turnover happen in rodents. For this analysis, I focus on *Muridae*, the largest rodent and mammalian family in terms of the number of species. The transcriptome data of mouse and rat are already available. In addition, I sequenced the transcriptome of 3-, 5- and 7-day postnatal wild large Japanese wood mouse species for the retrieval of large number of transcripts. Liver transcriptome analyses reveal that expression dynamics highlight phylogenetic relationships, and that CNSs seem to be less active in liver.

Mutation-driven hypothesis of evolution suggests that species with relatively high mutation rate would have higher ecological success rate because they would have abundant “raw materials” upon which evolutionary forces can act. Interestingly, previous studies and my own analyses show that rodents have higher evolutionary rates. I then asked if the high evolutionary rates in rodents have contributed to their phenotypic diversity and ecological success. If this is true, I hypothesized that rodent clades with higher number of species would have higher evolutionary rate. Using the published rodent genome data, I first show that though primates’ evolutionary rates are lower than rodents’, there is heterogeneity among rodents. Of particular interest is the observation that murids with the highest number of species indeed have the highest evolutionary rates. Unlike the previous reports however, I found limited evidence supporting the correlation between body size and evolutionary rate.

In conclusion, my study suggests that higher regulatory evolution in rodents might have contributed to rodent phenotypic diversity. This higher regulatory evolution is brought about by the high evolutionary rates among rodents. These results support mutation-driven hypothesis of evolution. The whole genome sequences of Japanese giant flying squirrel and capybara, with the transcriptome data of large Japanese wood mouse will be invaluable resources for future rodent evolutionary studies. The next line of analyses would be trait- and region- targeted to establish the genomic regions responsible for some interesting rodent phenotypes.

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Rodent diversity

Rodents are the mammalian species characterized by their continuously growing two pairs of incisors. The continuously growing incisors must be kept by gnawing (Nowak 1999). The gnawing not only keeps the incisors short, it also makes the incisors sharp. Among all mammalian orders, rodents are the most ecologically successful (Figure 1.1A). The distribution can be contrasted to that of non-human primates (Figure 1.1B). The diversity spans through the number of species, body size, morphological features, ecological success and lifespan. There are about 3,000 species of rodents, representing close to half of all the mammalian species (Wilson and Reeder, 2005). Although the actual number of individuals is not known, the numbers of rodent species, genera and families are unrivalled among mammalian lineages (Kay and Hoekstra 2007). The uniqueness can be appreciated when compared to primates (Table 1.1).

Their relatively small body size might have contributed to their migration with humans. However, they should have the genetic ability to quickly adapt to a new habitat. Apart from the wide-spread rodent species, there are also many rodent species that are endogenous to certain geographical location. An example is capybara which is found exclusively in South America.

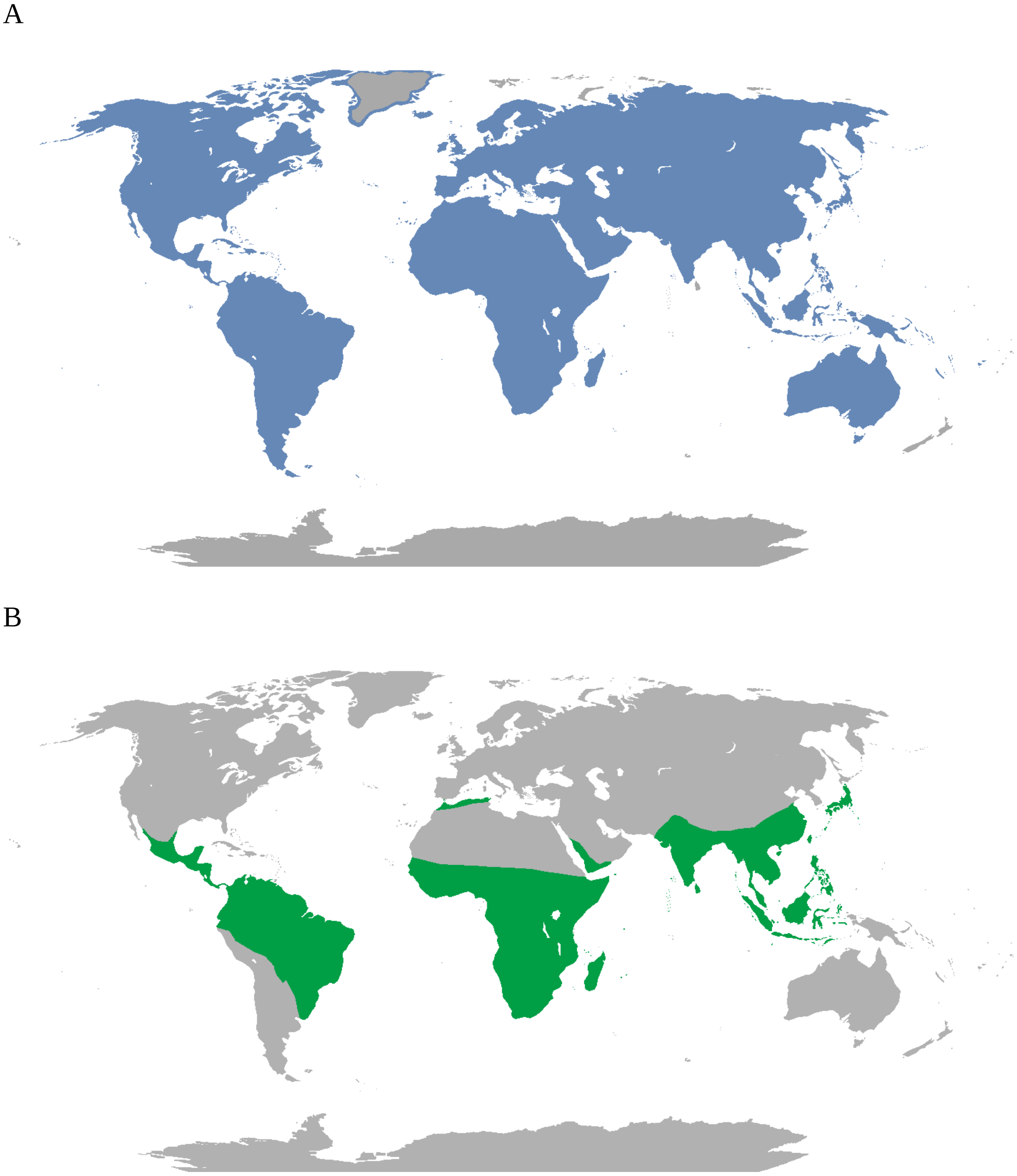


Figure 1.1: The distribution of rodent species is widespread. (A) The distribution of rodent species in the world. The blue areas represent regions where rodent species are found. [Source: https://commons.wikimedia.org/wiki/File:Rodent_range.png]. (B) Distribution of non-human Primates. Non-human primate species are found in green regions. [Source: https://commons.wikimedia.org/wiki/File:Range_of_Non-human_Primates.png].

Table 1.1: Comparison of selected features of rodents and primates

	Primates	Rodents
Number of species	300[a]	3,000[b]
Number of genera	71[a]	426[a]
Smallest extant	Pygmy marmoset (100g)[b]	Pygmy Jerboa (3.75 g)[c]
Largest extant	Gorilla (up to 350kg)[b]	Capybara (up to 81kg)[d]
Largest extinct	*Gigantopithecus blacki* (550kg)[e]	*Josephoartigasia monesi* (1000kg)[d]

[a] Nowak (1999)

[b] Wilson and Reeder (2005)

[c] Roberts (2006)

[d] Ferraz et al. (2005)

[e] Ciochon et al. (1990)

[f] Rinderknecht and Blanco (2008)

Large Japanese wood mouse is endemic to Japan, but is found in almost all Japanese islands. Also, Japanese giant flying squirrel is native to Japan. Interestingly, even for the rodent species that are endemic to certain geographical location, some are not threatened because they have quickly adapted to the environment in which they evolved.

Rodent diversity is not only restricted to the number of species and ecological success, they also have a huge morphological diversity. Although rodents are generally thought to be small-bodied mammals, some can weigh several tens of kilograms. Capybara, the largest extant rodent species, can weigh up to 81kg (Ferraz et al. 2005). This contradicts the body sizes of most rodent species which are usually less than 1kg. In fact, the adult female of the smallest rodent species weigh only 3.75g. In addition to the body size, rodent physiology has adapted to different ecological habitat. For example, although mammals are known to be homoeothermic (warm-blooded), naked mole rat is the only mammal that is currently thought to be poikilothermic or cold-blooded (Buffenstein and Yahav 1991). Some species, like flying squirrels, have the ability to glide. Others, like blind mole rat, have complex social structure. Therefore, rodent diversity is not only in the number of species, or the ecological distribution, it also spans through the physiological, morphological and behavioral traits.

Based on the position of masticatory muscles (the masseters), taxonomic studies have classified rodents into three major clades; (i) Castorimorpha, Anomaluromorpha and Myomorpha (mouse-related), (ii) Sciuromorpha (squirrel-related), and (iii) Hystricomorpha (guinea pig-related). This taxonomic classification has been supported by phylogenetic analyses

(Blanga-Kanfi et al. 2009). The phylogenetic relationship of the rodent species used in this study is presented in Figure 1.2. The phylogenetic tree presented in the figure was computed using amino acid sequences and the topological relationships were supported with more than 90% bootstrap values.

1.2 Identifying the genomic changes underlying phenotypic changes

Identifying the genomic changes responsible for certain phenotypic traits have been an important goal for many geneticists. According to Hitoshi Kihara (1946), "The history of the earth is recorded in the layers of its crust; the history of all organisms is inscribed in the chromosomes" (see Saitou 2014). Because there are thousands of protein-coding genes and probably hundreds of thousands of regulatory regions, linking particular genomic changes to certain phenotype remains a challenging task in genetics. An important question would be which genomic regions to focus on. Next, the appropriate approach to adopt would be decided.

In forward genetics approach, the phenotype of interest is first identified. Using a combination of comparative and statistical analyses, the genomic changes responsible for such phenotype is then identified. Because there are so many phenotypes and mutations in the natural population, there is an abundance of resources for the genetic studies. Unfortunately, these abundant resources make the linking of phenotype to genomic changes more complicated. Alternatively, the consequences of certain genomic changes can be investigated using reverse genetics. In this case, the phenotypic change is studied after inducing some genomic changes.

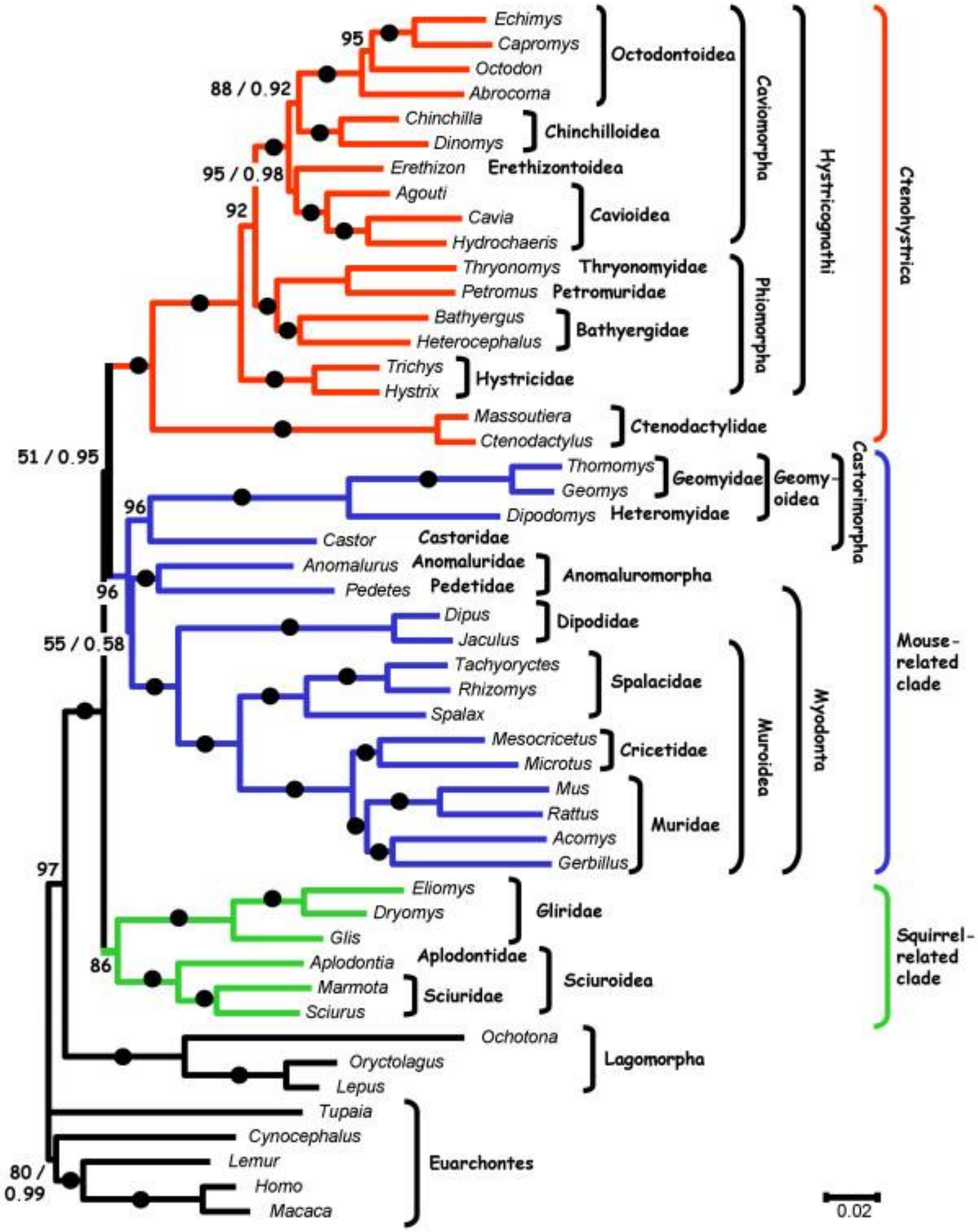


Figure 1.2: The phylogenetic relationships of selected rodent species. The maximum likelihood tree was made from six nuclear gene fragments. Capybara (*Hydrochaeris*) and guinea pig (*Cavia*) are shown to cluster. However, Japanese giant flying squirrel (Sciuridae family) and large Japanese wood mouse (Muridae family) are not represented in the tree. [Source: Blanga-Kanfi et al. (2009)].

Generally, when investigating the genomic changes responsible for certain phenotypic changes, three genomic changes can be investigated. The first and the most often investigated regions are the coding regions. Because most genes function at protein-level, changes in the protein coded by a gene may interrupt the function of the protein. The second regions are the regulatory regions. Regulatory changes have been reported to contribute to morphological evolution (Carroll 2008). The third type of change is the epigenetic changes. Recently, epigenetic changes are attracting the attention of many researchers.

1.3 Protein-coding evolution

Change in protein-coding genes is the most investigated type of genomic changes. The central dogma of DNA illustrates the flow of genetic information from DNA to protein via messenger RNA. Correct sequences of proteins are important for proper function. Therefore, if there is any type of amino acid disrupting mutations (nonsynonymous mutation) in the DNA sequences of the protein-coding genes, the function may be impaired. Based on the effect on protein product, mutations on protein-coding genes can be synonymous (no change of the amino acid sequences), missense (resulting in the change of the coded amino acid) or nonsense (resulting in the truncation of protein due to the premature stop codon).

Although protein-coding genes represent a tiny proportion (~2%) of mammalian genomes, they are evolutionarily very stable. The evolutionary stability is because of their functional importance. For example, about 66% of mutations in known mouse genes result in lethal phenotypes (Dickerson et al. 2011). This implies that many mutations on protein-coding genes are quickly removed, and thus may not contribute substantially to evolutionary changes. This stability of protein-coding genes was also observed between human and chimpanzee (King and Wilson 1975), and this hinted the importance of gene regulatory sequence in morphological evolution.

1.4 Regulatory evolution as a major contributor to phenotypic changes

The genomic revolution has made it possible to study wide range of genomic regions. In addition, with the technical and technological advancement in experimental approaches, genomic changes at many levels can be investigated. Specifically, it is now possible to investigate the regulatory changes. Based on the amount of support available, Carroll (2008) proposed a genetic theory of morphological evolution which states that form evolves largely by altering the expression of functionally conserved proteins and such alterations occur largely in the cis-regulatory regions. The theory establishes the previously suggested hypothesis about the contribution of regulatory changes to morphological evolution.

Although Carroll’s theory focuses more on cis-regulatory changes, gene expression regulation could be achieved at three different levels; (i) DNA (transcription) level, (ii) RNA level, and (iii) protein level. Regulation at DNA level involves the proper control of the transcription factors and the transcription machinery in achieving the optimal level of expression. The major DNA players are the promoters, enhancers, silencers and insulators. These bind to specific transcription factors and transcription machinery to achieve specific functions. Regulation can be positive or negative. Regulation at transcription level is probably the most economical regulation. The next level of regulation is at RNA level. The level of RNA can be controlled by some machinery like microRNA. Regulation of translational speed may be another form of control at RNA level. The last level is the protein layer. This regulation can be achieved by controlling the translational rate and protein stability.

Across many animal lineages, increasing line of evidences is supporting the importance of regulatory changes in morphological evolution across organisms. In *Drosophila melanogaster* for example, ectopic eyes could be induced by targeted expression of *eyeless* gene (Halder 1995). Also, loss of particular trichomes has been attributed to regulatory change of *shavenbaby/ovo* gene (Sucena et al. 2003). In sticklebacks, recurrent deletion of a *Pitx1* enhancer has been reported to be associated with pelvic reduction (Chan et al 2009). In mouse, *shh* enhancer has been linked to polydactyly (Lettice et al. 2003; Sagai et al. 2004). Furthermore, modification of mammalian limb lengths has been associated to regulatory divergence (Cretekos et al. 2008). It is interesting to note that morphological and physiological changes might have different molecular bases. While regulatory evolution might be more associated morphological changes, coding sequence evolution may be more important in physiological changes (Liao et al. 2010).

1.5 Conserved noncoding sequences as regulatory signatures

Unlike protein-coding genes which are well annotated, regulatory elements are not well annotated. One obvious reason is because they are not expressed. They are usually bound by transcription factors and machinery. Importantly, they are mostly located in the non-coding part of the genome. Therefore, any functional region in the noncoding region of the genome might be a regulatory element. Classically, functional genomic regions are identified by sequence constraint (Miyata et al. 1980; Loots et al. 2000; Woolfe et al. 2005). Any noncoding region with sequence constraint (conserved noncoding region or CNS), is therefore a potential regulatory element.

An important task is to correctly identify regions under evolutionary constraint. Most algorithms for identifying evolutionary constraint equate sequence similarity to sequence constraint. That is, any sequences found to be similar across species are thought to be under sequence constraint. However, regions that are in mutational cold-spots may also be found to be similar, even if they do not have any functional constraint. To investigate whether CNSs are just mutational cold-spot or not, Drake et al. (2006) and Takahashi and Saitou (2012) conducted derived allele frequency (DAF) spectrum analyses, and found that CNSs are under purifying selection and are not mutational cold-spots. However, the amount of retrieved CNSs depends critically on the computational thresholds, DAF analyses should be done to ascertain sequence constraint.

Being able to confirm that CNSs are under sequence constraint does not necessarily imply their activity as regulatory elements. For example, recent studies are emphasizing the importance of noncoding RNAs (Wilusz et al. 2009). Thus, there is a possibility that a noncoding region functions not as regulatory element, but as a noncoding RNA. Therefore, it is important to investigate if the CNSs actually have regulatory activities. Because regulatory elements function via looping (Ong and Corces 2009), techniques like chromosome conformation capture, also called 3C (Dekker et al. 2002), circularized chromosome conformation capture, also called 4C (Zhao et al. 2006) and carbon copy chromosome conformation capture, also called 5C (Dostie and Dekker 2007), which are able to identify interacting DNA regions, are of great importance. In addition, gene regulations are brought about by the interaction of transcription factors and machinery with the regulatory DNA regions. Techniques that can identify DNA-protein interaction are invaluable in proper identification of regulatory noncoding regions. Chromatin immunoprecipitation and massively parallel sequencing, ChIP-seq (Robertson et al. 2007) is one such technique. Extensive study by Viesel et al. (2009) shows that ChIP-seq analyses can accurately predict enhancer activity, and that CNSs are regulatory elements. Further integrative analyses by Hemberg et al. (2012) suggest that most CNSs are regulatory elements and not noncoding RNAs.

The strong evolutionary constraints and the functional importance of CNSs suggest that the deletion of the regions would give some obvious phenotypes. Indeed, that was observed in several studies (Loots et al. 2000; Lettice et al. 2003; Sagai et al. 2004). However, Nóbrega et al. (2004) and Ahituv et al. (2007) reported a surprising observation. The deletion of mouse genomic regions containing CNSs did not give any obvious phenotype. These two reports challenged the functionality of CNSs. The lack or obscurity of phenotypic change does not necessarily imply functional insignificance. Since the mice were raised under laboratory conditions, the selection force might not be in full operation. This possibility is supported by the Drosophila study of Frankel et al. (2010). In their study, the deletion of a *shavenbaby* regulatory element did not give any obvious phenotype under normal condition. In a stress condition however, obvious non-wildtype phenotypes were produced. Generally, most CNSs are likely to be functionally important as regulatory elements.

1.6 Goals of the study

The role of regulatory elements in morphological evolution has been introduced. If CNSs can be used as signatures for regulatory elements, then CNSs might be used to investigate the phenotypic diversity observed among many species. Particularly, I am interested in understanding rodent phenotypic diversity from the molecular evolution point of view. The evolutionary analyses were conducted using genome and expression data. This requires the generation of new genome and transcriptome data.

The specific goals of this study include;

1. Generating new genome and transcriptome sequence data for selected rodent species to improve rodent evolutionary study.
2. Investigating the roles of CNS as regulatory elements in conserved gene expression

3. Identifying unique features of rodent CNS evolution.
4. Analyzing the expression dynamics within rodent lineage.
5. Highlighting the evolutionary rate dynamics of rodent lineage.

The general aim of the study is to identify the genomic changes that are associated with rodent phenotypic diversity. Specifically, focusing on CNS evolutionary dynamics, I attempt to investigate the contribution of regulatory changes to rodent phenotypic diversity. To achieve these goals, I combined the knowledge of genomics, evolutionary biology, computational biology and certain degree of statistics. The images of Japanese giant flying squirrel, capybara and large Japanese wood mouse are presented in Figures A1.1, A1.2 and A1.3, respectively in Appendix 1.

1.7 Organization of the dissertation

The dissertation is divided into seven chapters. *Chapter One* (this chapter) lays the general background to the study. It introduces rodent diversity and CNS evolution and gives the justification to the study. In *Chapter Two*, the whole genome sequencing project of Japanese giant flying squirrel is reported. The genome sequence data is used in some of the subsequent analyses. *Chapter Three* reports extensive analyses using genomic, transcriptome and ChIP-seq data to demonstrate that CNSs are likely to be regulatory elements involved in conserved gene expression. Having demonstrated the involvement of CNSs in the gene regulation, I also compared CNS evolution in rodents and three other mammalian lineages. The unique nature of rodent CNS evolutionary dynamics is presented in *Chapter Four*. In addition, the CNS evolutionary dynamics over timescale is presented in Chapter Four. *Chapter Five* attempts to investigate how quickly gene expressions change by using transcriptome data of selected rodent species, including the newly determined large Japanese wood mouse. In *Chapter Six,* the draft of capybara genome sequences is reported. Using the genome data sequences, the evolutionary rate of rodents is investigated, and the rate is linked to the number of species in each clade. The final chapter, *Chapter Seven,* gives the general conclusions of the study. The impact and the contribution to science are discussed, before the future directions are presented.

From *Chapter Two* to *Chapter Three,* each project or study is presented. Each of the chapter addresses a particular question, starting with the *Chapter summary,* which gives the concise abstract of the chapter. *Introduction* gives the background information and the justification for the analyses in the chapter. Under *Materials and methods,* detailed description of the samples and data used is given. Also, the steps taken for the analyses are presented. The *Results* session presents the major findings of chapter. The implications of the results are discussed under *Discussion*. The supporting materials are presented in *Appendices,* at the end of the thesis. Each chapter is written to independently address specific questions. Although all chapters have the main objective of unravelling genomic changes associated with rodent evolution, each chapter can be understood without reference to another chapter.

CHAPTER TWO

JAPANESE GIANT FLYING SQUIRREL GENOME SEQUENCING

2.0 Chapter summary

Japanese giant flying squirrel, *Petaurista leucogenys,* is one of the rodent species that is endemic to Japan. This squirrel species, referred to as “musasabi” in Japanese has a highly restricted distribution. Despite the restricted distribution however, the species is not an endangered species, suggesting that the species has successfully adapted to its niche. The unavailability of genome data has limited the population genomics and evolutionary study of this species. The genome assembly, when completed, could be used to evaluate population size history and would be important for the investigation of genetic basis for the gliding ability of the species. In this Chapter, I report the sequencing of the musasabi genome. The genome sequences were determined to the estimated coverage of about 50×. Using the genome data, I established that the emergence of the three rodent lineages happened in rapid succession with squirrel-related lineage splitting before the divergence of mouse-related and Ctenohystrica lineages. The rate of conserved noncoding sequence (CNS) loss is heterogeneous, with squirrel-related lineage having the least and mouse-related lineage having the highest CNS loss.

2.1 Introduction

Among the three rodent lineages (Blanga-Kanfi et al. 2009), the mouse-related lineage is probably the most studied clade. The reason for this bias might be connected to the fact that mouse and rat which have been used as model mammalian species belong to this lineage. Recently, the genome sequence of naked mole rat was published (Kim et al. 2011). This, in addition to the previously determined guinea pig genome, makes the investigation into the Ctenohystrica lineage more feasible. Many efforts have not been put into the study of evolutionary genomics of squirrel-related lineage until now. The only available squirrel genome data (thirteen lined ground squirrel) in *Ensembl database* were determined at low coverage. Interestingly, squirrel-related lineage is the basal lineage in rodent evolution (Huchon et al. 2002; Blanga-Kanfi et al. 2009). Because this lineage is important for the comprehensive understanding of rodent evolution, I decided to determine the complete genome of Japanese giant flying squirrel.

Japanese giant flying squirrel, *Petaurista leucogenys*, is a rodent species belonging to Scuiridae. The species is native to Honshu, Kyushu and Shikoku islands of Japan, and is also found in Guangzhou in China (Ishii and Kaneko 2008). Japanese giant flying squirrel hereafter referred to as “musasabi”, like other flying squirrel, uses the tail and the web of skin between its legs for gliding. Musasabi is nocturnal, being active mostly in nights. It usually lives in burrowed holes on trees. For commuting between trees, musasabi usually glides. It can glide up to 115m horizontal distance (Ando and Shiraishi 1993). The species start gliding from young age. However, they start reproducing from age of 21-22 months (Kawamichi 1997).

Despite the fact that musasabi is endemic to restricted islands of Japan and China, it is classified as being least concerned by the International Union for Conservation of Nature, IUCN (Ishii and Kaneko 2008). The restricted distribution of musasabi suggested that it is adapted to certain environments. The restricted distribution might also be because of the migration limitation. Although the species can glide, gliding over kilometers has not been reported. Because the species are found on islands and are unable to swim, migration is heavily limited to their endemic islands. Also, its relatively big size makes migration by some agents (for example human transportation) difficult. Unfortunately, musasabi population genetics has not been extensively conducted. For example, whether they can survive in other environment has not been tested. In any case, the restricted distribution could be because of their ecological adaptation to a restricted niche (Carnaval et al. 2014) or because of the limitation in migration.

Of particular interest is the classification by the IUCN. The IUCN classifies musasabi as least concern with no threats of extinction. This suggests the ecological success of the species in the environment in which they are found. The ecological success of musasabi makes it an important species in understanding rodent phenotypic diversity. Particularly, musasabi population genetics would shed more lights into factors responsible for ecological distribution of rodents. Unlike mouse and rats which are ubiquitous, probably due to human migration, musasabi's restricted distribution highlights another layer of rodent phenotypic diversity. Species classified as being least concerned should have high effective population size. So far, to the best of my knowledge, there is no report of the effective population size of musasabi. With a single genome, it is possible to analyze the population size dynamics over timescale (Li and Durbin 2011). I was involved in analyzing crab-eating macaque genome data using this method in Osada et al. (2015). It would therefore be interesting to estimate the population history using musasabi genome sequences.

In this Chapter, I report the first sequencing effort of the musasabi genome. The general objective of the genome sequencing is to elucidate rodent evolution by considering the three main lineages of rodents. The sequencings were performed on Illumina platforms using Hiseq 2500 and Miseq. The genome is analyzed to extract important evolutionary and population genetics information.

2.2 Materials and methods

2.2.1 Sample description

An adult male musasabi individual was captured from Sannohe-gun, Aomori Prefecture, Japan on October 7th, 1989. The tissues of the individual were labelled and stored at the Laboratory of Wildlife Biology, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan. The first DNA extraction attempt was done using frozen liver sample. Because of the low DNA quality from liver, the second attempt was done using heart muscle. The stability of DNA from heart muscle was much higher than that of liver sample. Therefore, longer intact DNA could be extracted from heart tissue.

2.2.2 DNA extraction

The DNA was extracted from liver and heart muscle samples separately. A 50mg tissue section was thoroughly cut into small pieces using sterilized scissors. After cutting to small pieces, 300ul of Qiagen ATL reagent and 40ul Proteinase K were added. The mixture was vortexed and then incubated at 56^{o}C until total dissolution of the sample. The dissolved mixture was treated with RNase (8 µl of 100 mg/ml RNase A was added), followed by vortexing and then incubated for 30min. After RNase treatment, equal volume of phenol chloroform isoamyl alcohol (25:24:21) was added. The mixture was vortexed and centrifuged for 10min at 120,000G. The clear supernatant layer was transferred into a new tube and 1/10 volume of sodium acetate was added. DNA precipitation was done by adding twice the volume of 98% ethanol. The precipitated DNA was washed in 70% ethanol. After the washing, the pallet was air-dried and then eluded in 200ul of TE buffer. The quality of the DNA was assessed using gel electrophoresis, NanoDrop Spectrophotometer, Qubit® 3.0 Fluorometer and Agilent 2100 Bioanalyzer.

2.2.3 Library preparation for Miseq sequencing

The library was prepared with Agilent SureSelect QXT library preparation kit. For Miseq library preparation, DNA extracted from liver sample was used. The starting material was 30ng genomic DNA sample contained in 1ul of ultrapure water. The genomic DNA sample was then enzymatically fragmented and adaptor-tagged according to the manufacturer protocol. The adaptor-tagged library was then purified with AMPure XP beads before PCR amplification. The amplified library was then purified with AMPure XP beads after which library DNA quantity and quality were assessed using Qubit® 3.0 Fluorometer and Agilent 2100 Bioanalyzer and DNA 1000 Assay. The average fragment size was 700bp.

2.2.4 DNA sequencing

DNA was sequenced on two platforms. The first platform was Miseq. I did the whole genome sequencing using the library of 700bp fragment size that I prepared. Paired-end sequencing of reads 350bp (read 1) and 250bp (read 2) was conducted. The library preparation and sequencing was done in the Division of Human Genetics, National Institute of Genetics, Mishima, Japan. After the sequencing, the quality of the reads was assessed.

The second platform for sequencing was Hiseq2500. The library preparation and sequencing with Hiseq2500 was done in RIKEN, Yokohama, Japan. One set of library was prepared from liver DNA sample using Illumina TruSeq DNA PCR-Free Sample Prep Kit. The fragment size was 350bp. From the library, a lane of 100bp paired-end sequencing was conducted. Another set of library was prepared from heart muscle DNA sample using Illumina TruSeq DNA PCR-Free Sample Prep Kit. Longer fragment size of up to 2kb could be prepared. Another lane of 100bp paired-end sequencing was run. To improve the assembly, two mate pair library sets were prepared using Illumina Nextera Mate Pair Gel-Plus Sample Preparation Kit. The fragment size for the mate pair library sets were 2kb and 8kb. A lane of 100bp paired ends reads were prepared from the fragments made from each of the mate pair library sets. The quality of the reads from Hiseq2500 platform was assessed.

2.2.5 Genome assembly

Only paired-end reads from Miseq and Hiseq2500 platforms were used for contig formation. This was because the inclusion of mate pair reads resulted in poor contig formation. The contig formation was done using CLC-workbench available in Cell Innovation program (https://cell-innovation.nig.ac.jp/index_en.html). The default settings of CLC workbench were used.

2.2.6 Gene extraction

GMAP (Wu and Watanabe 2005) was used to extract musasabi genes using thirteen-lined ground squirrel coding sequences as the query. Although the quality of the thirteen-line ground squirrel is not excellent, it is the only squirrel species with annotation in Ensembl database. For higher sensitivity, I used 'cross-species' option. The GMAP commands are as follow;

```
gmap_build -D . -d Musasabi -k 8 -q 1

gmap -n 1 -t 100 --cross-species -D . -d Musasabi transcript.fa -Q > Musasabi_prot.fa

gmap -n 1 -t 100 --cross-species -D . -d Musasabi transcript.fa –E genomic >
Musasabi_exons.fa
```

The extracted exons were joined to form a single coding sequence per homologous gene of thirteen-line ground squirrel.

2.6.7 Evolutionary analyses of coding sequences

Reciprocal best hit analyses were done to establish gene orthology. Species used for evolutionary analyses include mouse and rat (mouse-related clade), naked mole rat and guinea pig (Ctenohystrica), thirteen-lined ground squirrel (another squirrel-related clade), human and gorilla (outgroup species), in addition to musasabi. Multiple sequence alignments of the genes were done using ClustalW (Larkin et al. 2007). The details of homologous gene extraction and multiple sequences alignments are presented in sections 6.2.7 and 6.2.8 of Chapter Six.

2.6.8 Conserved noncoding sequence analyses

The ancestral conserved noncoding sequences (CNSs) were retrieved from chicken, human, dog and cow genomes. Repeat-masked genomes and annotation data of the four species were downloaded from Ensembl database. The coding regions of the genomes were masked to “N”. After masking, pairwise homology search was done using BLASTN (Altschul et al. 1997), with chicken as the query. Regions overlapping or with homology to any annotated gene were removed. Regions of the chicken genome found in all the three mammalian species are termed “common”. Regions of chicken genome found in at least one of the three mammalian species are termed “union”. To investigate the dynamics of CNS loss across rodent lineages, I checked the presence of both common and union ancestral CNSs in each of the rodent genome.

2.3 Results

2.3.1 Sequencing and assembly

Good quality DNA was extracted from individual collected more than 25years ago. The quality of heart muscle appeared to be better than that of liver. High-depth genome sequencing was performed using Illumina Miseq and Hiseq2500 platforms. The details of the reads produced from each platform and run are presented in Table 2.1. Two types of libraries were prepared for Hiseq2500 sequencing. The first library type was the usual paired-end library. The second library was mate-pair library, designed for longer insert size and more effective scaffolding. To get longer read lengths, three runs of Miseq sequencings were also done. For reads from Miseq platform, read1 were 350bp while read2 were 250bp. Altogether, about 148Gbp nucleotides from ~1.2 billion reads were available for assembly. Assuming the genome size of ~3Gbp, the expected read coverage would be about 50×. The quality distributions of different library types are presented in Figure A2.1 in Appendix 2.

The inclusion of mate-pair libraries impairs contig formation. The reason for the impairment may be due to the library orientation, read composition or other reasons. Therefore, I decided to use pair-end libraries only for contig formation. As a result, mate-pair libraries were excluded during contig formation stage. CLC workbench on Cell Innovation project (https://cell-innovation.nig.ac.jp) was used for the contig formation. The total number of contigs produced was 1,253,204 (minimum length of 200bp). The summary of contig formation is presented in Table 2.2. Percent GC content of 40.6 is comparable to that of other rodent species.

The contigs were used for gene prediction with GMAP (Wu and Watanabe 2005). Of the total 18,826 annotated genes of thirteen-lined ground squirrels, 17,337 genes have musasabi homologs. The number with established orthology relationship from reciprocal best hit was 14,472. In all the eight species used, 9,221 had minimum of 100 score between thirteen-lined ground squirrel and all other species.

2.3.2 Phylogenetic relationships of the three rodent lineages

The first question addressed with musasabi genome is the phylogenetic relationship of the three rodent lineages. To establish the phylogenetic relationship, the multiple sequence alignments of 9,221 protein-coding genes were concatenated. In each lineage, two species were analyzed. All sites with gap in any species were excluded. In total, ~2.9 million amino acid sites were used for the analyses. Figure 2.1A shows the phylogenetic tree computed from the amino acids. The two species in each lineage were adequately paired with 100% bootstrap values.

The phylogenetic tree suggests that the three rodent lineages split in rapid succession. Clearly, the squirrel-related lineage appeared to have emerged before the divergence of mouse-related and Ctenohystrica lineages. The topological relationship was supported with 100% bootstrap values. Indeed, similar results were found when using first and second codon positions (Figure A2.2A and B in Appendix 2). Whereas amino acid sequences, as well as first and second codon positions are under strong influence of selection constraints, third codon positions are mostly neutrally evolving. Even for the mostly neutrally evolving third codon positions, the topology is the same (Figure 2.1B; Table A2.2 in Appendix 2). The only observed difference was in the branch lengths. Therefore, the topological relationship is well supported.

Table 2.1: Reads from different library types

ID	Platform	#Lane /run	Library type	Insert size (bp)	#Reads (mi)*	Read length (bp)	Yield (Gbp)*
Miseq	Miseq	3	PE	700	144	350,250	40
GShi10486	Hiseq	1	MP	2000	277	100	28
GShi10487	Hiseq	1	PE	2000	285	100	29
GShi10488	Hiseq	1	MP	8000	219	100	22
GShi10489	Hiseq	1	PE	350	289	100	29

PE = paired-end library

MP = Mate-paired library

*Total values for read 1 and read 2 of paired-end and mate-pair libraries.

Table 2.2: Summary of contig assembly

Parameter	Value
Total positions (Gbp)*	2.83
Determined positions (Gbp)**	2.82
Percent GC content	40.6
Minimum contig (bp)	200
Longest contig (kbp)	172.37
N20 length (bp)	26,378
N50 length (bp)	10,192
N80 length (bp)	1,664

*Total length of all contigs

**Total length of all contigs minus total number of undetermined sites (Ns)

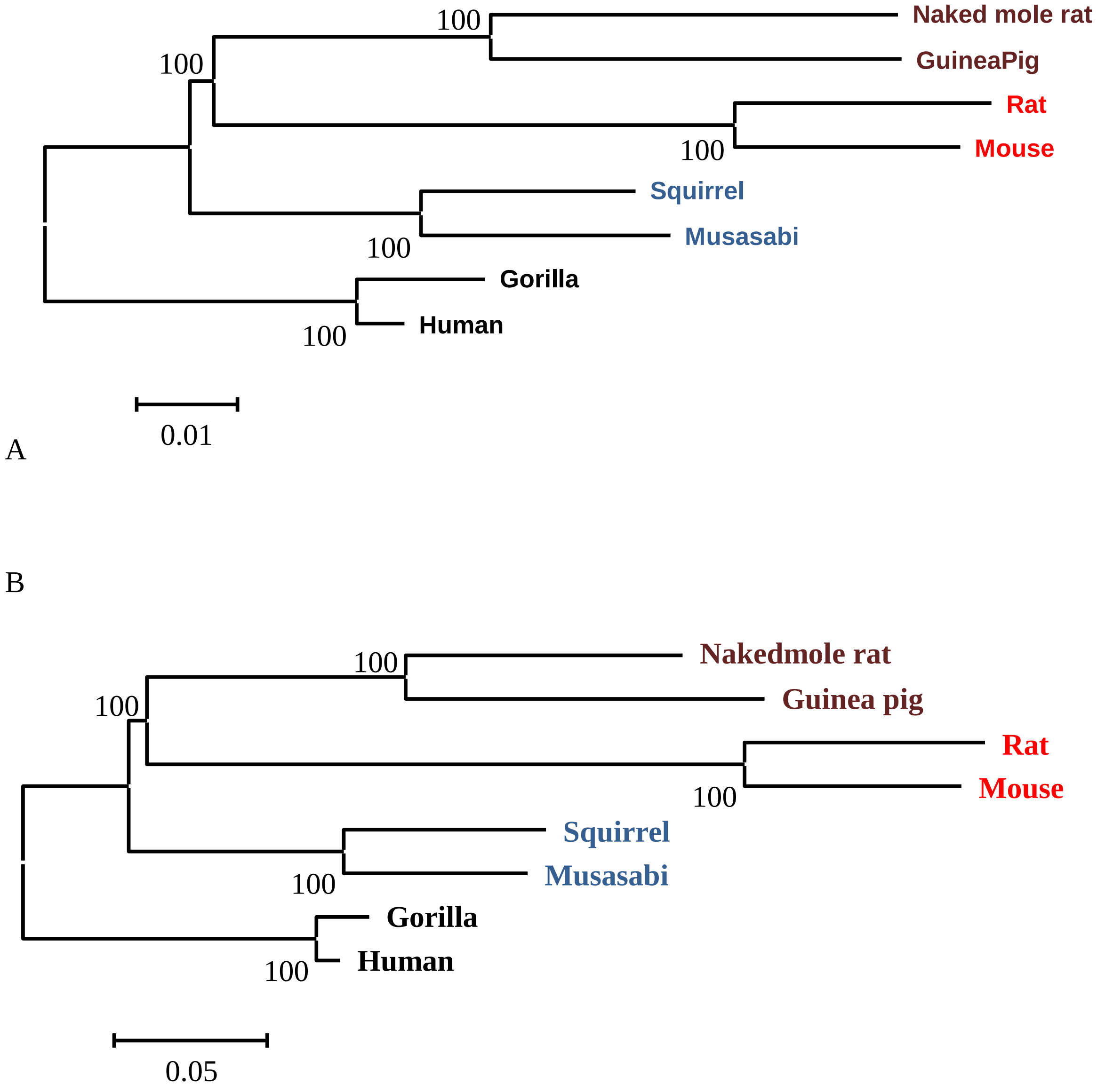


Figure 2.1: The phylogenetic relationships of the three rodent lineages. (A) NJ tree constructed from ~2.9 million amino acid sequences. (B) NJ tree constructed from ~2.9million third codon positions.

2.3.3 Squirrel-related lineage has relatively lower evolutionary rate

The genetic distance of each species to musasabi was estimated using PAML baseml. Figure 2.2A shows the median distance to musasabi using third codon positions. Expectedly, thirteen-lined ground squirrel has the shortest distance. The median distance between musasabi and thirteen-lined ground squirrel was about 0.15. The highest distances were found in mouse-related lineage. The distances were 0.96 and 0.92 for rat and mouse, respectively. The distances to human and gorilla, which are primate species, are significantly lower than the distances to rodent species outside the squirrel-related lineage, as expected from the phylogenetic tree shown in Figure 2.1. This highlights the difference in primate and rodent evolutionary rates.

To put it in better perspective, Figure 2.2B shows the genetic distance between human and other rodent species. The differences among rodent lineages become more obvious. Whereas the highest distance was in mouse-related lineage, squirrel-related lineage has the shortest distance. The differences between the lineages are significant (p value < 0.00001, Mann Whitney U test). This difference is not because of the genome quality because species in squirrel-related lineage with relatively lower coverage had the lowest distances. This observation suggests that there was a large heterogeneity in evolutionary rates in rodent history.

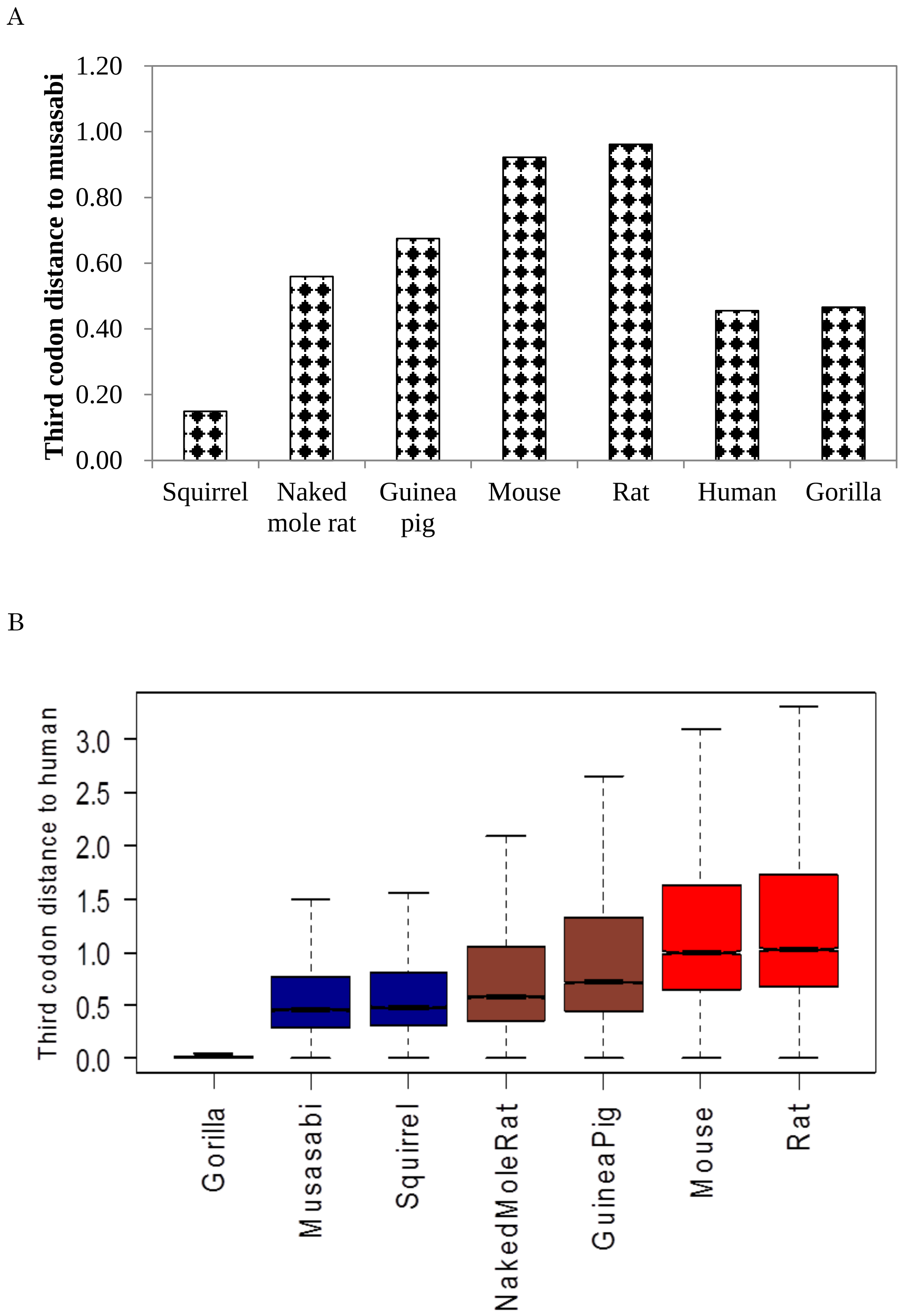


Figure 2.2: Relatively lower evolutionary rates in squirrel-related lineage. (A) The median values of the genetic distance between musasabi and other used species. (B) Heterogeneity of rodent evolutionary rate. Outliers are not shown.

2.3.4 Heterogeneity of CNS retention among rodents

Having established the phylogenetic relationships among the three rodent lineages, it is important to examine the evolutionary rate dynamics in each of the three lineages and the implications in CNS evolution. Figures 2.1B, 2.2A and 2.2B show that one of the differences is the evolutionary rate. Although the genome quality of thirteen-lined ground squirrel is not excellent, and the genome of musasabi is only in contigs, mouse, rat, guinea pig and naked mole rat have relatively good quality genomes. Species in mouse-related lineage have the highest evolutionary rate. The lowest evolutionary rate among rodents was observed in squirrel-related lineage. To evaluate the impact of the evolutionary rate on regulatory evolution, I investigated the retention of ancestral CNSs.

In Babarinde and Saitou (2013), which is presented in Chapter Four, I reported that rodents have higher rates of CNS turnover. Specifically, I found that the loss of CNS is higher in rodents than in other examined mammalian orders. I then asked whether the loss of CNSs is uniformly high in all the three rodent lineages. Figure 2.3 shows that the retention of CNSs is heterogeneous among the three rodent lineages. Notably, squirrel related lineages retained more ancestral CNSs than other lineages. The rate of loss is highest in mouse-related lineage. The pattern is the same in both common and union CNSs. It is important to note that the results would not be significantly affected by relatively recent laboratory selection and inbreeding because the evolutionary signatures span millions of years. Because the genome assembly of musasabi is still on-going, protein-coding sequences could not be adequately analyzed.

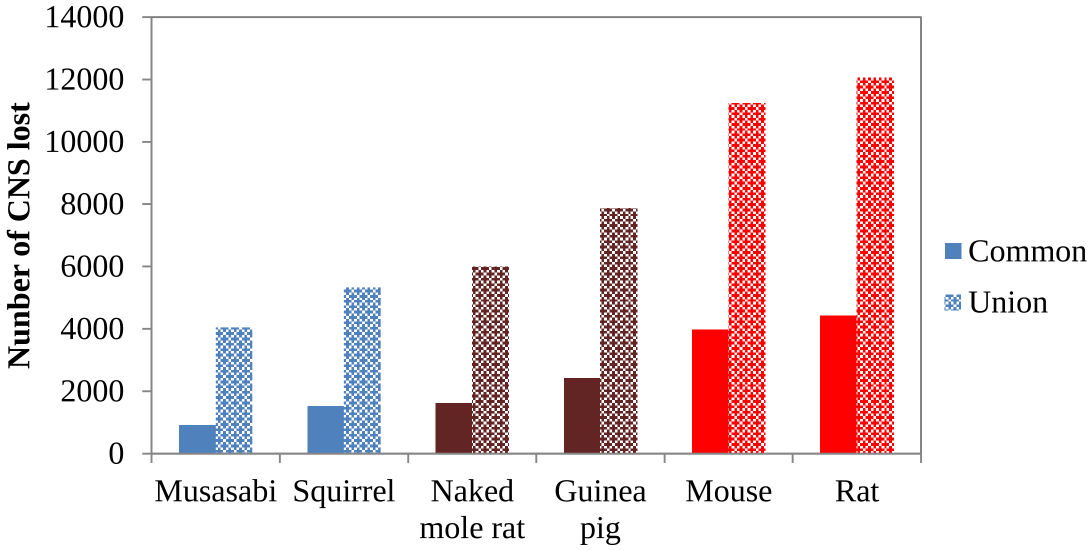


Figure 2.3: The loss of ancestral CNSs in the three rodent lineages. The ancestral sets of CNSs were extracted using chicken, human, dog and cow. The common CNSs are found in all the four species. Union CNSs are conserved between chicken and at least one of human, dog or cow genomes. The vertical axis shows the number of ancestral CNSs that are not detected in each species. Squirrel is the thirteen-lined ground squirrel.

2.4 Discussion

In this Chapter, I have reported *de novo* genome sequences of Japanese giant flying squirrel (musasabi). Good quality DNA sample could be extracted from the individual that was collected and kept more than 25 years ago. DNA extracted from heart muscle was found to be of better quality than the one extracted from liver. This suggests that DNA stability depends on the nature of the tissue. Using the extracted DNA, whole genome nucleotide sequence of musasabi was determined. Although scaffolding of the species is ongoing, the GC content is similar to what is found in other mammalian species.

Using musasabi *de novo* genome sequences with other species, I have established the phylogenetic relationships of the three rodent lineages. Similar to the reports of Blanga-Kanfi et al. (2009) and dos Reis (2012), squirrel-related lineage is basal. This result contradicts the study of Huchon et al. (2002) that reported that Ctenohystrica was basal. Also, the estimation of divergence times by Hedges et al. (2015) puts the divergence times between mouse and thirteen-lined ground squirrel at 74.0MYA. On the contrary, mouse-capybara split was put at 77.2MYA. These reports suggest that that Ctenohystrica lineage (containing capybara) was basal. However, all the data in this study supports that squirrel is basal with the highest possible bootstrap support. This shows that the accurate estimation of phylogenetic relationships in rapidly splitting lineages requires sufficient genomic data.

The analyses of the genetic distance between the examined species show an interesting evolutionary phenomenon. First, the genetic distances between musasabi and two primate species were shorter than the genetic distances between musasabi and the species from the two other rodent lineages. This supports previous reports of high evolutionary rates in rodents (WU and Li 1985; Li and Wu 1987). Second, the distance between human and examined rodent species reveal heterogeneity in the evolutionary rates of rodents. Specifically, the lowest distance was found in squirrel-related lineage. The distance between human and musasabi is even lower than the distance between human and naked mole rat. This is interesting because naked mole rat has been reported to have slower evolutionary rate (Kim et al. 2011). It is important to note that the genetic distance between human and rodent species covers more than 90 million years of history, and may not represent the extant evolutionary rate. More in-depth analyses of rodent evolutionary rate are presented in Chapter Six.

After the split, each rodent lineage has gone through independent evolutionary processes. Although the quality of the genome of some species did not allow more definite conclusion, the heterogeneity of evolutionary rates among the lineages cannot be explained by the quality of genome data. The focus on third codon position also minimizes the effect of selection constraint. Therefore, the observed differences are closely related to the actual substitution rate differences. The evaluation of the impact of the evolutionary rate differences on regulatory evolution is important. In Babarinde and Saitou (2013), I reported that the loss of ancestral CNSs was higher in rodents, compared to other examined mammalian orders. Probably as the consequence of the heterogeneity of evolutionary rates, the retention of ancestral CNSs varied among lineages. Specifically, mouse-related lineage has lost the highest number of ancestral CNSs, while the squirrel-related lineage lost the least.

The analyses of the population genomics of musasabi genome are still ongoing. The population dynamics of the species would be well investigated. The analysis in this Chapter was done using assembled contigs. The scaffolding is yet to be completed. After the completion of the genome assembly, transcriptome sequencing would be done to improve the annotation. Frozen sample of a musasabi individual has already been obtained. With the analysis, more detailed information of musasabi genome would be obtained. Possibly, combining the transcriptome analyses with the genome evolutionary analyses, the candidate genomic regions responsible for musasabi flying ability can be uncovered. An important strategy would be to computationally investigate musasabi-specific regulatory sequences that may be related to the gliding ability. By collaborating with researchers doing wet experiments, the candidate regions can then be experimentally tested by genome editing. If the candidate regions are inserted into mouse genome, similar morphological feature may be found in mouse.

CHAPTER THREE

ANALYSES OF CONSERVED NONCODING SEQUENCE REGULATORY FUNCTIONS

3.0 Chapter summary

Experimental studies have found the involvement of certain conserved noncoding sequences (CNSs) in the regulation of the proximal protein-coding genes in mammals. However, whether these observations are general features of CNSs or not should be properly investigated. For in-depth analyses of CNSs, I used CNSs conserved among chicken, human, rodent, cattle and dog. These species were chosen based on the quality of genome data and the need to avoid setting thresholds as any sequence conserved among these species is less likely to be neutrally evolving. I show that CNSs are indeed under purifying selection and are less likely to be noncoding RNAs. The physical distances between CNSs and proximal genes tend to be conserved, suggesting functional importance. Combining RNA-seq and ChIP-seq data, I show that the CNSs are more likely to be regulatory elements associated with more conserved protein-coding gene expression. The content of this chapter has been accepted for publication with the title “Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics” in Molecular Biology and Evolution (Babarinde and Saitou, 2016).

3.1 Introduction

Conserved noncoding sequences (CNSs) are the noncoding parts of the genome that are under sequence constraint probably due to the functional importance. Generally identified by computational searches, the exact numbers of CNSs are difficult to estimate. This is because the number of CNSs retrieved from a computational search depends on the threshold used (e.g., Bejerano et al. 2004; Takahashi and Saitou 2012; Babarinde and Saitou, 2013). However, some consistent properties of CNSs were found despite the difference in thresholds used. One such property is the general tendency to cluster around certain types of genes (Woolfe et al. 2005; Takahashi and Saitou 2012; Babarinde and Saitou, 2013; Bhatia et al. 2014). Specifically, CNSs are found to be overrepresented around genes involved with transcription, development and nervous system. On the contrary, genes associated with defense, immunity and response to stimulus have been shown to have lower number of mammalian CNSs around them (Babarinde and Saitou, 2013). Based on the proximity to the protein-coding genes, mammalian lineage-specific CNSs were suggested to be important for the regulation of the genes around which they are found (Takahashi and Saitou 2012; Babarinde and Saitou, 2013). In fact, computational analyses and experimental analyses have shown the involvement of certain CNSs in the regulation of the closest genes (e.g., Sumiyama et al. 2002; Bhatia et al. 2014). This understanding has been classically employed in identifying the potential regulatory elements of genes of interest (e.g., Göttgens et al 1999; Sumiyama et al. 2002). Basically, certain lengths of the flanking regions of genes of interest are searched to identify the conserved regions. The identified potential regulatory regions are then tested experimentally (e.g., Göttgens et al. 1999; Antoniv et al. 2001; Sumiyama et al. 2002, 2003; Visel et al. 2009). This approach fundamentally assumes that the regulatory elements are found at a reasonable distance from the gene of interest.

Location of genes in syntenic blocks has also been suggested to be important for gene regulation (Irimia et al. 2012). In this case, the regulatory element of a gene is resident in the intron of a neighboring gene (see Lettice et al. 2003; Sagai et al. 2004, 2009 for specific examples). Unlike housekeeping genes which tend to have shorter introns (Eisenberg and Levanon 2003; Rao et al. 2010), some genes may have larger introns so that they can "house" their regulatory elements or those of their neighboring genes (Calle-Mustienes et al. 2005). The location of regulatory elements in the introns of genes might be to ensure proper genomic location. These observations imply that the function of one gene may be affected if the gene and its regulatory elements are not located at the specified location.

Some observations, however, challenge the importance of the coexistence of the regulatory elements and the target genes. One example is the mega base pair long range regulatory activities (Lettice et al. 2003; Sagai et al. 2004, 2009). The distal regulatory element of *shh* gene was not only located far away from the gene, but also found inside intron 5 of another gene, *Lmbr1* (Lettice at al. 2003). This long range interaction can be brought about by the DNA looping structure (Ong and Corces 2009). In addition, recent chromosome conformation capture, also called 3C (Dekker et al. 2002), circularized chromosome conformation capture, also called 4C (Zhao et al. 2006) and carbon copy chromosome conformation capture, also called 5C (Dostie and Dekker 2007) technologies have revealed inter-chromosomal DNA interaction suggesting that the regulatory element and the target genes may be located on different chromosomes. This observation is not entirely new, as one promoter on a chromosome was reported to initiate the transcription on the separate chromosome (Morris et al. 1998).

I previously reported that CNSs are not always in homologous positions with respect to genes (Babarinde and Saitou 2013, presented in Chapter Four). For example, I found cases in which intergenic CNSs in one species is intronic in another species. These observations suggest that regulatory elements can interact with their target genes from any part of the genome. They only have to be brought in contact during their activity. We can thus hypothesize that the clustering of regulatory elements is not because they regulate nearby gene expression, but because they are more stable and/or active in that region.

Regarding the importance of the genomic location of CNSs with respect to the protein-coding genes, two hypotheses can be tested. The null hypothesis is that the genomic location of CNSs is not related to their regulatory function. The alternative hypothesis is that the genomic location is important, and that CNSs function best at specific location. To investigate these hypotheses, I extracted the CNSs conserved among chicken, human, mouse, dog and cattle. Employing a combination of evolutionary and statistical approaches, I found series of evidences supporting the second hypothesis that the genomic location is important for the regulatory activities of most mammalian CNSs.

3.2 Materials and methods

3.2.1 Datasets used

The repeat-masked genome data of chicken, human, mouse, dog and cow used in this study were downloaded from *Ensembl* database build 72. The protein-coding and lincRNA annotation data were retrieved from *Ensembl* biomart. Orthology data were downloaded from *Ensembl biomart* (Vilella et al. 2009). The phastcons and phyloP conservation scores were downloaded from UCSC table browser. For human conservation scores, 100-vertebrate conservation score table was used. For mouse, 60-vertebrate conservation score table was used. Tissue expression data were obtained from Necsulea et al. (2014) and the ChIP-Seq data were retrieved from LICR Histone track of UCSC table browser for mouse. The SNP data of Asian population of Phase 1 of 1000 Genome Project were used for DAF analyses. The liftover chain files were retrieved from UCSC database. For the analysis of CNS-gene distance conservation, I retrieved the coordinates of vista elements with positive enhancer activity from VISTA enhancer browser (Viesel et al. 2007).

3.2.2 Retrieval of CNSs

The coding regions of the repeat-masked genomes of chicken, human, mouse, dog and cattle were masked using a custom made script. Using the chicken noncoding genome as the query and the noncoding genomes of the four mammals as subjects, independent pairwise homology search was run using BLASTN. The E-value threshold of 0.00001 was used for the search. The common regions conserved in all the five species were extracted from the blast output using python script. The sequences of the conserved regions were blasted against Refseq and Uniprot databases in addition to the *Ensembl* amino acid sequences of each of the five species. The coordinates of annotated protein-coding genes were also checked. The protein-coding regions were discarded.

3.2.3 Random coordinates

Random coordinates were picked such that for each CNS, corresponding random coordinate has the same length and is found on the same chromosome. If a coordinate overlaps with previously picked random coordinate, another coordinate is randomly picked. For intergenic random coordinates, random sequences did not overlap any protein-coding gene. For intragenic random coordinates, random coordinates were picked from the intragenic region if they did not overlap any coding region. For coding random coordinates, random coordinates were picked from the coding regions. For each class of CNSs, random coordinates were picked in ten independent samplings.

3.2.4 Derived allele frequency spectrum

The single nucleotide polymorphism coordinates overlapping with CNSs were extracted from Asian Population of 1000 Genome Project (McVean et al. 2012). The corresponding coordinates of chimpanzee, gorilla and orangutan were extracted from the coordinates of human biallelic SNPs using UCSC liftover tool. For biallelic human SNPs, the ancestral states of the

SNPs were parsimoniously determined. Specifically, an allele found in at least two outgroup species was determined as the ancestral state. Any SNP in which the ancestral state could not be determined was discarded. The frequency of the derived alleles for each SNP position was extracted from the VCF file of Asian Population of 1000 Genome Project (McVean et al. 2012). For each coordinate category, the distribution of the derived allele frequencies was computed for each category of coordinates.

3.2.5 CNS-gene association

For intragenic CNS, the CNS is assumed to be associated with the gene inside which it is found. For intergenic CNS, the CNS is assumed to be associated with the gene with the closest transcription start site (TSS).

3.2.6 Distance conservation

For this analysis, I used CNSs and protein-coding genes with only one copy in human and mouse genomes. Cases in which the closest gene overlaps with another gene were discarded. I determined the distance between the CNS and the closest gene in human for examining the CNS-gene distance conservation. I also determined the distance between the orthologous CNS and the orthologous closest genes in the mouse genome. The difference between human and mouse CNS-gene distance was normalized by the average of the distances to give the “relative distance difference” (RDD). As control, I also computed the distance conservation of similar number of closest pairs of protein-coding genes and vista enhancer element (Visel et al. 2007).

Because of the limited number, intragenic and intergenic vista enhancer elements were not separated. The RDD between human and mouse were computed from the following equation:

$$\mathrm{RDD} = \frac{|Xh - Xm|}{mean}, \tag{1}$$

where *Xh* and *Xm* are the CNS-gene physical distance in the human genome and the mouse genome, respectively, and *mean* = (*Xh*+*Xm*)/2. Normalized difference of the TSS genomic coordinates of the two adjacent protein-coding genes is defined as the gene-gene distance.

3.2.7 Retrieval of housekeeping genes

I followed the definition of housekeeping genes reported by Eisenberg and Levanon (2013). For this analysis, I used the expression data of Necsulea et al. (2014). The criteria used to define housekeeping genes as reported by Eisenberg and Levanon (2013) include; (i) expression observed in all 15 tissues; (ii) low variance over tissue expression values: standard-deviation [$\log_2$(RPKM)]<1; and (iii) no exceptional expression in any single tissue; that is, no log-expression value differed from the averaged $\log_2$(RPKM) by two (fourfold) or more. This returned 4,161 genes in human and 3,902 genes in mouse.

3.2.8 Gene enrichment test

I first made a list of all genes in GO terms with A_{total} elements. I then made a list of CNS-associated genes with GO terms with A_{CNS} elements. Genes associated with multiple CNSs were represented multiple times in the CNS-associated gene list. For each tested GO term, I counted the number (T_{CNS}) of CNS-associated genes. I also counted the number (T_{total}) of the genes with the term in the all gene list. The expression for the enrichment is given as:

$$\text{Enrichment (fold change)} = (T_{CNS}\times A_{total})/\ (A_{CNS}\times T_{total}). \qquad (2)$$

Statistical significance was calculated with a binomial test with Bonferroni correction as implemented in Panther (Thomas et al. 2003).

3.2.9 Modified CGN computation

Conservation of genomic neighborhood score which quantifies the stability of a gene neighborhood was proposed by De et al. (2009). The CGN score of a given gene in human is simply the fraction of genes within a window (of size *2Mbp*) surrounding it, which have orthologs in an equivalent window in its mouse ortholog. The score was computed for genes with one-to-one orthology in human and mouse. For a human window, the original CGN expression was;

$$\text{CGN} = \frac{Number\ of\ human\ genes\ conserved\ in\ a\ window}{Total\ number\ of\ human\ genes\ in\ the\ window}. \qquad (3)$$

“Conserved” in this context implies that orthologous genes are found in the orthologous windows in the two species being considered. Incorporating the conservation in mouse neighborhood, the expression for the modified conservation of genomic neighborhood (modCGN) in this study is;

$$\text{modCGN} = \frac{Number\ of\ human\ genes\ conserved\ in\ a\ window}{Avearge\ number\ of\ human\ and\ mouse\ genes\ in\ the\ window}. \qquad (4)$$

To calculate modCGN, I focused on human gene with one-to-one correspondence in mouse. I searched 1Mbp distance upstream and downstream of the gene in human and retrieved all genes found in the human window. I performed similar search in the mouse genome to get all the genes found in mouse windows. I then counted the number of genes in human window found in the corresponding mouse window. The ratio of this count to the average number of genes for human and mouse windows give the modCGN.

3.2.10 Expression conservation

I used twelve tissues (see Table A3.2 in Appendix 3) with expression data in human and mouse. For every gene with one-to-one orthology in human and mouse, I computed the Spearman's correlation coefficient from the expression data of the twelve tissues. For each category (0CNS, 1-3CNSs, 4-9CNSs, >9CNSs) of genes, I calculated the average Spearman's correlation coefficient.

3.3 Results

3.3.1 Acquisition of mammalian CNSs and their basic characteristics

I conducted homology search using BLASTN (Altschul et al. 1997) on repeat- and coding sequence-masked genomes of chicken, human, mouse, dog and cattle (see section 3.2 of Material and methods) to identify mammalian CNSs. The chicken genome was used as the query. Since the nucleotide divergence between mammals and chicken is sufficiently large (synonymous substitution rate > 1), I did not have to set percent identity threshold. Therefore, I defined "CNS" in this study as a noncoding region conserved between chicken and the four mammalian species with the minimum length of 100bp (see Materials and Methods). The searches gave 21,584 chicken CNSs that are conserved in all the four mammalian species.

When the coordinates were mapped to human and mouse genomes, 21,191 and 21,026 CNSs were found, respectively (Figure A3.1 and Table A3.1 in Appendix 3). The slight difference in the numbers of CNSs is due to lineage-specific duplications. I focused on human and mouse genomes because of the quality of the genomes and the availability of data. Out of the human CNSs, 10,120 were found to overlap with protein-coding genes (mostly intronic), hereafter referred to as intragenic CNSs. The remaining 11,071 CNSs which do not overlap with any protein-coding gene are referred to as intergenic CNSs. I found 9,482 intragenic and 11,544 intergenic CNSs from the mouse genome. The conservation levels of the retrieved CNSs, random sequences, lincRNAs and various codon positions of protein coding genes were tested using the phastcons (Figures 3.1A and A3.2A in Appendix 3) and phyloP (Figures 3.1B and A3.2B in Appendix 3) conservation scores. The average phastcons scores for CNSs are more than 7-fold higher than random sequences. Interestingly, conservation scores clearly distinguished lincRNAs from CNSs (Figures 3.1A, 3.1B, A3.2A and A3.2B in Appendix 3). The phastcons scores for CNSs are higher than all the codon positions of the protein coding genes. The distribution patterns of phyloP conservation scores in Figure 3.1B show that the distribution of CNSs divides the distribution of second codon position into three parts. In the first part (scores $\leq$ 2), there is a higher proportion of second codon datasets. In the second part (scores 2-5), CNSs dominate. In the third part (scores $\geq$ 5), second codon dominates.

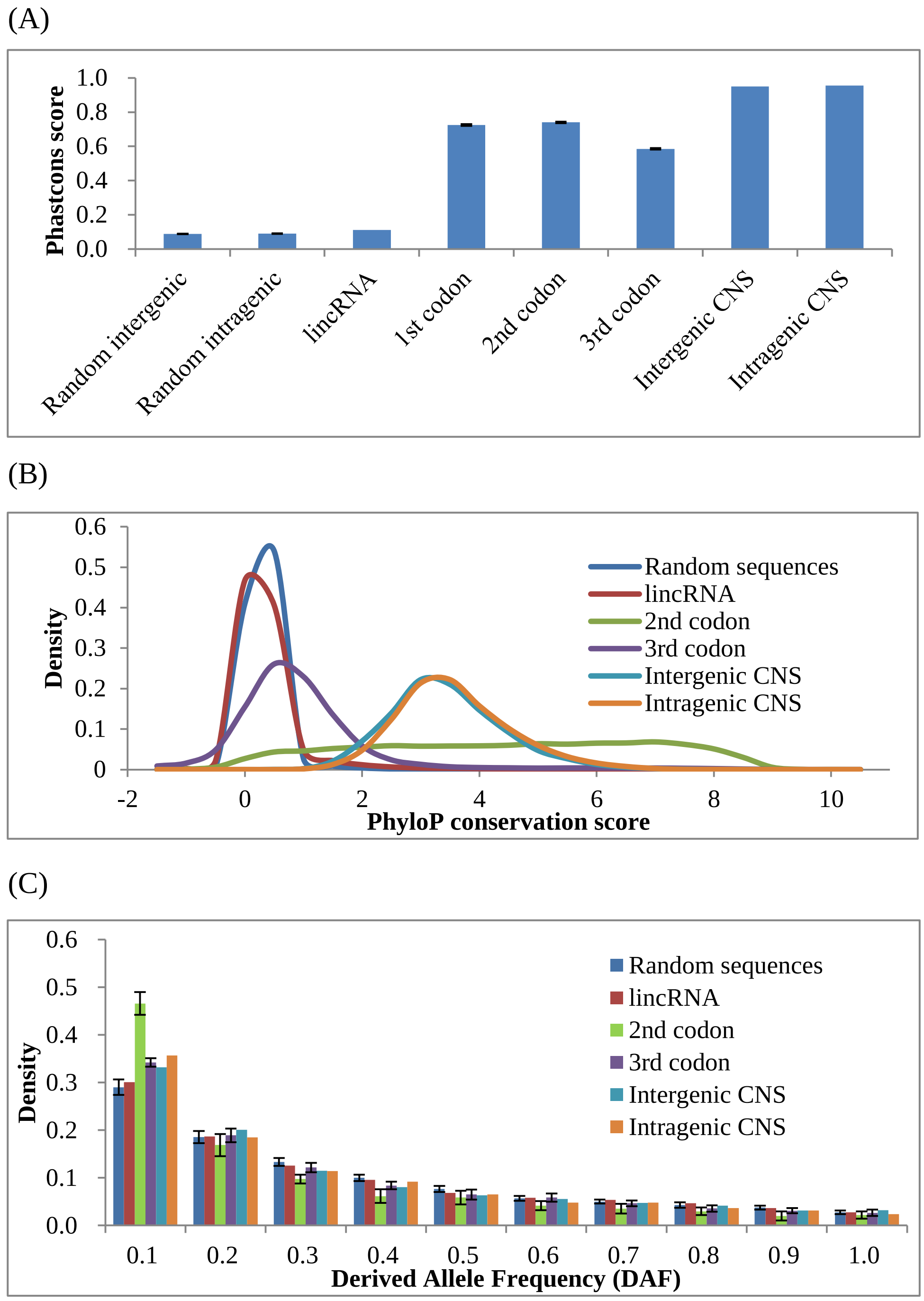


Figure 3.1: Retrieved CNSs have strong constraints in the human genome. (A) CNSs have the highest phastcons score. (B) CNSs have higher phyloP conservation score than random sequences. (C) CNSs are under purifying selection, and not mutational cold spots. (Chi square p value < 0.001). Error bars are 99.99% CI from ten independent random samplings.

This wide distribution pattern of second codon positions probably indicates the wide spectrum of selection forces acting on protein-coding gene evolution. Some protein-coding regions are poorly conserved while other regions are highly conserved. CNSs on the other hands have less poorly conserved regions. In fact, they are more conserved than random sequences, lincRNAs and third codon positions.

However, higher conservation scores do not necessarily imply functional importance. The higher similarity revealed by the conservation scores might just be the result of the low substitution rate in the region (mutation cold-spot hypothesis). To confirm if the regions are actually under purifying selection, I performed derived allele frequency (DAF) spectrum analyses (Drake et al. 2006). For regions that are under purifying selection, derived alleles would be quickly removed and would not be able to spread in the population. Therefore, there would be excess of low frequency alleles in the regions. Using genome sequences of chimpanzee, gorilla and orangutan to determine the ancestral states and SNP data from the Asian population of 1000 Genome Project (McVean et al. 2012), I compared the DAF spectra in random sequences, lincRNAs, protein-coding genes and CNSs (see Material and Methods). I found an excess of low frequency allele in protein-coding genes and CNSs, compared to the random expectation (Figure 3.1C). This clearly demonstrates that the CNSs, like the protein-coding genes, are under purifying selection.

3.3.2 Examination of CNS locations

Having established that the CNSs are under purifying selection, I then analyzed the genomic distribution of intergenic CNSs. I asked whether the CNSs often occur in clusters or in isolation. To answer this, I made 2Mbp sliding windows with 500kbp step size, and counted the numbers of coordinates in each window. Figures 3.2A and A3.3A in Appendix 3 show that, compared to the random intergenic sequences, CNSs often exist in clusters in human and mouse genomes, respectively (Chi square p value < 0.001). This shows that the distribution of CNSs is not random.

Are the CNSs preferentially located around protein-coding genes? I associated intergenic CNSs to the gene with the closest TSS and compared the distribution of the distance to the closest genes. Figures 3.2B and A3.3B in Appendix 3 show that the intergenic CNSs tend to be located far away from the TSS when compared to the random intergenic sequences both for human and mouse genomes (chi square p value <0.001). On the contrary, lincRNAs are closer to the protein-coding genes than random expectation. The location of intergenic CNSs far away from genes suggests that proximity may not be important. However, it does not say much about the importance of the actual genomic location.

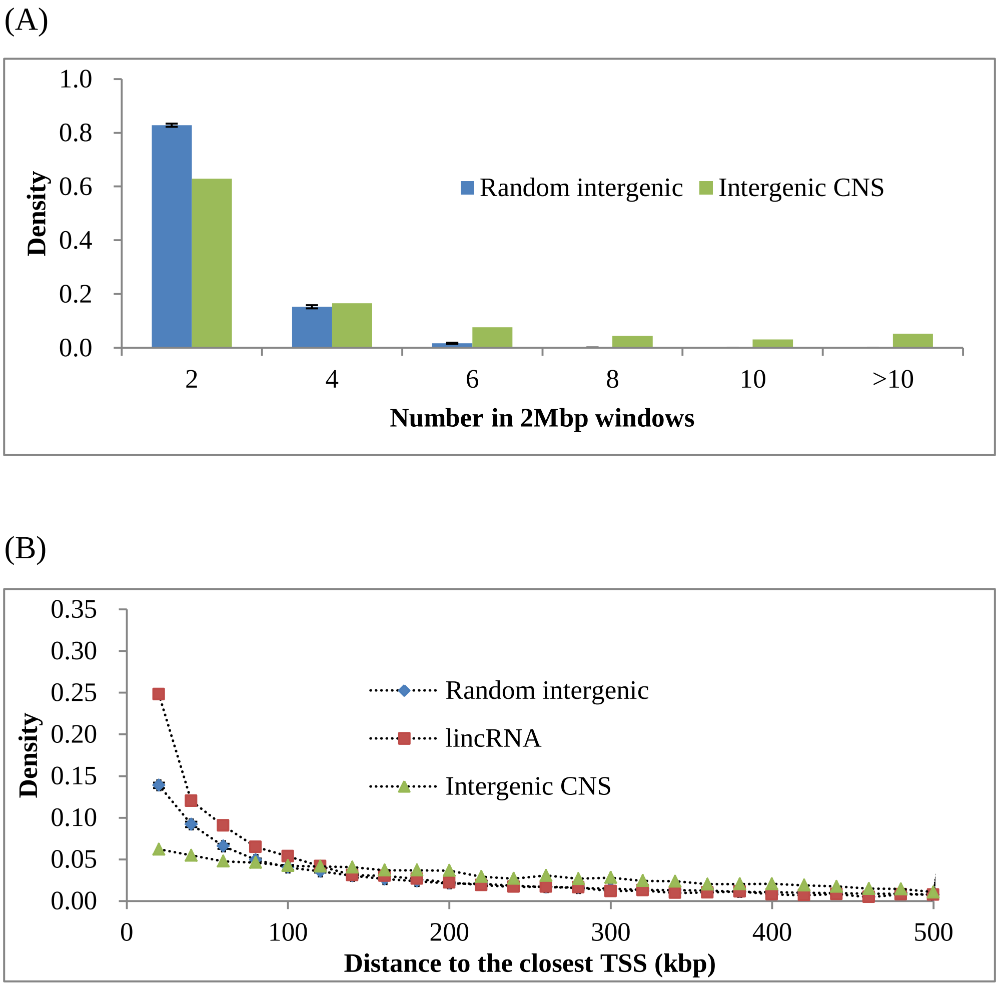


Figure 3.2: Nonrandom genomic location of CNSs in the human genome. (A) Compared to random sequences and lincRNAs, CNSs tend to exist in genomic clusters. (B) Intergenic CNSs tend to be located far away from protein-coding genes. Error bars are 99.999% CI from ten independent random samplings. For the two charts, chi square p value < 0.001.

To probe the importance of the genomic location between genes and CNSs, I investigated the evolutionary stability of the CNS-gene distance. Conservation of the distance would suggest that their genomic physical distance is important. If the CNS-gene distance is evolutionarily conserved during the mammalian evolution, the distance between human CNS and the closest human gene would be similar to the distance between the orthologous CNS and orthologous gene in the mouse genome. The normalized difference (distance relative difference) between human distance and the orthologous mouse distance would be close to zero. I thus devised a new measure, the relative distance difference (RDD) for this purpose (see Materials and Methods for definition of RDD). I computed RDDs for both intergenic and intragenic CNSs and the corresponding closest protein-coding genes. As control, I also computed RDDs for closest gene pairs (gene-gene). Figure 3.3A shows that CNS-gene RDDs are lower than gene-gene RDDs (Chi square p value < 0.001). The mean of gene-gene RDDs was 0.55. This value is more than twice the mean of CNS-gene RDD both for intergenic and intragenic CNSs (0.22 and 0.26, respectively). Similar results were found when the RDD values between human and dog, as well as between human and cow were computed. This suggests that an evolutionary force may be acting to stabilize CNS-gene distances over time, and that the genomic physical distance may be important for the integrity of the regulatory function. Interestingly, the corresponding mean RDD value (0.23) for experimentally confirmed vista enhancer elements (Viesel et al. 2007) is comparable to those of CNS-gene. If the distance conservation can be used as an important feature of functionally active enhancer elements, my results suggest that most of the CNSs in my dataset have regulatory activity.

If the conservation of CNS-gene distance is important, we would expect to find more stable expressions in genes with more conserved CNS-gene distance. For this analysis, the expression data of 12 human and mouse tissues were used. Spearman's expression correlation across the 12 tissues was computed for each protein-coding gene with one-to-one orthology between human and mouse. The genes were then grouped into four classes based on the RDD values. For genes associated with more than one CNS, the median values of the RDD were used. Figure 3.3B shows that genes with lower RDD values tend to have higher expression correlation. For example, the median correlation of coefficient of genes with first quartile RDD values (< 0.092) was 0.59, whereas the median correlation coefficient of genes with fourth quartile RDD values (> 0.362) was 0.54. This difference is statistically significant (Mann Whitney U p-value < 0.01). The result suggests that genes with more conserved CNS distance (lower RDD value) tend to have more stable expression across evolutionary timescale. This highlights the importance of CNS-gene distance conservation.

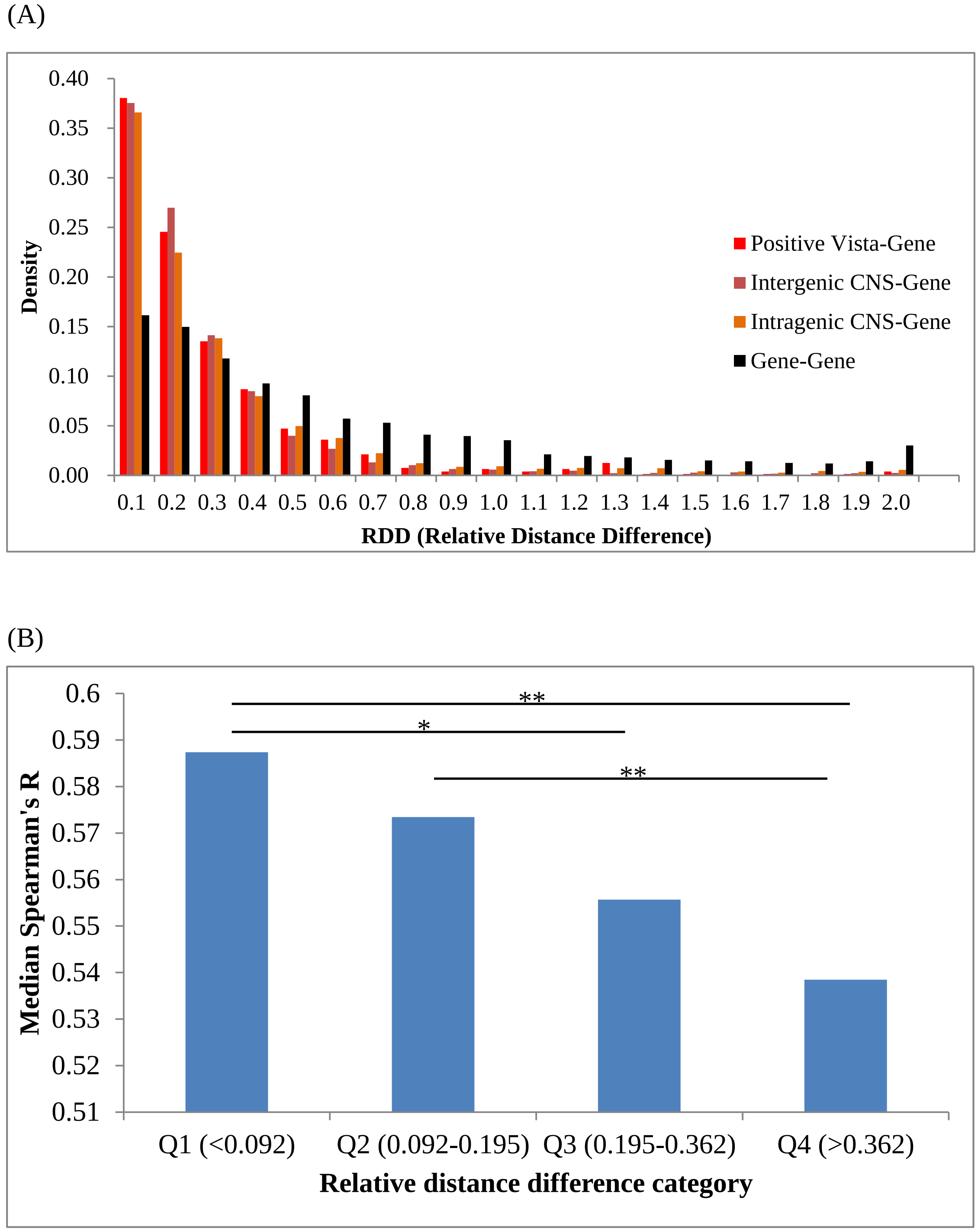


Figure 3.3: The importance of CNS-gene distance. (A) The CNS-gene distance is evolutionarily more conserved than the gene-gene distance (Chi square p value < 0.001). RDD measures the relative difference between human CNS-gene or gene-gene physical distance and mouse orthologous distance. (B) Genes with lower RDD values tend to have higher expression correlation. The genes with CNSs (n = 2,847) were ordered by RDD values, and were divided into quartiles. * p value < 0.05; ** p value < 0.01 (Mann-Whitney U test).

3.3.3 Features of CNS-associated genes

Having established the nonrandom distribution of CNSs, I then asked if genes flanked by CNSs have unique features. It has been previously reported that CNSs tend to cluster around genes involved in the nervous system, transcription regulation and development (McEwen et al. 2009; Takahashi and Saitou 2012; Babarinde and Saitou 2013), while they tend to be underrepresented around genes involved in response to stimuli as well as defense and immunity (Babarinde and Saitou 2013). I first established that these gene ontology enrichment patterns also apply to my CNS dataset (Figure 3.4A in Appendix 3). The biased genomic location around certain gene functional categories suggests that CNSs may be preferentially located for realizing specific tissue expression patterns. The enrichment patterns of gene functional categories suggest that genes expressed in embryonic brain would have more CNSs because genes involved with development, nervous system and transcription regulation would be expressed at that stage. On the contrary, testis, which is functional at adult stage, may not express many CNS-associated genes. To probe the hypothesis that testis and embryonic brain expression patterns follow the gene ontology enrichment prediction, I used the RNA-Seq data of embryonic brain and testis (Necsulea et al. 2014). I set tissue expression cutoff at 5RPKM and performed enrichment test. As expected, genes expressed in embryonic brain tended to be associated with more CNSs while genes highly expressed in testis as well as housekeeping genes were associated with fewer CNSs (Figure 3.4B). At different expression levels (0.5RPKM and 1RPKM), the results are similar (Figure A3.4 in Appendix 3).

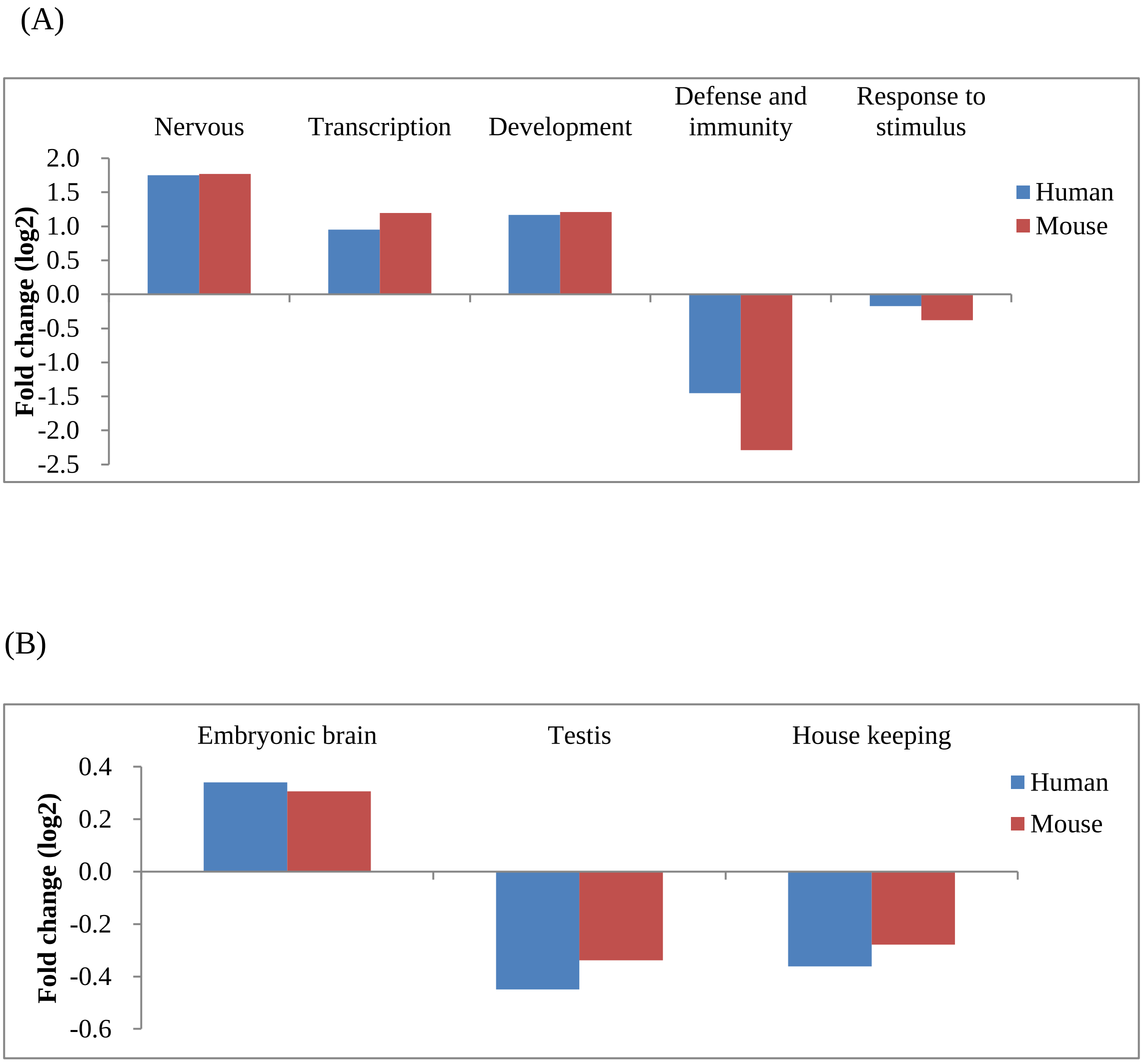


Figure 3.4: Enrichment of CNS-associated genes. (A) Enrichment test of selected gene ontology terms. (B) Enrichment test of genes expressed in certain tissues. The threshold expression level was 5rpkm. All the enrichment directions were statistically significant (Bonferroni corrected binomial p value < 0.001).

Focusing on gene expression patterns of embryonic brain, I asked whether genes with more CNSs have higher expression than those with fewer CNSs. My analyses indeed revealed that genes with more CNSs have higher expression. Genes with no CNS closest to them have lower expression level (Figures A3.5A and A3.5B in Appendix 3). The reverse is found in testis-related expression; genes with no associated CNSs tend to have higher testis expression (Figures A3.5A and A3.5B in Appendix 3).

In terms of the gene structure and genomic background, are there features that distinguish CNS-associated genes from others? For genes with more intragenic CNSs, I would expect them to have larger noncoding proportions. To keep the CNSs in one gene, some evolutionary force should act to prevent loss of noncoding regions (e.g. intron). Therefore, they would be expected to have larger noncoding proportion than genes with no CNSs residing in them. Indeed, that is what I observed (Figure 3.5A). Human genes flanked with no CNSs have significantly lower noncoding proportion (86.95%) compared to 96.68%, 98.44% and 99.18% for genes with 1-3 CNS, 4-9 CNSs and genes with 10 or more CNSs, respectively. The result is similar for the mouse genome (Figure 3.5A). Previous studies have suggested that the evolutionary force may be acting on housekeeping genes such that they would always have shorter intron size (Castillo-Davis et al. 2002; Eisenberg and Levanon 2003; Rao et al. 2010). The prediction of this model is that housekeeping genes would have more stable intron size over evolutionary time. Since the intron size is related to the proportion of noncoding regions, the noncoding region proportion of housekeeping genes should be stable over evolutionary time.

Hence, the correlation between human and mouse noncoding region proportions for housekeeping genes should be higher than for tissue-specific genes. As shown in Figure 3.4B, CNSs are underrepresented in housekeeping genes. Therefore, I would expect to see higher correlation of noncoding proportion in genes with no CNSs, which are enriched in housekeeping genes. Although the correlation of noncoding proportion between human and mouse genomes is high, this feature is not unique to the genes with no CNS (Figure A3.6A in Appendix 3). Therefore, the evolutionary force acting on short-intron genes may be similar in magnitude to the force acting on large-intron genes.

To understand the nature of the intergenic regions of CNS-associated genes, I analyzed the distance of the CNSs to the nearest genes. The distance to the nearest gene tells whether the genes exist in clusters, or in isolation. Genes in isolation are located far away from other genes. As would be predicted from Figure 3.2B, genes with no CNSs are significantly closer to the next gene than for genes with many CNSs (Figure 3.5B). To understand the effect of evolution on the distance between closest genes, I computed the correlation of the distances for the genes with one-to-one correspondence between human and mouse genomes. Interestingly, the correlation coefficient (0.59) for the genes with no CNSs is lower than that (>0.68 depending on the number of CNSs) for genes with CNSs (Figure A3.6B in Appendix 3). This higher correlation of the distance to the nearest genes in CNS-associated genes suggests that the flanking regions of CNS-associated genes might be conserved in structure.

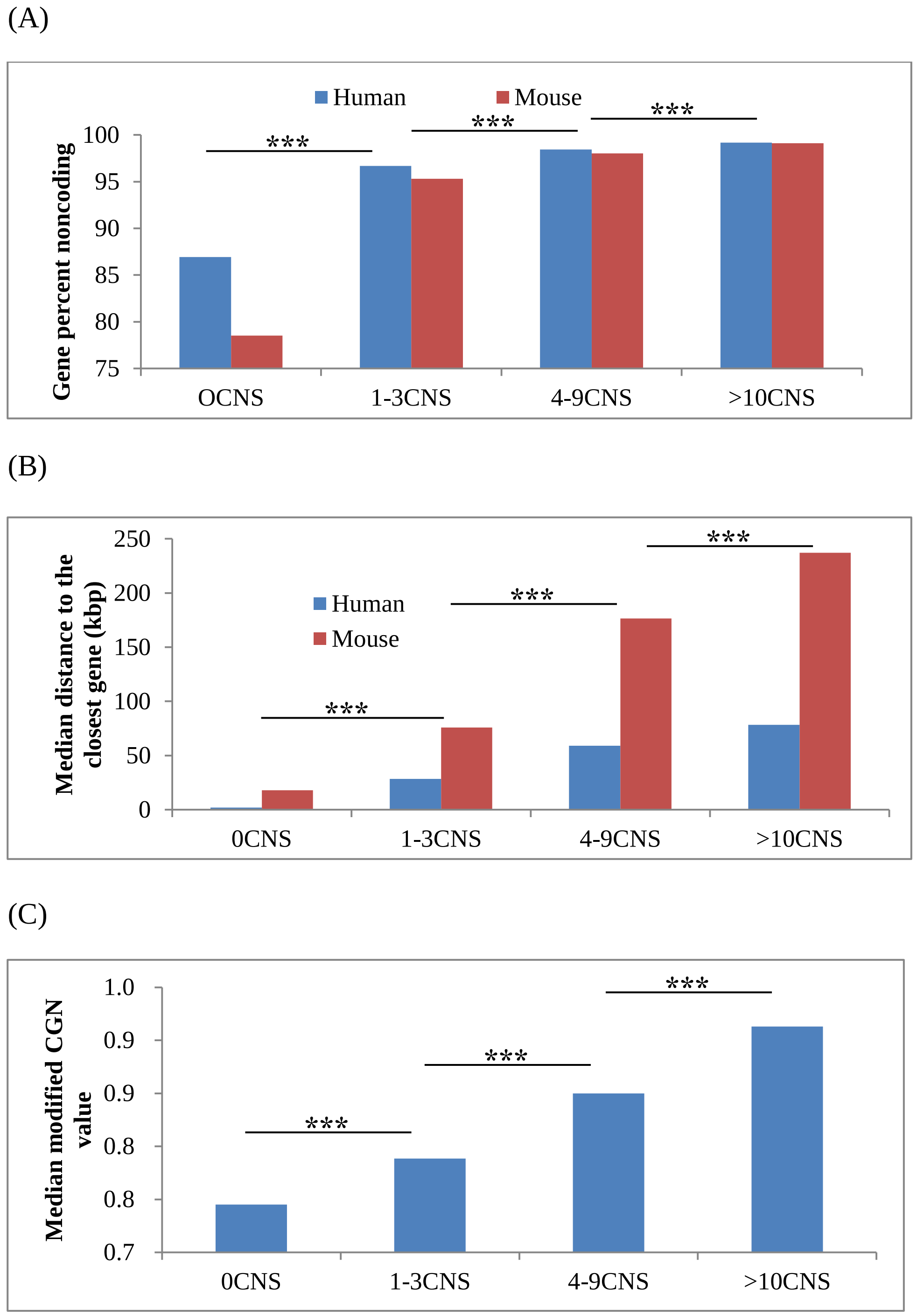


Figure 3.5: Unique features of genes associated with CNSs. (A) Genes with no associated CNS have lower noncoding percent (*** T-test p value < 0.001). (B) Genes with more CNSs tend to be located far away from other genes (*** Mann WhitneyU test p value < 0.001). (C) Genes with more CNSs are located on more conserved genomic neighborhood (*** Mann WhitneyU test p value < 0.001).

To further probe this, I computed the modified conservation of genomic neighborhood (modCGN) score for the genes. Conservation of gene neighborhood (CGN), originally proposed by De et al. (2009), is the proportion of the number of genes in a window that are found in the homologous window in another species (see Material and Methods for more details). I computed the modCGN value for genes with one-to-one correspondence between human and mouse. Figure 3.5C shows that genes associated with no CNS have lower values than CNS-associated genes. This result further shows that the genomic neighborhood of genes associated with CNSs is conserved.

3.3.4 CNS association with gene expression dynamics

So far, I have shown that the distribution of CNSs is nonrandom and that genes with CNSs have unique features. However, I have not examined direct involvement of CNSs in the gene expression dynamics. To address this issue, I analyzed the ChIP-Seq and RNA-Seq data. As shown in Figures 3.4B and A3.5 in Appendix 3, and as would be expected from the GO enrichment test (Figure 3.4A), CNS-associated genes are more expressed in embryonic brain, and less expressed in testis (Figure A3.5 in Appendix 3). If CNSs are associated with enhancer activity, higher signal of H3k4me1 and H3k27ac which are marks of enhancer elements (Akhtar-Zaidi 2005; Kim et al. 2010; Creyghton et al. 2010) would be expected in brain than testis. For this analysis, ChIP-Seq data for H3k4me1 and H3k4me3 were retrieved from UCSC table. As would be expected, CNSs have higher H3k4me1 signal in embryonic brain than random sequences (Figures 6A and A3.7A in Appendix 3). The difference between the CNSs and random sequences is not as obvious in testis and liver (Figures A3.7A and C in Appendix 3). The higher signal of CNSs is not observed in H3k4me3, which is the mark of active promoter (e.g. Cain et al. 2010) as shown in Figures 3.6A and A3.7B in Appendix 3. As would be expected, lincRNAs have higher signals of the three examined marks (Figures A3.7A, B, and C in Appendix 3), because of their transcription and proximity to protein-coding genes. H3k4me3 specifically differentiates lincRNAs from other coordinates (Figure A3.7B in Appendix 3). This analysis shows that CNSs are more associated with gene regulation in embryonic brain than in testis.

Finally, I probed the expression conservation of protein-coding genes with respect to number of flanking CNSs. Genes with more conserved expression would have a higher correlation coefficient. To do this, I calculated the correlation coefficient for each gene using RNA-Seq data of 12 human and mouse tissues from Necsulea et al. (2014). As shown in Figure 3.6B, genes with more CNSs tend to have higher expression correlation coefficient. Specifically, genes with no CNSs have the lowest expression correlation coefficient, while genes with at least 10 CNSs have the highest correlation coefficient. This shows the involvement of CNS in conserved gene expression of the neighboring genes. Interestingly, genes associated with more CNSs tend to have lower dN/dS values (Figure A3.8 in Appendix 3).

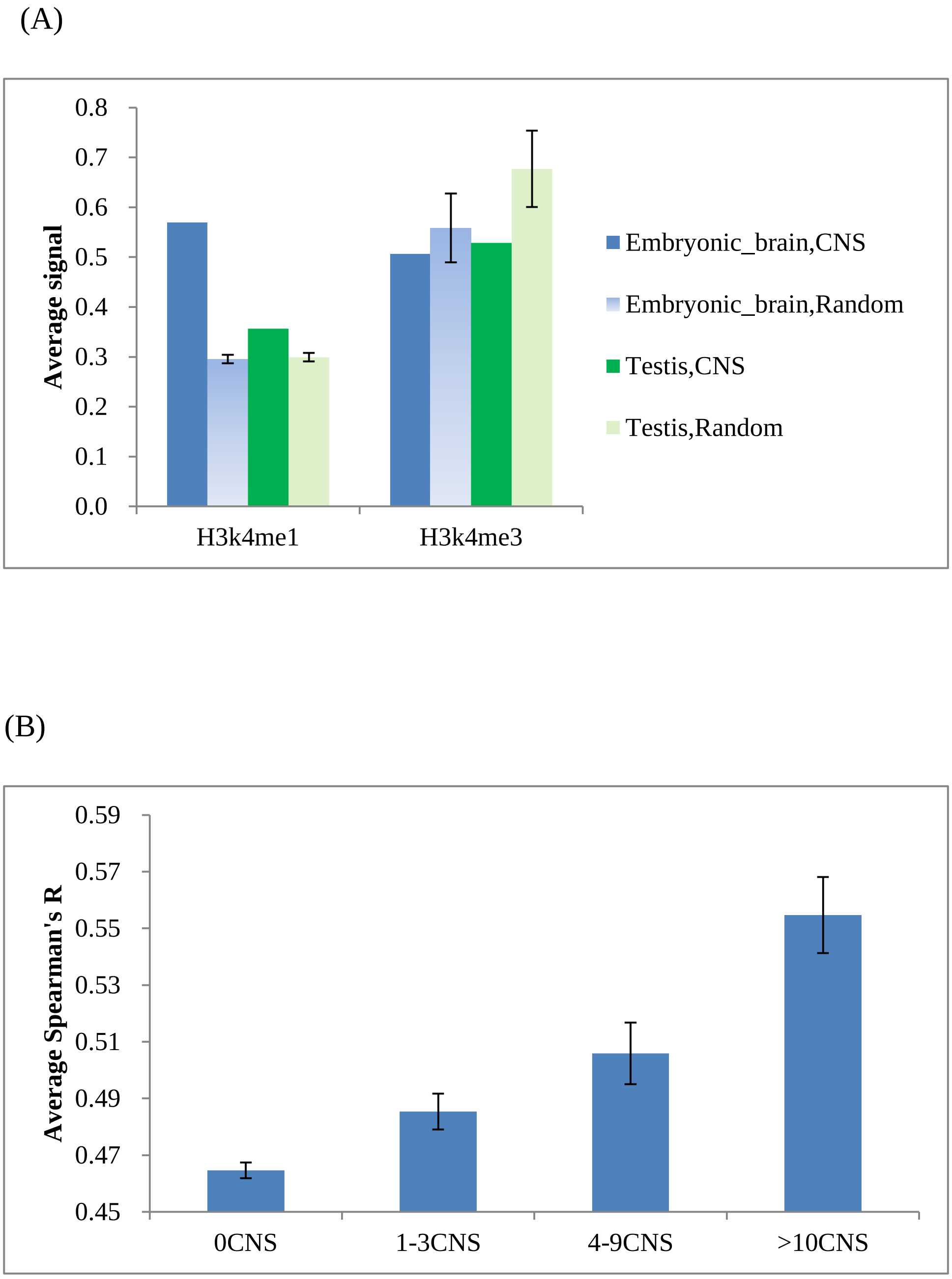


Figure 3.6: CNSs are associated with more conserved expression. (A) CNSs have stronger signals for higher H3k4me1 (enhancer) signal than random sequences. The difference is more obvious in embryonic brain. Such differences are not found in H3k4me3 (promoter) signal. Error bars are 99.99% CI from ten independent random samplings. (B) Genes with more CNSs tend to have more conserved expression. Error bars are 99.999% CI from the genes for which correlation coefficient was calculated.

3.4 Discussion

Using computational searches, I have identified ~20,000 CNSs that are conserved among chicken and four mammalian species. The conservation levels of the CNSs are significantly higher than those of random sequences and lincRNA exons. Purifying selection on CNSs is clearly stronger than observed in random sequences or lincRNAs. Interestingly, intragenic CNSs tend to have stronger constraint than their intergenic counterparts. In fact, there is overrepresentation of intragenic CNSs. For example, while the human genome gene percent is 41.54%, intragenic CNSs represent 47.76% of the total CNSs. For mouse, genome gene percent and intragenic CNS percent are 36.55% and 45.10%, respectively (these proportions are based on values given in Table A3.1). This suggests that intragenic CNSs are more stable than intergenic ones.

Intergenic CNSs tend to be located in clusters, far away from protein-coding genes. This clustering of CNSs in gene deserts suggests that the proximity to gene is not very important. In addition, CNSs are overrepresented in certain gene categories. Specifically, genes associated with development, transcription and nervous system, and/or genes expressed in embryonic brain tend to have more CNSs. Focusing on protein-coding genes, I found that genes associated with more CNSs tend to be located far away from other genes, demonstrating the bias in CNS-gene genomic location. In fact, I showed that the distance between CNS and closest genes tend to be conserved between human and mouse genomes in terms of RDD measures. This suggests that the evolutionary constraints are acting on the genomic location of the CNSs with regards to the closest target genes. Specifically, genes that require more strict regulation may have to be located in such a genomic location as to allow controlled access to the regulatory regions.

My results highlight some differences between lincRNAs and CNSs. First, the sequence constraint on lincRNAs is much weaker than those on CNSs (Figures 3.1 and A3.2 in Appendix 3). Second, while lincRNAs tend to be located close to the TSS of the next protein-coding genes, intergenic CNSs tend to be located far away from the TSS (Figures 3.2B and A3.3B). Third, the histone modification signals are different (Figure A3.7B in Appendix 3). Specifically, lincRNA H3k4me3 (promoter mark) is more than four-fold than that of CNSs (Figure A3.7B in Appendix 3). This may reflect the transcriptional activity of lincRNAs or the overlap of many lincRNAs with protein-coding promoter regions. These differences suggest that CNSs do not function as lincRNAs. I then investigated the regulatory activity of CNSs. I compared RDDs of the functionally verified vista enhancer elements and that of the identified CNS. I discovered that distance conservation of vista enhancer elements is comparable to those of my CNSs (Figure 3.3). If the distance conservation is due to the regulatory activity, the comparable strength of distance conservation between functionally verified vista enhancer elements and my CNS dataset suggests that majority of the identified CNSs have regulatory function. For example, I checked the previously reported enhancers in *Pax6* locus (Bhatia et al. 2014) and found that three of the CNSs overlapped with Id855, agCNE1 and agCNE4. These three elements drive conserved expression in forebrain, trigeminal ganglia and hindbrain, respectively. In addition, I found that genes associated with CNSs tend to be under stronger purifying selection, as revealed by dN/dS ratio (Figure A3.8 in Appendix 3). Comparing the distance conservation across selected GO terms, I found that CNS-gene RDDs of genes involved in nervous system, development and transcription tend be higher than CNS-gene RDDs without the ontology terms (Figure A3.9 in Appendix 3). The level of significance of the difference is not as high in genes involved in response to stimulus, defense and immunity (see the p values in Figure A3.9 in Appendix 3).

Furthermore, using the ChIP-Seq data, I have shown that CNSs have higher enhancer signals than random coordinates, especially in embryonic brain. The expression patterns of CNS-proximal protein-coding genes showed unique properties, demonstrating that CNSs are preferentially located close to certain protein-coding genes. My results suggest that even for long-range enhancer elements, the physically closest gene might be a target. For cases like *shh* enhancer in which the target gene of a CNS is not the closest gene, it is possible that such enhancers have multiple targets. Cases in which the proximal genes are not the target genes seem to be rare. My results therefore suggest that the genomic location of one CNS is important for its regulatory function. This further implies that phenotypic changes could be observed from genomic deletion experiment even if the regulatory element is intact. If such deletion is large enough, the genomic location of the regulatory element with respect to the target gene may be affected and that may produce certain phenotype changes.

My genomic and evolutionary analyses have highlighted some important features of

CNSs. Additionally, unique properties of CNS-proximal genes were revealed. I thus have demonstrated that the genomic location of CNSs with respect to genes is important for the proper gene regulation, and that the evolutionary force acts to maintain the genomic location. The previously reported non-homologous location of CNS with respect to gene (Babarinde and Saitou 2013) may be due to the change in gene structure (see Figure A3.10 in Appendix 3 for one example case). While the actual genomic locations of CNSs are relatively fixed, change in gene structure, such as exon loss or gain or *de novo* gene evolution may lead to such observed different location. Because all CNSs are homologous, the numbers of intragenic CNSs should be similar across species. However, there is variation in the number (Table A3.1 in Appendix 3), implying that some intragenic CNSs in one species are intergenic in another species. This suggests that inter- or intragenic location of CNSs may not affect their function. Also, I found that intergenic CNSs are located far away from genes (Figures 3.2B and A3.3B in Appendix 3), suggesting that proximity may also not be important. However, I found evidence for conservation of CNS-gene distance. The conservation of CNS-gene distance may be because of the looping structure of DNA. In the loop, only certain regions could be brought in contact with the promoter regions. A regulatory element should therefore sit on such genomic region that could be easily brought in contact with its gene promoter for effective regulation. While regions that are too close may be difficult to bend, the location of the CNS inside or outside the gene may not adversely affect the looping. In conclusion, I have shown the importance of CNS genomic location and demonstrated that the CNSs are likely regulatory elements associated with conserved expression of the proximal genes.

CHAPTER FOUR

HETROGENEOUS MODE AND TEMPO OF CONSERVED NONCODING SEQUENCES AMONG FOUR MAMMALIAN ORDERS

4.0 Chapter summary

This chapter assesses the evolutionary dynamics of CNSs across four mammalian orders and evolutionary timescales. Rodents have lost more ancestral CNSs and gained least compared to primates, carnivores and certartiodactyls. This suggests higher regulatory turnover in rodents. The identified CNSs tend to be under purifying selection, enriched in transcription, development and nervous system related genes, and underrepresented in genes involved in immunity, defense and response to stimulus. Older CNSs are shown to be under stronger constraints. Some CNSs are shown to be in non-homologous positions with respect to gene body, suggesting that whether a CNS is inside or outside a gene body is not always important. The content of this chapter was already published in Babarinde and Saitou (2013).

4.1 Introduction

Conserved noncoding sequence (CNS) analyses have been proved to be computationally powerful in the detection of regulatory elements (Hardison 2000; Levy et al. 2001). Although some noncoding messenger RNA (mRNA) sequences have been found to be conserved, Hemberg et al. (2012) reported that it is four times more likely that a conserved noncoding island is a regulatory element than a noncoding mRNA. ChIP-seq has been reported to be accurate in predicting enhancer activity (Visel et al. 2009). Schmidt et al. (2010), focusing on two transcription factors in liver tissues of five vertebrate species for ChIP-seq analysis, reported that very few regulatory elements are shared by all the species used. However, a mouse ChIP-seq study that examined five transcription factors in 19 tissues and cell types of mouse shows that more than 70% of CNSs function in gene regulation (Shen et al. 2012).

These reports suggest that CNS functions could be specific to tissue, cell type, transcription factor, and/or species. In fact, Shen et al. (2012) clearly demonstrated that the regulatory elements recovered increases with number of tissues and cell types. Some regulatory elements are not conserved (Schmidt et al. 2010), whereas others are conserved. Therefore, the analysis of CNSs should give an idea of the shared regulatory elements. Although Meader et al. (2010) reported high turnover in mammalian functional sequences, many of CNSs are conserved over long evolutionary time (Woolfe et al. 2004) and some are even more conserved than the coding regions (Bejerano et al. 2004; Katzman et al. 2007; Takahashi and Saitou 2012). This is in concordance with the result of several studies that both conserved and nonconserved regions can function as regulatory elements (Chen et al. 2008; McGaughey et al. 2009; Schmidt et al. 2010; Shen et al. 2012).

These reports showed that although some important gene regulations are indispensable and hence conserved, some are experiencing higher turnover rates and thus they are less conserved over a long evolutionary time. Previous reports have shown that arrays of ultra-conserved noncoding regions span through the key developmental genes in vertebrate genomes, and those ultra-conserved regions have strong positive positional correlation with genes encoding transcription factors (Sandelin et al. 2004; Woolfe et al. 2004). Therefore, CNSs conserved in all members of a lineage are functionally important for the lineage. Among the lineage shared CNSs, a subset of CNSs that are unique to the lineage might be functionally important for the lineage-specific features. Such CNSs that are conserved or lost in an order but not in any other outgroup might hold the key to explain the phenotypic diversity among orders.

It is possible that two species would have alignable sequences, not because the sequences are under functional constraint, but because they diverged recently. One of the important tasks in CNS analyses is setting the appropriate thresholds. It is critical to differentiate sequences that are under real selective constraints from those that have simply not had enough time to accumulate enough mutations that will make them distinguishable. For example, more than 95% of human genomes can be aligned to the chimpanzee genome of which only about 10% has been reported to be under selective constraints (Meader et al. 2010; Ponting and Hardison 2011) although the ENCODE Project Consortium (2012) reported a much higher proportion (80%) of biochemically functional elements in the human genome (see Graur et al. 2013 on the ENCODE paper). One way to handle this task is to use more distantly related species such as human and fugu (Woolfe et al. 2005) or elephant shark and some other vertebrates (Lee et al. 2011). With more distantly related species, any sequence conserved over such a long evolutionary time must be functionally important. The other option is to set a threshold that will filter off hits that are not under functional constraints. In setting the threshold, length and percent identity are usually considered. For example, ultra-conserved elements are 100% identical over at least 200 bp length among human, mouse, and rat genomes (Bejerano et al. 2004). Some of the other criteria that have been used are 70% over 100 bp (Duret et al. 1993; Lee et al. 2011), 95% over 50 bp (Sandelin et al. 2004), 95% over 500 bp (Janes et al. 2011), and 98% over 100 bp (Takahashi and Saitou 2012). In this option, however, we have to take cognizance of the different evolutionary rates and the divergence time of the species. Species with higher substitution rates and those that diverged earlier would have lower percent identity.

Takahashi and Saitou (2012) previously compared CNSs of primates (human and marmoset) and rodents (mouse and rat) and found various differences on CNSs and their flanking protein-coding genes. The stringency of the threshold ensured that only sequences with extremely high selectively constraints are studied. However, functional elements that are not highly conserved would be excluded. In this study, using less stringent but reasonable thresholds, I compared genome sequences of five primate species, three rodent species, three carnivore species, and three cetartiodactyl species. I adjusted for the differential evolutionary rate and divergence time to define CNSs that are under functional constraint. I defined a CNS of one mammalian order as a noncoding part of the genome with at least 100 bp length and the percent divergence similar to or higher than that of protein-coding genes of that order. In setting divergence thresholds, I used whole-coding gene sequences as well as third codon-skipped coding region sequences. These criteria are different from Takahashi and Saitou (2012) who used 98% identity over 100 bp. I discovered that the tempo and mode of CNS evolution differed from order to order among mammals and that recent and more ancestral CNSs are under different constraints.

4.2 Materials and methods

4.2.1 Homology Search

The repeat-masked genomes of 24 species were retrieved from the Ensembl genome database, except the sheep genome (oviAri1) that was retrieved from the UCSC genome. The species used are listed in Table A4.1 in Appendix 4, and their phylogenetic relationship is shown in Figure A4.1 in Appendix 4. The genomic coding coordinates were retrieved from Ensembl biomart and UCSC table browser. For most of the species, the genome coverage is at least 6×. To ensure unbiased comparison, I selected the species such that the most distantly related species within every order diverged from the reference species 50–60Ma. The coding regions of all the species were masked. I focused on four different mammalian orders; Primates, Rodentia, Carnivora, and Cetartiodactyla. For each order, a reference species was selected based on the quality and availability of genome information. Human, mouse, dog, and cow were used as reference genomes for primates, rodents, carnivores, and cetartiodactyls, respectively.

After masking the coding sequences in each genome, I searched for the sequences that are conserved in each member of a lineage, using the reference genome as the query. BlastN 2.2.25+ (Altschul et al. 1997) was used for whole genome pairwise homology search. The thresholds used were e value of 10^{-5} and the database size of 3×10^{9}. Nonchromosomal sequences (such as mitochondrial DNA, unmapped DNA, and variant DNA) were not included. If two hits are completely overlapping, the shortest hit is discarded. The threshold percent divergences were set for each group (see details later). Hits above the threshold were first retrieved in as much as they are at least 100 bp long. For the remaining hits, I searched for the core in the alignment with highly conserved regions of at least the threshold percent identity and 100bp length by using sliding windows. This procedure ensures that only alignments of at least the threshold percent identity and 100 bp long in the pairwise search between the reference genome and other members of the group were retained. These conserved sequences are expected to be under functional constraint. Regions of the reference genome that are conserved in all the members of the lineage are potential “group-common” CNSs. Regions of any CNS that overlap RNA gene, pseudogene, or the region that contains masked region is discarded. The resulting CNSs, irrespective of their presence in other species, are referred to as group-common CNSs. I then searched all other species to examine whether these group-common CNSs of one group have homologous sequences. By discarding regions conserved in nonmembers, I detected regions that are unique to that group if they are at least 100 bp long. CNSs thus obtained are common to all members of the group but not found in any nonmember species. I refer to these as “group-unique” CNSs.

4.2. 2 Setting the percent identity threshold

It is important to differentiate between homologous regions that are under functional constraint and those that are similar because they have not had enough time to accumulate enough mutation to distinguish the sequences. This is especially important because this study is on lineage-specific CNSs that diverged around 50–60 Ma. As protein-coding genes are under functional constraint, I decided to use gene-based approach to set the threshold. I did not use protein percent identity because this analysis is nucleotide based. I first considered using nonsynonymous substitution and obtained the values for genes with one-to-one correspondence between the reference species and the most diverged species from Ensembl biomart. However, the standard deviation (SD) for nonsynonymous substitutions is very large (see Table A4.2 in Appendix 4) with some genes having very high values. I then considered setting the threshold using coding sequences. I retrieved the cDNA sequences of one-to-one (with one-to-one correspondence in Ensembl biomart) orthologous protein-coding genes for the reference and most diverged species of each group from Ensembl biomart (Vilella et al. 2009). I used the longest transcript for each gene in each species. For this estimation, I used human and marmoset for primates, mouse and guinea pig for rodents, dog and cat for carnivores, and cow and pig for cetartiodactyls (see Figure A4.1 in Appendix 4). I used BlastN search to calculate the percent identity of the nucleotide sequences in the two species. If there is more than one local alignment in a gene pair, I used the most conserved alignment. Because of the e-value threshold in the homology search, poorly conserved alignments are filtered off. This filtering results in lower SD.

Nucleotide divergences of the coding regions were normally distributed ($P < 10^{-10}$) in all lineages. The test for normality is an omnibus test that combines skew and kurtosis test using Scipy package (Jones et al. 2011). The mean divergence value was significantly lower than the mean value of synonymous substitutions and the genomic average (t-test $P < 10^{-20}$) but higher than the mean value of nonsynonymous substitutions (t-test $P < 10^{-15}$) in all lineages. This is reasonable because although most synonymous sites are evolving neutrally, nonsynonymous sites are mostly under selective constraint. I also compared the mean divergence value of the coding regions with the average divergence of all gene and repeat-masked noncoding regions. To do this, I did homology search using BlastN with the gene and repeat-masked genome of the reference species as the query and most diverged species as the subject using the threshold e-value of 10^{-5}. I used only regions with no duplicates in any of the species. The mean divergence value of the coding regions was significantly lower than the average genomic noncoding divergence (t-test $P = 0$) in each lineage. Finally, I considered the proportion of the mean divergence value of the coding regions to the nonsynonymous substitutions (Table A4.2

in Appendix 4). This value is similar in each lineage (~0.06), suggesting that the mean divergence value of the coding regions is a reasonable threshold. I therefore focused on CNSs that have at most the mean divergence value of the protein-coding genes of the most diverged species to the reference genome. I subsequently refer to this threshold as “whole coding threshold.”

As majority of the third codon sites are synonymous, I decide to set more stringent thresholds using the alignment of coding sequences without using third codons. I obtained the coding regions for the reference genome and the most diverged species for each lineage. I removed bases on third codon positions and concatenated the remaining sequences. I then aligned the concatenated third codon-skipped sequences. The means of percent divergence for each lineage were used as thresholds. These are subsequently referred to as “skip3 thresholds.” Skip3 thresholds are therefore more stringent than whole coding thresholds (Table A4.3 in Appendix 4). For each threshold, only sequences with at least 100 bp were considered. It should be noted that these criteria are different from that of Takahashi and Saitou (2012).

4.2.3 Retention of ancestral CNSs

The abundance of CNSs in a group is partly a result of the retention and loss of ancestral CNSs. This might be an important force in lineage-specific evolution. To study the retention of ancestral CNSs, I used chicken as a basal species. This is because birds have been reported to have sequences closer to the ancestral genome of amniotes (Bourque et al. 2005). I did independent homology search for genomes of human, mouse, dog, cow, African elephant, opossum, and platypus, with the chicken genome as query. Portions of hits overlapping any known gene were discarded. Using the whole coding and the skip3 thresholds (see Table A4.3 in Appendix 4) for sequences with at least 100bp length, I obtained the pairwise CNSs between chicken and each of the seven species. To obtain the total picture of amniote ancestral CNSs, I made a union set of all the CNSs found in all the species used by merging overlapping hits. These CNSs are the total amniotic ancestral CNSs that are still retained in chicken. I then found CNSs lost at each phylogenetic branch starting from the ancestral CNSs.

To investigate the dynamics of more recent CNSs (those found in eutherian mammal common ancestor), I used the genome sequences of African elephant, which is the immediate outgroup species in my analysis as the reference genome, and searched for CNSs in human, mouse, cow, and dog as representative species for each order. Use of African elephant gave more CNSs because many CNSs that evolved after the split of chicken and mammals are included. I repeated the procedure above to obtain the number of CNSs lost in each lineage.

4.2.4 Phylogenetic tree reconstruction

I extracted tetrapod common CNSs in all the 24 species using whole coding thresholds. For each CNS, I got all the orthologs in all the species and aligned the CNSs using ClustalW (Larkin et al. 2007). These alignments were concatenated and blocks with gaps were removed. A Neighbor-Joining tree (Saitou and Nei 1987) was constructed using MEGA version 5

(Tamura et al. 2011).

4.2.5 Conservation levels and guanine–cytosine contents of flanking regions of CNSs

I extracted CNSs together with 1,500-bp upstream and downstream flanking sequences and then aligned the sequences using BlastN. For each alignment, I made sliding windows of 50 bp and a step size of 20 bp starting from 30 bp inside the CNSs and calculated the percent identity in each window. I then calculated the average of the mean of the percent identity for each window. I also calculated the average percent identity of 100 bp in the center of the CNSs. For computation of guanine–cytosine (GC) contents, I similarly obtained CNSs with the 1,500-bp upstream and downstream. I made sliding windows of 200-bp and 10-bp step sizes, starting from 50 bp into the CNSs for calculating the mean GC content values for each window.

4.2.6 Single nucleotide polymorphism and derived allele frequency analyses

I downloaded human single nucleotide polymorphism (SNP 135) database from Ensembl site. I used all SNPs as well as the single nucleotide variants (SNVs) and found the coverage of SNPs in CNSs and random sequences. For derived allele frequency (DAF) analysis, I retrieved Hapmap SNP frequency data of Yoruba population in Ibadan, Nigeria, from UCSC table browser. The ancestral alleles of SNPs overlapping the CNSs or random sequences were determined using chimpanzee sequences.

4.2.7 Gene ontology analysis

I used a modified form of closest gene model for the gene ontology analysis. For a CNS to regulate the closest gene in a reference species, I assumed that in another species, the ortholog of the gene must be the closest to the putative ortholog of the CNS. I downloaded orthologous genes for the reference species and the most distantly related species in the group from Ensembl biomart (Vilella et al. 2009). For each group-unique CNS, I retrieved the list of genes found 1Mbp upstream and downstream. Lettice et al. (2003) and El-Kasti et al. (2012) reported 1Mbp regulation and Vavouri et al. (2005) reported that focusing on 1 Mb range is suitable to obtain the likely target genes. I matched the orthologous genes to each homologous CNS pair and calculated the average distance, defined as the sum of the gene-CNS distance in species 1 and gene-CNS distance in species 2 divided by 2. The orthologs with the shortest average distance are considered as the likely target genes for the homologous CNSs. I checked the functional classification of the tetrapod common CNS-associated genes using PANTHER 7.0 software (Thomas et al. 2003). Because of the limited gene number in the PANTHER database, I manually checked the enrichment of genes using the binomial test as described in PANTHER. Statistical analyses were done using R (R Core Team 2013) and Scipy (Jones et al. 2011).

4.2.8 Genomic Distribution

I downloaded gene coordinates as well as the corresponding 5′ untranslated region (UTR) and 3′ UTR coordinates for the reference species from Ensembl Biomart. For this analysis, I rely almost exclusively on Ensembl gene annotations. Although I used Ensembl build 66 for other species, I used Ensembl build 70 for dog because the gene annotation of dog has recently been improved using cDNA sequences from several tissues. I extracted the promoter coordinates from the gene coordinates. I defined promoter region as the region within 1,000-bp upstream of transcription start site. I then found CNSs that are located on UTR, promoter, intronic, and intergenic regions. To see whether the genomic location of orthologous CNSs is always constant, I first searched whether any of the CNSs has duplicates in the whole genome. Because CNSs with duplicates may have more than one genomic position, I excluded all the CNSs with duplicates in any of the four reference species. I then mapped the genomic location of single-copy orthologous CNSs in each species.

To have the overall picture of the distribution of CNSs and genes in a chromosome, I used sliding windows of 1Mbp size and step size of 100kbp. To investigate whether there is a correlation between gene density and CNS density, I counted the number of CNSs and genes in each window. I calculated the Pearson correlation coefficient between CNSs and genes for all the bins with at least one CNS and at least one gene. I used only windows with at least a gene and a CNS because some windows, especially, around the centromeres do not have any gene or CNS.

4.3 Results

4.3.1 Lineage common CNSs

This analysis takes the evolutionary rate differences among lineages into consideration (see

Materials and Methods). The results using the whole coding threshold and the skip3 threshold are presented in Table A4.3 in Appendix 4. Rodents have a significantly high percent difference and primates have the lowest percent difference (t-test $P < 10^{-180}$), consistent with Li et al. (1996) and several other studies that showed higher substitution rates in rodents. Many experimentally verified functional elements, such as vista enhancer elements and transcription factor ChIP sequences, have been reported for human. Vista enhancers are highly conserved noncoding regions that have been experimentally confirmed to have regulatory function. Also, DNase clusters have been reported to have regulatory signatures (Crawford et al. 2006). To examine the suitability of the thresholds, I obtained the human sequences of these elements from UCSC table browser. Using BlastN, I searched for the homologous marmoset sequence and calculated the percent divergence of each element. I then found the distributions of the percent divergence between these human elements and the corresponding marmoset orthologs. This analysis shows that my thresholds in primates are reasonable (Figure 4.1). The thresholds are close to the peak of vista enhancers. As the same procedure was used for the four lineages, I assumed that the thresholds are generally reasonable.

Furthermore, considering the average genomic noncoding divergence (Table A4.2 in Appendix 4), alignments of the whole coding threshold divergence levels are less likely to be observed by chance (binomial $P < 0.05$) in all lineages. The chance is even less likely under skip3 threshold (binomial $P < 0.005$; Table A4.3 in Appendix 4). Using the whole coding threshold and the skip3 threshold, as well as minimum length of 100bp, I searched for the lineage common CNSs. They include CNSs that originated before the emergence or in the common ancestor of each lineage. Primates have the highest number of common CNSs (861,183) compared to 148,848, 491,078, and 257,051 CNSs in rodents, carnivores, and cetartiodactyls, respectively, using the whole coding threshold. The lineage common CNSs covered 5.52%, 1.28%, 2.14%, and 2.12% of human, mouse, dog, and cow chromosome-mapped genomes, respectively. When more stringent skip3 thresholds were used, the numbers of CNSs retrieved were 323,351, 62,985, 220,009, and 130,381 for primates, rodents, carnivores, and cetartiodactyls, respectively (see Figure A4.2 in Appendix 4). It is important to note that five primate species were used compared with three each in other lineages (Figure A4.1 in Appendix 4). Because of higher turnover rate of CNSs (Meader et al. 2010) and high independent losses of conserved noncoding DNA elements (CNEs; Hiller et al. 2012), lineages with higher number of species should have fewer CNSs. However, primate CNSs are more than 5-fold that of rodents and about 3-fold that of cetartiodactyls. The carnivore CNSs are about twice that of cetartiodactyls and 3-fold that of rodents despite similar genome sizes. Even with more stringent thresholds (e.g., coding divergence minus 1 SD or half of coding divergence) between the reference genomes and the most diverged species, although the differences become smaller, the patterns remain essentially the same (see Figure A4.2 in Appendix 4). However, the differences between the numbers of primate and carnivore CNSs become smaller.

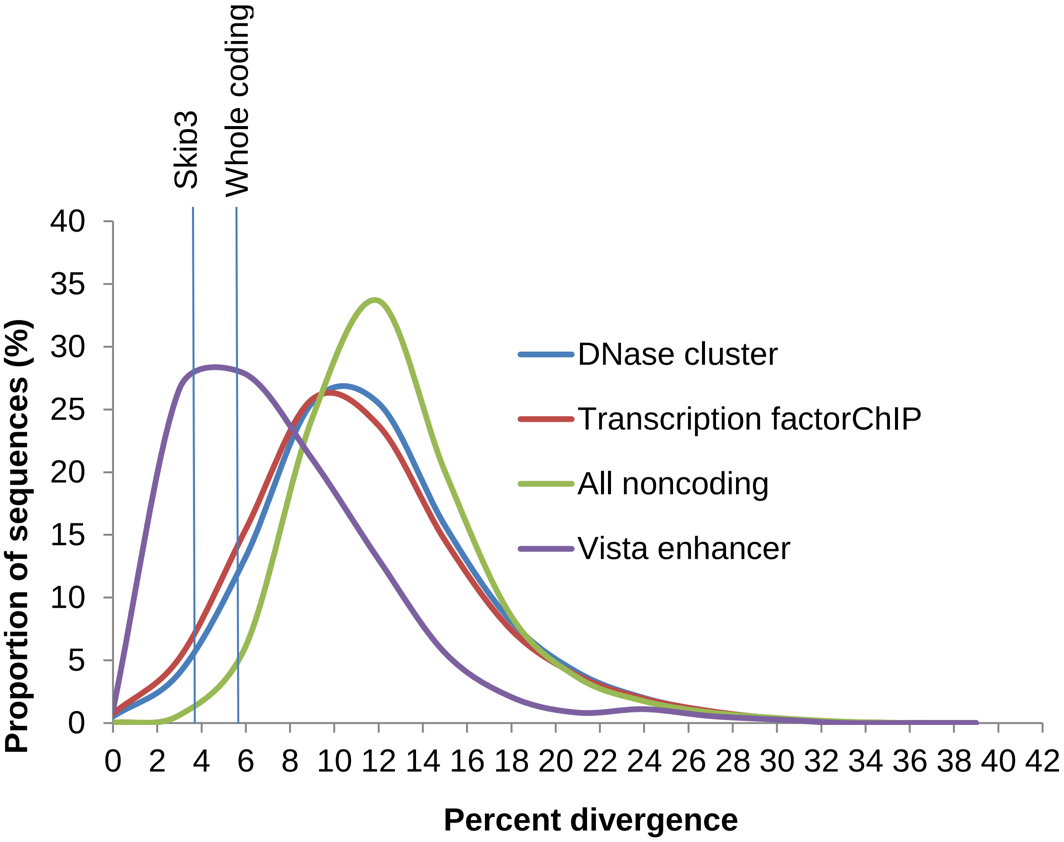


Figure 4.1: Assessing the suitability of percent divergence thresholds. The two thresholds are comparable to the divergence of vista enhancer elements and higher than the average of all alignable noncoding sequences.

The differences in the abundance may be due to the difference in genome data quality. I relied on the genome data coverage information available in the database. All genomes have more than 6× coverage. Although human has the best genome quality, it is important to note that CNSs are conserved in all members of a lineage, and therefore, the genome qualities of all members contribute to the number of CNSs. Although mouse and rat have very good genome qualities, rodents have the least number of CNSs. I further checked the effect of genome quality by searching for the number of CNSs conserved with chicken in all the species. In all the lineages, the number of elements conserved between each species and chicken is similar (data not shown). Human and dog have similar number of CNSs conserved with chicken (see Table A4.4 in Appendix 4) but the number of primate common CNSs is higher than that of carnivores. This suggests that genome quality and repeat masking database do not significantly influence the results.

I next checked the distribution of CNSs by length. I compared the abundance of CNSs at various length categories using the whole coding threshold. This result shows that primates have the highest number whereas rodents have the least, irrespective of the length criterion used (see Figure A4.3 in Appendix 4). The average lengths of CNSs are 195, 219, 228, and 204 bp for primates, rodents, carnivores, and cetartiodactyls, respectively, though the standard deviations are high. This result shows that although primates have higher proportion of short CNSs, irrespective of the length threshold used, primates have more CNSs and rodents have the least. Therefore, choosing shorter length threshold would not alter the pattern of the result.

4.3.2 Phylogenetic origin and abundance of CNSs

The difference in the abundance of CNSs obtained across lineages may be due to the difference in the amount of CNSs that were gained or lost at each branch. I therefore checked the CNSs that are unique to each lineage. These are sequences that are reasoned to have gained new functional constraint on that branch and all the extant members still retain the function. With the whole coding threshold, the amount of primate-unique CNSs (52,124) is more than 100-fold of 353 rodent-unique CNSs (Figure 4.2). Among 861,183 primate-common CNSs, only 52,124 were found to be primate-unique and only 1,779 are shared by all the species used. Of the total 37,709 eutherian common CNSs, only 12,378 evolved in eutherian common ancestor. This dynamics indicates that many CNSs are of older origin but have been lost in some species, making it difficult to trace their evolutionary origin. If I consider the number of shared CNSs, the more diverged the species included, the less the number of CNSs that can be retrieved. Although primates have the highest number of shared CNSs, 73,641 euarchontoglire (primates, rodents, and rabbit) common CNSs are lower than 111,705 laurasiatheria (carnivores, cetartiodactls, and microbat) common. The number of CNSs conserved in all eutherian species (euarchontoglire, laurasiatheria, and African elephant) is 37,709. When I included opossum and platypus, the number of CNSs became 11,693. These patterns are similar when the skip3 threshold was used. This is expected due to evolution of new CNSs as well as high turnover rate of CNSs.

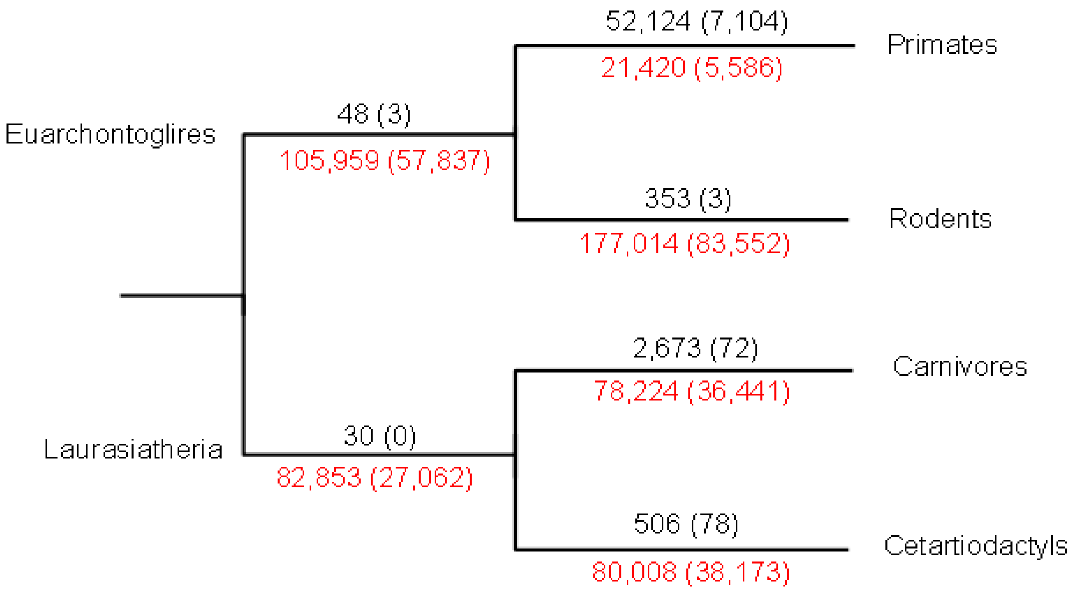


Figure 4.2: The phylogenetic gain and loss of CNSs. The values on a branch (in black font) are the numbers of CNSs gained on the branch, whereas the values under a branch (in red font) are the number of CNSs lost on that branch with African elephant as the reference. For each point, the values are the numbers if whole coding thresholds were used, whereas the values in parentheses are the numbers if skip3 thresholds were used.

What might be responsible for the high number of CNSs in primate lineage? One possibility is that there were many duplications of the CNSs in the primate common ancestor. To check this, I searched for duplicates in primate unique CNSs. Out of 52,124 primate unique CNSs retrieved using the whole coding threshold, 49,888 (96%) are single copies with no duplicates. I then checked how old the primate unique CNSs are. I did not include mouse lemur, a primate which diverged earlier from human, in this analysis due to lower genome quality and divergence time consideration of its genome. I therefore checked how many of the primate unique CNSs have copies in its genome. I found 18,308 CNSs with hits in the mouse lemur genome, out of which 9,528 are 100 bp or longer. This value is much higher than the unique CNSs in other lineages, suggesting that the observed higher number of CNSs in primates is not only solely because of the divergence time but also because of functional constraints. On the other hand, most of the CNSs evolved after the emergence of mouse lemur are suggested to acquire new functional constraints.

4.3.3 Lineage-specific loss of CNSs

The second contributing factor to the difference in number of CNSs in each lineage may be the retention or loss of ancestral CNSs. The loss of ancestral CNSs could happen through two processes. The first process is through high-sequence divergence. The constraint on a sequence might be relaxed in a lineage from loss of function or adaptive evolution. In this case, although the homologous sequence is retained in the genome, the sequence has diverged beyond recognition. The second process that may lead to CNS loss is sequence deletion as reported by Hiller et al. (2012). In this case, the whole region or a part of the region might be deleted from the genome. By lowering the threshold, I could retrieve some of the CNSs that have been lost by sequence divergence. However, not being able to retrieve lost CNSs, even after lowering the threshold, may not necessarily mean they were lost by deletion. Some CNSs might have gone through high-sequence divergence such that sequence similarity would have been completely lost or lost to the level that it cannot be identified by homology search.

In my definition of lost CNSs, the specific mechanism of loss is not considered. A CNS may thus have been lost through high-sequence divergence or deletion. I assume that, in either case, the function would have been modified if not completely lost. I therefore checked the rate of loss of amniote ancestral CNSs conserved between chicken and at least one other mammalian species. For this analysis, I used eight species, including a representative species from each of the four lineages (see Figure 4.2; Table A4.4 in Appendix 4; Materials and Methods). As chicken and other species have lost some CNSs, complete set of amniote ancestral CNSs cannot be accounted for. However, assuming an unbiased parallel loss between chicken and the mammalian species used, this analysis would give a true pattern of lineage-specific rate of loss. Chicken has 66,210 CNSs, conserved in one or more of the species used and 11,472 conserved in all, implying that 54,738 CNSs have been lost in one or more species or lineages. I then calculated the number of CNSs lost in each branch. Mouse has lost more number of CNSs than any other species (Table 4.1 and Table A4.4 in Appendix 4). In euarchontoglire common ancestor, 7,699 CNSs were lost, whereas 3,833 CNSs were lost after the divergence of euarchontoglire and laurasiatheria and before the divergence of cow and dog. However, 1,861 CNSs lost in the human lineage after the divergence from mouse is lower, and the 13,015 CNSs lost in the mouse lineage is higher than those lost in cow and dog lineages after their divergence. Similar patterns are observed with the skip3 threshold (Table 4.1 and Table A4.4 in Appendix 4). Out of the total 41,465 CNSs in chicken, only 7,889 are conserved in all species. This suggests that the rates of CNS loss are heterogeneous in euarchontoglire lineages (see Figure A4.1 in Appendix 4), with human lineage having a lower rate and mouse lineage having a higher rate.

I then used African elephant as the outgroup genome so as to include more recent CNSs in my focus. The results in Figure 4.2, Table 4.1 and see Table A4.4 in Appendix 4, indicate a very rapid loss of CNSs in a species-specific manner. With the whole coding threshold, out of 439,034 CNSs present before the eutherian common ancestor, only 109,475 (24. 9%) was retained in all the four species. A significant proportion of the CNSs (36.4%) have been lost in three of the four species; 19.3% in two species; and 19.3% in just one species. When the skip3 threshold was used, of the 195,926 CNSs in the eutherian common ancestor, 35%, 21%, 22%, and 22% are found in one, two, three, and all of the four species, respectively. This suggests rapid independent loss events of CNSs. Even when the number of CNSs per unit branch length was considered, the patterns did not change. This suggests that real functions, in addition to evolutionary rate differences, account for the dynamics of CNSs.

Table 4.1: The loss of ancestral CNSs with different thresholds and reference species

	Number of CNSs lost			
	Chicken as reference		African elephant as reference	
Phylogenetic branch	Whole coding	Skip3	Whole coding	Skip3
Primates	1,861	1,243	21,420	5,586
Rodents	13,015	7,240	177,014	83,552
Carnivores	4,806	3,300	78,224	36,441
Cetartiodactyls	6,884	3,882	80,008	38,173
Euarchontoglires	7,699	5,236	105,959	57,837
Laurasiatheria	3,833	1,786	82,853	27,062
Euarchontoglires and Laurasia	1,888	1,476	-	-
African elephant	12,605	7,938	-	-
Eutheria	7,468	5,171	-	-
Opossum	21,348	13,142	-	-
Theria	7,437	3,798	-	-
Platypus	30,514	18,698	-	-

4.3.4 CNSs are under functional constraint

I investigated whether the CNSs of different ages are under similar functional constraint. I compared the conservation level of CNSs that evolved in primate-, eutherian-, mammalian-common ancestor, and those that were found in tetrapod common ancestor. The numbers of CNSs retrieved using whole coding thresholds were 52,124, 12,378, 4,059, and 1,779, whereas the numbers retrieved using skip3 thresholds were 7,104, 2,118, 1,390, and 1,733, for primate unique, eutherian unique, mammalian unique, and tetrapod common, respectively. The focus on primate order was because of higher number of primate unique CNSs. I extracted human and marmoset sequences for each of the classes and calculated the pairwise divergence using ClustalW (Larkin et al. 2007). As shown in Figure 4.3A, percent difference of tetrapod common CNSs is the lowest while that of primate unique CNSs is the highest (t-test P $< 10^{-20}$). This suggests that more ancestral CNSs are under stronger constraint than the recent ones.

Is the difference in conservation level due to age alone or the result of indispensability of the function? To answer this, I compare the conservation level of CNSs that have been lost in mouse but conserved among human, marmoset, and chicken (see Table 4.1). If age is the sole determining factor, these mouse-lost CNSs should have conservation level similar to that of tetrapod common CNSs. However, the CNSs lost in mouse are not significantly different from primate unique CNSs, even though primate unique CNSs are more recent. This suggests that the difference in conservation levels is due to the indispensability of the function. The lineage-specific evolution and loss of CNSs are major players in shaping the abundance of CNSs in a lineage (Figure 4.2; Table 4.1). Because recently evolved CNSs and those that have been lost in some lineages are under relatively lower constraint (Figures 4.3 and 4.4A), too stringent threshold would imply that most of the CNSs retrieved would be ancestral ones, and therefore lineage difference will be reduced. Indeed, that is the observation, especially between primates and carnivores, when I used more stringent thresholds (coding divergence minus 1 SD or half of coding divergence; see Figure A4.2 in Appendix 4).

As primate unique CNSs are under lowest functional constraint, I checked whether the conservation level is different from that of randomly retrieved sequences. I first randomly retrieved noncoding sequences of the same number and lengths as the primate unique CNSs retrieved using the whole coding threshold. I did homology search and used sequences with unique identifiable homolog. Only 16,934 (34%) have unique identifiable homolog of at least 100 bp long. The average percent difference of randomly retrieved sequences (12%) is significantly higher than 5.8% of primate unique CNSs (t-test P = 0).

I also checked the conservation level of the CNS flanking regions. Figure 4.3B (for the whole coding threshold) and see Figure A4.4 in Appendix 4 (for the skip3 threshold) show that CNSs are under stronger conservation level, compared with the flanking regions. Also, the conservation of the flanking regions of tetrapod common CNSs is higher than that of the primate unique CNSs as predicted from Figure 4.3A. Although more ancestral CNSs are peaks of long conserved regions, younger CNSs seem to be peaks of poorly conserved regions.

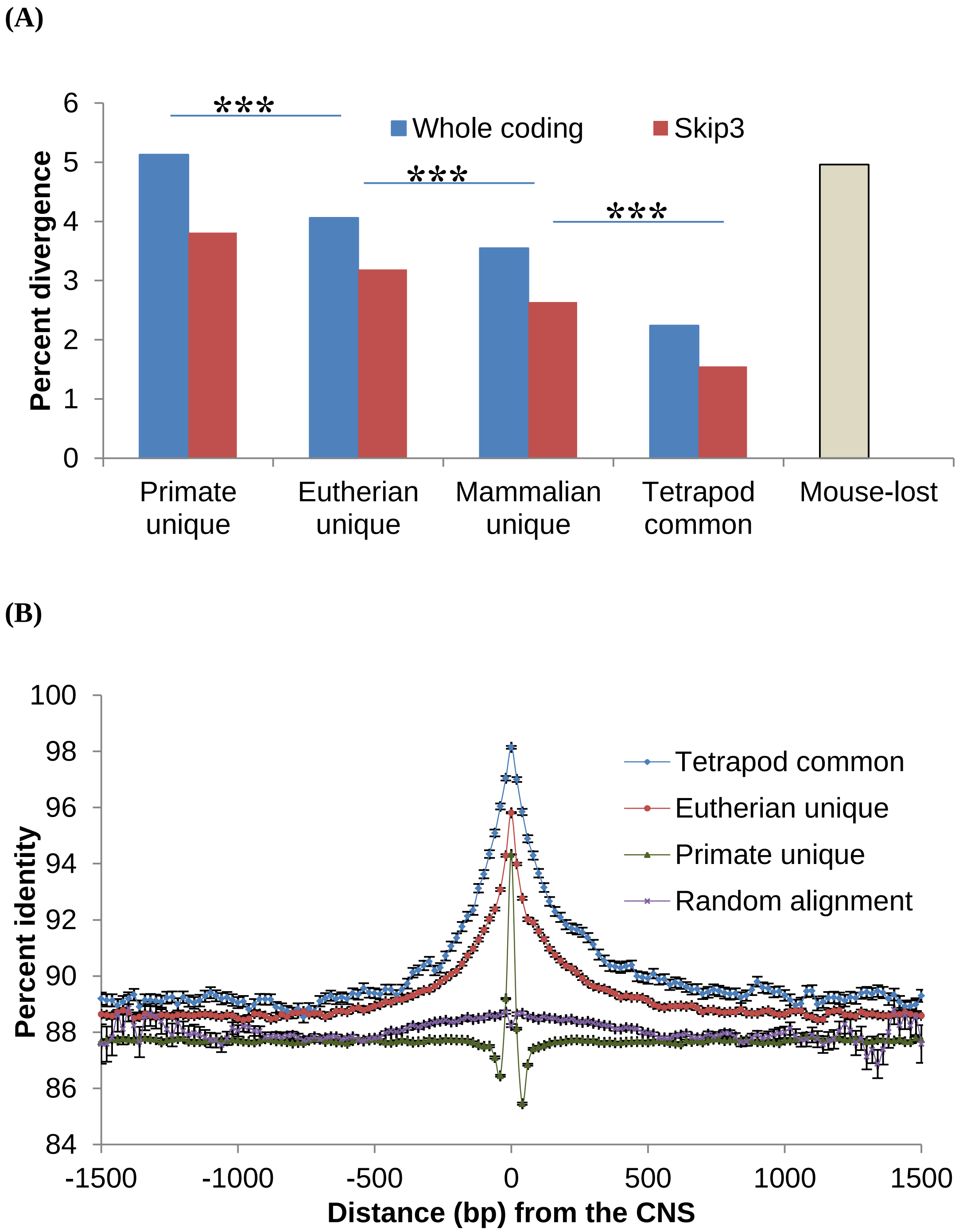


Figure 4.3: The conservation levels in and around CNSs. (A) The divergence levels of CNSs (***t-test P-value < 10−20). (B) The conservation levels of flanking regions of CNSs with whole coding thresholds. Point 0 is the average percent identity of 100 bp at the center of the CNSs, whereas other points are the average of 50-bp windows moved at 20-bp steps starting from 30pb inside the CNSs. The bars are the standard error of the mean for each window.

These patterns are not observed in random sequences. Note that for random sequences, I used unfiltered alignments of at least 1,200 bp long. Because the sequences tend to contain CNSs, there seems to be slight elevation around the CNSs. Even with that, the center is not the peak.

As another measure of functional constraint, I examined the coverage of SNPs and SNVs as found in an Ensembl database of human SNP. Primate unique CNSs had higher percentage of SNPs and SNVs, and tetrapod common CNSs have lower SNPs and SNVs as expected (Figure 4.4). Again, mouse-lost CNSs are similar to primate unique CNSs supporting the hypothesis that the indispensability of the CNS, rather than just the age, determines the strength of functional constraint. It is also important to note that CNSs cover less SNPs and SNVs, compared with random sequences (t-test $P < 10^{-15}$).

Another test of functionality of a sequence is DAF analysis (Drake et al. 2006; Takahashi and Saitou 2012). The frequency of derived alleles of functional regions is expected to be lower than the genomic average because of purifying selection. The result (see Figure A4.5 in Appendix 4) supports that those CNSs are under purifying selection. Higher proportion of CNSs has lower derived alleles than random expectation. For example, for alleles with frequency of <0.1, the number of alleles with frequency of <0.1 in CNSs is significantly higher than random expectation (binomial $P < 10^{-15}$). On the other hand, at higher frequencies, CNSs have slightly lower proportion than the genomic average.

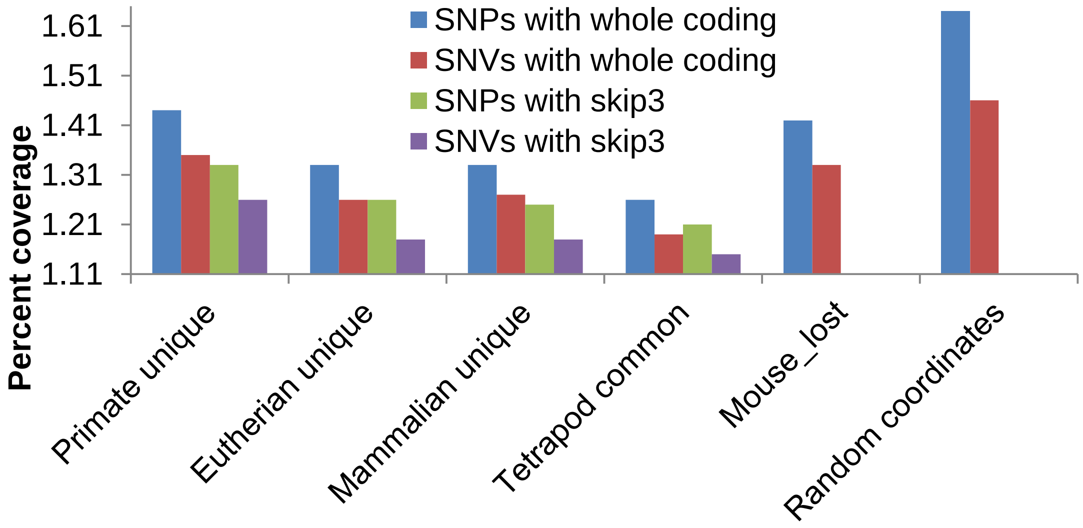


Figure 4.4: The SNP coverage of CNSs. The average numbers of SNPs found in 100 bp of CNS are presented for each age category. Complete SNP data as well as SNV data were used. Random coordinates of number and lengths similar to whole coding primate unique CNSs were used (***t-test P-value < 0.001; **t-test P-value < 0.005).

I checked the nucleotide composition of the CNSs, with emphasis on GC content. The GC content of the CNSs is significantly lower than that of protein-coding regions (t-test $P < 0.001$). Except for the primate unique CNSs, GC contents of CNSs unique to the other three mammalian orders are lower than that of genomic averages reported by Karro et al. (2008). This might be because primate unique CNSs evolved recently and did not have enough time to lower their GC contents. As in the conservation level and SNV coverage, the GC content tends to correlate with the age of the CNSs (Figure 4.5). This result suggests that the GC content of CNSs decays as CNSs become old. In fact, Table A4.5 in Appendix 4 shows that GC→AT substitutions are more than AT→GC substitutions.

In addition, I conducted sliding window analysis to examine the GC content of the flanking regions of the CNSs. I focused on mammalian unique CNSs with the lowest GC content (Figure 4.5). With the sliding windows of size 200 bp, 10 bp step size starting from 50 bp inside the CNSs, I found a sharp decrease toward the CNSs (Figure 4.5B). Similar pattern was observed when I considered all skip3 primate common CNSs (see Figure A4.6 in Appendix 4). The observed decrease in the GC contents toward CNSs might be related to nucleosome occupancy as reported by Gupta et al. (2008).

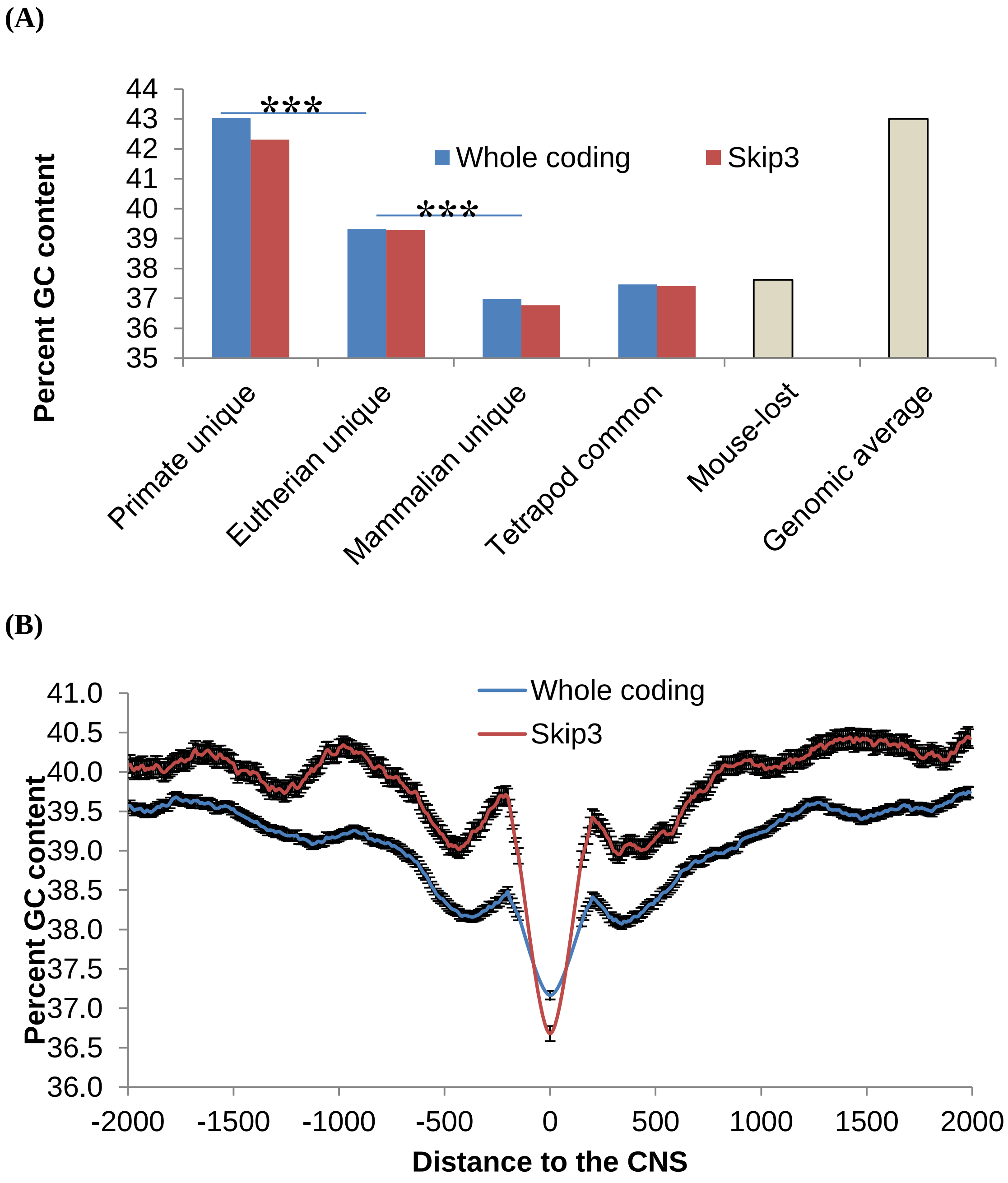


Figure 4.5: GC contents of CNSs and flanking regions. (A) The GC contents of CNSs. The genomic average is from Karro et al. (2008) (***t-test P-value < 0.001). (B) Using sliding windows of 200-bp size and sliding steps of 10 bp, the percent GC contents of the mammalian unique CNSs and flanking regions were computed. Position 0 is the 100 bp in the center of the CNSs and the first window starts from 50 bp into the CNSs.

4.3.5 Genomic Distribution

The genomic distribution of the CNSs varies across lineages. This is especially obvious when considering CNSs found in intergenic and intronic regions in each lineage (Figure 4.6 and Figure A4.7 in Appendix 4). Primates and rodents (Euarchontoglires) have higher proportion of intronic CNSs than random expectation (binomial $P < 10^{-100}$). On the contrary, carnivores and cetartiodactyls (Laurasiatheria) do not have such a high proportion of intronic CNSs. The distribution of carnivore and cetartiodactyl CNSs is very close to the random expectation, which might be due to the quality of genome data, especially of cow. Primates also have notably higher promoter unique CNSs. Genomic distribution of recently evolved CNSs and older ones are different, especially when I focused on UTR CNSs. The UTR proportion of the eutherian common CNSs in primates and rodents agrees well with Siepel et al. (2005), who reported that vertebrate highly conserved elements were associated with 3′ UTR of regulatory genes. My analysis suggests that the locations of functional elements are dynamic even among eutherian mammals.

Figure 4.6 shows that the proportions of UTR CNSs in unique and ancestral CNSs are different. For eutherian common CNSs in primates, for example, 7% of the CNSs are located in the UTRs, compared with 2% of the unique CNSs located in the same location (binomial $p < 10^{-100}$). The same pattern was observed across the four orders. This suggests that more ancestral CNSs are more associated with UTRs. Apart from age-related dynamics in genomic distribution of CNSs, I also observed species-related dynamics (Figure 4.6).

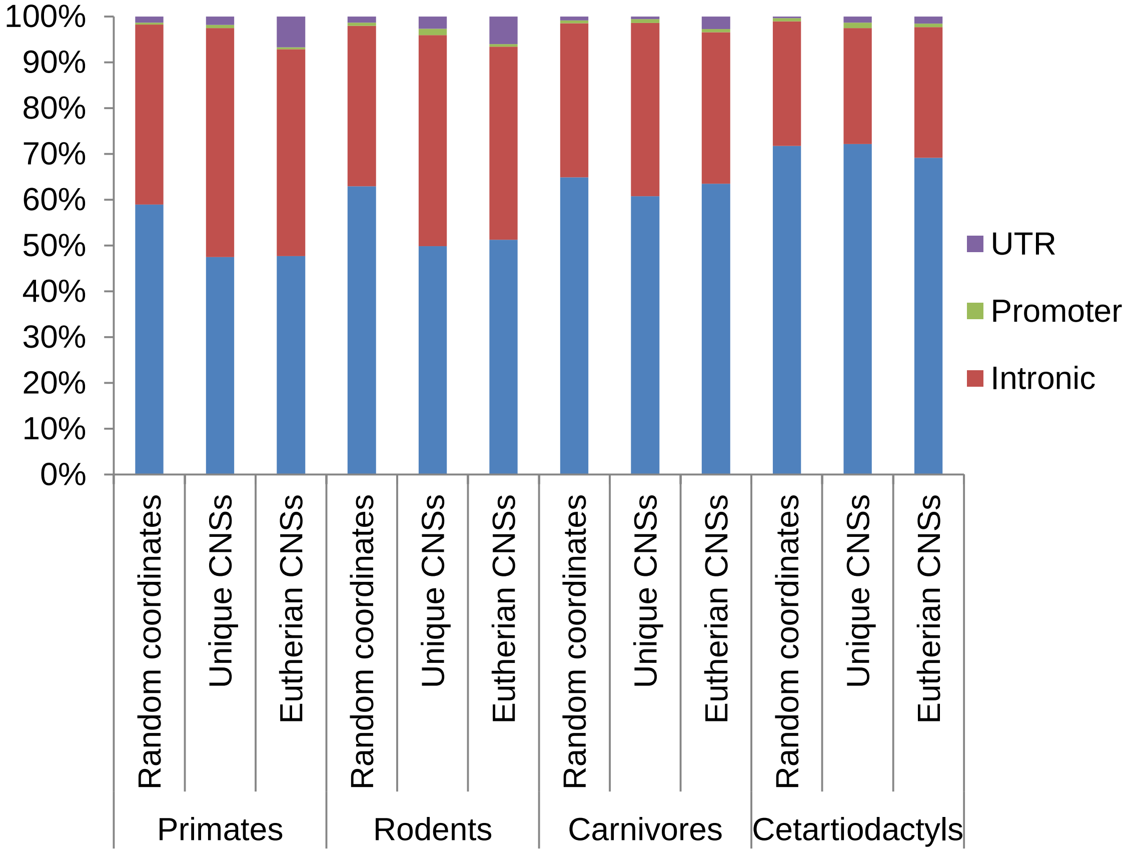


Figure 4.6: The genomic distribution of the CNSs. Using whole coding thresholds, percentage of CNSs found in each genomic region for each lineage and category are presented. Eutherian CNSs are the 35,906 single-copy CNSs conserved in all the four lineages. For random coordinates, same number and lengths of eutherian CNSs were randomly selected from noncoding sequences of each reference species.

To examine whether ancestral CNSs are always fixed in location or may be "relocated," in terms of location with respect to genes, I considered the genomic location of eutherian common CNSs, using CNSs with single copy, such that CNSs with duplicates in any of the species representing each lineage are not included. It turned out that the location of orthologous CNSs is not always fixed (Figure 4.5 and see Figure A4.8 in Appendix 4; Table 4.2 and see Tables A4.6 and A4.7 in Appendix 4). Out of 35,906 homologous single-copy CNSs considered, 10,701 (about 30%) are located in a different region in one or more species. In human and dog, for example, 19.38% of homologous CNSs are not located in the same genomic location in the two species (Table 4.2). The events of difference in locations between mouse and human genomes are not as many. I acknowledge the fact that this pattern may be caused by the heterogeneity in genome annotation quality. Human and mouse genomes have been reasonably well annotated, and dog annotation has been recently improved with cDNA data. Even between human and mouse, 8% of the CNSs are located on different locations. Interestingly, the closest genes tend to be the same even when the actual locations of CNSs with respect to gene body are different.

Table 4.2: The similarity of the genomic location of orthologous CNSs

	Human	Mouse	Dog	Cow
Human		2,883 (8.03%)	6,960 (19.38%)	8,387 (23.36%)
Mouse	33,023		5,951 (16.57%)	7,221 (20.11%)
Dog	28,946	29,955		5,099 (14.20%)
Cow	27,519	28,685	30,807	

The genomic locations (intergenic, intronic, UTR or promoter) of the 35906 single-copy CNSs that are shared by all eutherian species used were compared in the four representative species. The numbers in grey shade are located on the same genomic location while the numbers in the upper part (without shade) are located on different locations.

I further examined whether the CNSs are uniformly distributed over the chromosome by using 1-Mbp windows with 100kbp step size. CNSs and genes are distributed nonuniformly across the chromosomes but not found at all around centromeres (see Figure A4.9 in Appendix 4). This might be because centromeres have many repeat elements. Some conserved regulatory elements have been reported to be found in the gene desert (Ovcharenko et al. 2005), and Siepel et al (2005) reported that vertebrate highly conserved elements are associated with stable gene desert. I therefore examined the correlation between the number of CNSs and the genes in 1-Mbp bin with at least a gene and a CNS. The Pearson correlation coefficients (and P values) are −0.2234 (8.069×10^{-299}), −0.2425 (2.18×10^{-300}), −0.3737 (0.00), and −0.2486 (0.00) for primates, rodents, carnivores, and cetartiodactyls, respectively. The negative correlations in all the lineages indicate that more CNSs cluster where there are few genes (gene deserts). However, considering the proximity of CNSs to genes, more than 90% of group-common and lineage-specific CNSs have the closest gene within 400-kbp upstream or downstream in all lineages except in cetartiodactyls. In fact, more than 95% of CNSs have the closest gene within 700kbp (see Figure A4.10 in Appendix 4).

4.3.6 Functional analysis of the CNSs

Takahashi and Saitou (2012) previously reported lineage-specific highly conserved noncoding sequences (HCNSs), which were associated with some protein-coding genes. As the same primate species were used in both studies, I checked the overlap of the primate sequences found in both studies. Of the total 8,198 HCNSs reported, 6,643 (81%) were retrieved in this study. Among nonoverlapping 1,555 (19%), 1,005 were nonprotein-coding genes that were removed from the present analysis.

To check whether identified CNSs have regulatory functions, I checked for the overlap with reported human transcription factor ChIP sequences from UCSC table and found 166,451 primate common CNSs overlap with previously reported ChIP-seq data. As another signature of regulatory activity, I found that 317,590 overlap with DNase clustered sites (see Figure A4.11 in Appendix 4). I also checked for the possibility of transcription of the identified CNSs using the GENCODE comprehensive gene data from UCSC table. Only 51,261 of the CNSs retrieved with the whole coding threshold were found to overlap any ENCODE gene. It is important to note that GENCODE genes cover about 48% human genome while the ChIP-seq data cover 7.78%. These results suggest that identified CNSs are 20 times more likely to be regulatory elements, compared with genes. In fact, among the 51,261 CNSs found to overlap ENCODE genes, 16,278 also overlap ChIP-seq data. If I checked for more regulatory signatures such as histone modification marks, DNA-hypersensitive sites, or focus on more cell or tissue types as well as more developmental stages, more CNSs may be found to have regulatory signatures. Thus, many of the identified CNSs most likely have regulatory functions. Figure A4.12A in Appendix 4 shows an example of a multiple alignment of an identified CNS. In addition, Figure A4.12B and C in Appendix 4 show examples of identified CNSs that overlap some regulatory signatures and experimentally confirmed brain enhancers, as reported by Viesel et al (2013).

To further investigate the likely functions of the CNSs, I studied the enrichment of the biological process of the genes that are likely regulated by the CNSs using PANTHER. Several studies have reported that CNSs are associated with genes involved in transcription regulation (e.g., Vavouri et al. 2007; Elgar 2009) and developmental genes (Hardison 2000; Levy et al. 2001). I studied the enrichment of the likely target genes of tetrapod common CNSs using PANTHER and considered the biological process of the 20 most enriched ontology groups. The gene ontology result shows that the ancestral CNSs are enriched in genes involved with transcription regulation and development (Table 4.3), suggesting that the CNSs play an important role in phenotypic diversity. It is important to note here that nervous system related genes are also enriched, suggesting that many regulatory elements associated with nervous system are conserved. This observation is consistent with the report of Matsunami and Saitou (2013) that vertebrate paralogous CNSs may be related to gene expression in the brain. PANTHER database uses a limited number of genes. Moreover, I analyzed the enrichment by CNS-weighted genes. I therefore calculated the enrichment and P value under binomial distribution as described by PANTHER for the top three overrepresented terms and top two underrepresented terms (see Table A4.8 in Appendix 4). The three overrepresented terms are transcription, development, and nervous system, whereas the two underrepresented terms are response to stimulus and immune and defense. This observation suggests that CNSs tend to be associated with genes that are under negative selection and underrepresented in genes under positive selection.

4.4 Discussion

Understanding the molecular mechanism underlying the phenotypic diversity observed among species has been of interest to many scientists. Before the invention of high technology that is now available for molecular studies, the taxonomists have classified the organisms according to the phenotypes of each species, classifying species with similar features into the same group. As phenotypes have genetic background, it is expected that each taxonomic group should share some unique genetic features. In fact, many phylogenetic studies have been in agreement with taxonomic classification (e.g., Meyer and Zardoya 2003; Blanga-Kanfi et al. 2009). Therefore, the knowledge of some unique molecular features should shed more light into the understanding of lineage evolution. However, the phenotypic diversity observed among species and taxonomic groups could not be sufficiently explained by mere presence or absence of a particular set of genes (Stern 2000; Wray 2003). This is because many genes are highly conserved in many species. The regulation of spatiotemporal gene expression has long been suggested to be important in phenotypic diversity (Zuckerkandl and Pauling 1965; King and Wilson 1975). Therefore, one way to molecularly explain the phenotypic diversity may be in terms of gene regulation. Not just the presence or absence of a gene but also when, how, and where certain genes are expressed are also important in the evolution of phenotypic diversity observed among and within orders (Zhang and Peterson 2006; Takahashi and Saitou 2012; Wittkopp and Kalay 2012). A genetic theory of morphological evolution states that form evolves largely by mutations in cis-regulatory sequences that alter the expression of functionally conserved proteins (Carroll 2008). In fact, Cretekos et al. (2008) reported that regulatory divergence modifies limb length between mammals. CNSs are therefore good proxy for conserved regulatory elements.

In this study, I defined CNSs as homologous regions with at least 100bp length and average conservation level of the protein-coding genes with one-to-one correspondence. Divergence thresholds were set by using whole coding genes as well as third codon-skipped coding sequences. Although it is possible that some functional CNSs may have less degree of conservation, I assume that only CNSs with threshold or lower divergence levels have a function that is important for the group. I therefore did not discard the possibility of functional sequences with conservation lower than the threshold, as it is known that many regulatory elements are not evolutionarily conserved (Schmidt et al. 2010; Shen et al. 2012). The percent identity threshold minimizes the false negatives and also allows for correction for difference in evolutionary rates among lineages. The conservation level, coverage of SNV, GC content, and overlap with previously reported regulatory signatures as well as DAF analysis support that the CNSs identified are under functional constraint and may have regulatory functions.

To check the orthology of the CNSs, I constructed the phylogenetic tree using concatenated CNSs. If the CNSs are orthologous, I would expect that the species tree should be recapitulated with high statistical support. Indeed, all branches of the tree (see Figure A4.13 in Appendix 4) had 100% bootstrap probabilities, and it had the identical topology with the established mammalian phylogeny (e.g., Beck et al. 2006). This result is consistent with the hypothesis that the CNSs of different species are orthologous. In addition, the tree clearly shows that CNSs can be useful for producing species trees.

Although there are several reports on CNS evolution in vertebrates and tetrapods, my study shows a significant difference in the evolutionary dynamics of CNSs among the four mammalian lineages that diverged <120 Ma. There is an obvious difference in the abundance of CNSs in each lineage. Although primates have more CNSs, rodents have the least CNSs. Also, the rate of gain of function of noncoding regions, detectable as regions uniquely conserved in a lineage, varies across lineages with primates having much more than any other lineages. Takahashi and Saitou (2012) also found that the numbers of HCNSs are different with rodents having more than primates. It is important to note that Takahashi and Saitou (2012) used MEGABLAST for homology search, whereas I used BlastN, which is more sensitive but much slower. Apart from the difference in methodology, especially the definitions of HCNSs (Takahashi and Saitou 2012) versus CNSs (this study), the different pattern may be due to the species included especially in the rodent lineage. I included guinea pig with higher genetic distance to mouse. As I have shown that species-specific loss of CNSs occurs, the number of species included would affect the abundance of CNSs that could be recovered. The loss of ancestral CNSs was higher in mouse than in human. My study also showed that more CNSs originated in the common ancestor of primates compared with that of rodents. These differences may contribute to the lineage-specific phenotypic dynamics.

There are still many uncertain areas in the evolution and actual function of CNSs.

Although there are many functional signatures on the sequences, some experiments involving deletions of such conserved elements yielded viable mice (Nóbrega et al. 2004; Ahituv et al. 2007). These reports raised some concerns about the functions of CNSs and the actual reasons behind their conservations. However, the fact that the deletion of the regions did not produce any phenotype does not necessarily imply that they are functionless. Their functions might be subtle and related to a weak fitness in the wild. As a matter of fact, deletion of some enhancer elements of Drosophila *shavenbaby* genes produced no phenotype under optimal temperatures but under low or high temperatures, non-wild-type phenotypes were produced (Frankel et al. 2010). Although it is possible that some CNSs may have other functions, they are more likely to function as regulatory elements (Hardison 2000; Levy et al. 2001; Hemberg et al. 2012). The fact that the CNSs cluster around genes involved in transcription regulation and development implies that they may contribute to phenotypic differences. There is a possibility that some of the identified CNSs might be some genes that are yet to be annotated. However, the overlap of the CNSs with reported regulatory signatures suggests that many have gene regulatory functions. Attributing more CNSs to more regulatory elements, this study suggests that primates have more shared regulatory elements implying higher complexity in primates compared with other lineages. Although the human genome has about the same number of protein-coding genes as the mouse genome, the higher number of human regulatory elements (inferred from the higher number of CNSs) suggests that primates have complex shared gene regulatory system that deals with nervous system and brain development. This is reasonable because I have shown CNSs to be enriched in genes associated with the nervous system as previously reported (e.g. Takahashi and Saitou 2012, Matsunami and Saitou 2013).

Schmidt et al. (2010) reported 11,588 CEBPA bindings in human cells and 19,212 transcription factor bindings for the same protein in mouse cells using ChIP-seq analysis in livers. This suggests that, for some proteins, human may not necessarily have higher transcription factor binding sites. This means that fewer CNSs do not always imply fewer regulatory elements. It may just show fewer shared regulatory elements as a result of high turnover rates. This makes more sense with my discovery that more ancestral CNSs are under stronger constraint because of their indispensability and that newly gained sites have higher turnover rates. Because functional elements have higher turnover rates (Meader et al. 2010), I interpret fewer CNSs in rodents, not as an effect of just fewer regulatory elements but as a result of high turnover rates. This implies that many regulatory elements are not conserved in rodents. This result suggests a high morphological diversity in the rodent lineage.

The genomic location may give a hint about the function of a CNS. For example, CNSs located several hundred kbp to a gene is less likely to regulate that gene as a proximal promoter element. My study indicates that the genomic distribution of CNSs is different in each lineage. Phylogenetically close lineages, such as primates and rodents, have more similar distribution. Also, about 30% of single-copy orthologous CNSs are located on different genomic positions in one or more different species. Although lower quality of cow genome annotation may contribute to this difference, even between human and mouse that have been well studied, 8% orthologous CNSs are located on different locations. This suggests that in the course of evolution, the genomic location of CNSs with respect to genes might change. One mechanism through which such "relocation" can happen is through gene loss. The gene harboring a CNS might be lost and the CNS kept if the CNS regulates a neighboring active gene. In this way, CNS previously in an intron is "relocated" to an intergenic region. Other gene restructuring processes such as translocation, inversion, gene fusion, and duplication followed by loss might also lead to this relocation. It is possible that such relocation may be correlated with local evolutionary rates as Liu et al. (2006) reported difference in evolutionary rates among regions. This assertion has to be further investigated.

My gene ontology analysis indicates that the putative target genes of the ancestral CNSs are enriched in genes involved with transcription and development. As the phenotypic differences observed among species arise during development, the CNSs may contribute to phenotypic differences. Hiller et al. (2012), for example, showed that the CNEs function as a spinal cord enhancer. Although CNSs may have other functions apart from regulation, they are more likely to function as regulatory elements (Hemberg et al. 2012; Shen et al. 2012). Considering the high dynamics in abundance, retention, and loss of ancestral CNSs and gain of new ones as well as the difference in genomic distribution, many CNSs are expected to hold the key to the understanding of the phenotypic diversity observed among species, via their activities as regulatory elements or other mechanisms that are yet to be fully understood.

CHAPTER FIVE

LARGE JAPANESE WOOD MOUSE TRANSCRIPTOME ANALYSES

5.0 Chapter summary

Large Japanese wood mouse or akanezumi (アカネズミ in Japanese), *Apodemus speciosus*, is a murine species endemic to Japan. The species, though not found in many other countries is found in all the major islands in Japan. In addition, the species have been reported to have adapted to temperate climate. Although the phylogenetic relationships among *Apodemus* have been studied, relationships with other murines are still under debate. Using transcriptome data, I show that *Apodemus* is closer to mouse than rat. The expression data also supports this phylogenetic relationship, demonstrating the power of expression dynamics in establishing phylogenetic relationship of rapidly diverged species. Finally, I show that CNS-associated genes have lowly expressed and lower expression correlation in liver demonstrating that CNSs are not active in liver tissues.

5.1 Introduction

Large Japanese wood mouse (*Apodemus speciosus*) is a murine species belonging to Muridae, the family to which rat and mouse belong. The family Muridae has the highest number of species, not only among the rodents, but also among mammals with 281 genera and 1326 species (Musser and Carleton 1993). Large Japanese wood mouse, hereafter referred to as akanezumi ("アカネズミ", the Japanese name of the species), is endemic to Japan (Kaneko and Ishii 2008). Although there is limited report of the detailed and extensive genomic evolutionary and phylogenetic analyses to understand the emergence or introduction of the species, the widespread distribution of the species across most islands in Japan is indisputable. Reported analyses using mitochondrial DNA have revealed some interesting population structures (e.g. Serizawa et al. 2000; Hirota et al. 2004; Suzuki et al 2014; Okano et al. 2015). The species belongs to the genus *Apodemus*. There are about 20 members of the genus across the world with four members found in Japan (Musser and Carleton 1993).

Unlike evolutionary and phylogenetic studies which are in paucity, ecological studies have been extensive (e.g. Sato et al. 2014; Sakamoto et al. 2012). One of the factors that have contributed to the ecological success of akanezumi is probably their ability to adapt very quickly to new environment. Akanezumi is one of the few rodent species that have evolved the molecular mechanisms to colonize temperate environment (Suzuki et al. 2004; Hirota et al. 2004).

The phylogenetic relationship among rat, mouse and akanezumi is still under debate.

Analyses of some genes suggest that akanezumi is more closely related to mouse than it is to rat (Liu et al. 2004; Steppan et al. 2005, Suzuki et al. 2008). Other analyses show the contrary, with akanezumi being closer to rat than mouse (Steppan et al. 2005). However, the lack of akanezumi genome and transcriptome data places an embargo on the decisive conclusion about the phylogenetic relationship among the species. Whereas the genome and numerous transcriptome data for a number of tissues of mouse and rat are available, there has not been any reported genome or transcriptome data for akanezumi. In collaboration with some Japanese researchers, we are now determining the whole genome sequences of akanezumi.

In this chapter, I present transcriptome analyses of akanezumi caught in the wild. Being a member of the genus *Apodemus*, the third most species rich genus among Murinae, the evolutionary analyses of akanezumi will illuminate some aspects of rodent evolution as it relates to the number of species. The decision to use the transcriptome data is also to investigate the evolutionary dynamics of gene expression across murines. Using the previously published mouse (Merkin et al. 2012), rat (Merkin et al. 2012) and naked mole rat (Kim et al. 2011) data, I report murine evolution from gene expression data. I show from both nucleotide and expression data that akanezumi is closer to mouse than rat. Also, I further demonstrate that CNSs are not active in liver.

5.2 Materials and methods

5.2.1 Sample collection

The samples for this study were kindly provided by Dr Tomozawa Morihiko of Keio University. Akanezumi samples were collected on May 30th, 2013 from Miyake Island in Japan. One individual each of ages 3-, 5- and 7-days postnatal was sacrificed on the field, and their whole bodies were immediately preserved in RNAlater® solution. In addition, the liver, kidney, heart, spleen and muscle samples of an adult sample was harvested and immediately preserved in RNA later solution. The harvested samples were then transported to the laboratory, and were preserved at -80°C until RNA extraction.

5.2.2 RNA extraction and purification

Total RNA was extracted from the whole body and various tissue samples. RNA extraction was done using TRIzol® reagent. The sample was first homogenized in 9 times TRIzol® reagent. The homogenized samples were then phased according to the manufacturer’s protocol. The aqueous layer was extracted from the phased homogenate. Total RNA was precipitated and washed from the extracted aqueous layer. The extracted total RNA was then purified with Agencourt AMPure XP®. The qualities of the extracted RNA were evaluated using bioanalyzer and gel electrophoresis. Unfortunately, the qualities of RNA extracted from kidney, heart, spleen and muscle were not good enough for sequencing. The purified RNA was immediately preserved in -80°C. The purified RNA samples were then sent to BGI for transcriptome RNA-seq using Illumina HiSeq 2000.

5.2.3 Transcriptome assembly

Using Hiseq2000, 90bp paired-end reads of the trasncriptomes were generated. The quality of the fastq files was examined using Fast QC. The first 11bp of the reads were trimmed using Fastx toolkit to exclude any possibility of adapter contamination. The decision to remove the first 11bp was reached after closer examination of the nucleotide composition distribution of the reads. Because the genome sequences of wood mouse have not been published, *de novo* assembly of the transcripts was done using release 2013-02-25 of Trinity (Grabherr et al. 2011). To make a comprehensive gene list, the reads of 3-, 5- and 7-day postnatal transcriptomes, in addition to liver transcriptome were first combined. After the combination, the reads were then assembled. The command used was follow,

```
Trinity.pl --seqType fq --SS_lib_type RF --left read_1.fq --right read_2.fq --CPU 8 --JM 40G
--min_kmer_cov 3 -bflyHeapSpaceMax 40G
```

The unavailability of akanezumi genome data makes mapping of the reads to a reference genome unfeasible. That is the reason why *de novo* assembly of transcripts was done. For comparative analyses, published transcriptome data of liver samples of mouse, rat and naked mole rat were retrieved. Although the genomes of the species have been published, I also made *de novo* assemblies of the transcripts of the species to ensure consistency. The first 11bp of the reads were trimmed before *de novo* assembly. The command used for the Trinity assembly of liver samples was;

```
Trinity.pl --seqType fq --SS_lib_type RF --left read_1.fq --right read_2.fq --CPU 8 --JM 40G
-bflyHeapSpaceMax 40G
```

The default value of —min_kmer_cov option was used for the *de novo* transcriptome assembly of liver sample of all the species. The decision was based on the limited number of reads. Whereas four samples were combined in the first assembly, only liver sample was used in this assembly. Because whole bodies were used, there is a possibility of microbial RNA contamination in whole body samples. This suggests that many lowly expressed genes might not be akanezumi genes. The possibility of microbial contamination is relatively low in liver sample. Therefore, there was no need to filter off lowly expressed assemblies.

5.2.4 Retrieval of data used

Complete sets of protein-coding gene sequences of mouse and rat were downloaded from *Ensembl* database version 82. The genome sequence data of naked mole rat was downloaded from UCSC database. The list of the species used and the sources of protein-coding sequence data used for sequence evolutionary analyses are shown in Table A5.1 of Appendix 5. For the comparative analyses of gene expression patterns, raw reads of liver samples of mouse, rat and naked mole rat were downloaded from Gene Expression Omnibus (GEO) database. The accession numbers of the reads are presented in Table A5.2 of Appendix 5.

5.2.5 Homologous gene search

Homologous genes were extracted by reciprocal blast searches. Because of the availability of good quality data, mouse was used as the reference species. For all multiple-transcript genes, the longest transcripts were used for evolutionary analyses. To ensure consistency in liver expression data analyses, *de novo* assembled data were used. For each species, the first run of blast search was run with mouse amino acid sequence as query and the Trinity-assembled transcripts as the subject. TBLASTN searches were run with E-value of 0.00001. For the reciprocal blast search, BLASTX was used with the Trinity-assembled transcripts as the query and mouse as the subject and the E-value of 0.00001. After the reciprocal blast searches, reciprocal best hits were extracted using bit scores. In this context, two genes are considered homologous if the homology search between them gave the highest score irrespective of which was used as the query.

After extracting the homologous gene pairs between mouse and all the Trinity-assembled transcripts, gene clustering was done using the mouse gene as reference. For clustering, I extracted the corresponding transcripts of all mouse genes. For every gene, I extracted transcripts from all Trinity-assembled transcripts. Only transcripts with identifiable homologs in all samples were used for subsequent analyses.

5.2.6 Sequence evolutionary analyses

The protein-coding sequences of rat and mouse were retrieved from Ensembl database build 82. For naked mole rat, the available gene sequences were predicted with limited homology with mouse. To obtain more genes that could be used for evolutionary analyses, I used GMAP (Wu and Watanabe 2005) to extract protein-coding genes using guinea pig as the query. Guinea pig was used because it is the closest species to naked mole rat with gene annotation in Ensembl database. The details of GMAP analyses are shown in Chapter Six. Similarly, the protein-coding sequences of akanezumi were extracted using GMAP. The assembled transcripts were used for coding-gene prediction using GMAP. GMAP performed better than BLASTX or TBLASTN.

As described above, the homologous gene pairs were extracted using mouse as the reference. In this case, reciprocal BLASTP searches were used. After establishing the homology information, the genes were clustered. Note that each cluster contains one gene per species. Independent amino acid multiple alignments (with CLUSTALW) were run for each homologous cluster with genes from all species used. The amino acid alignments were transformed to codons for further evolutionary analyses.

5.2.7 Analyses of expression dynamics

The first step in evolutionary analyses was to estimate transcript abundance in order to know the level of expression of certain transcripts. To avoid bias due to quality of annotation, I did *de* *novo* assembly of liver transcripts in each species (as described above). The abundance was then estimated using RSEM (Li and Dewey 2011) in Trinity package. Briefly, RSEM estimates the abundance by using Bowtie (Langmead et al. 2009) to align raw reads to the transcripts. Under certain assumptions, the expression levels are proportional to the amount of reads mapped, expressed in Fragments Per Kilobase per Million mapped reads (FPKM).

For expression dynamics, the abundance estimates of homologous transcripts were compared across species. The expression correlations of the transcripts were used for phylogenetic and principal component analyses. Briefly, for each pair of species, Spearman’s correlation coefficients were computed from the transcript abundance estimates. Only transcripts with non-zero expression in all species were used. The expression distance is simply given as;

$$\text{Expression distance} = 1 - \text{Spearman's correlation coefficient} \qquad (5)$$

Expression distance matrix was then computed for the species used. Note that two individuals of naked mole rats (4 years old and 20 years old) were independently analyzed. Neighbor joining (Saitou and Nei 1989) tree was constructed from the expression distance matrix using MEGA6 (Tamura et al. 2012). For principal component analyses, the expression distance matrix was used for the analyses in R statistical package.

5.3 Results

5.3.1 Akanezumi transcriptome data

Transcriptome sequencings were performed for RNA extracted from whole bodies of 3-, 5- and 7 days postnatal and adult liver of wild individuals. The sequencing statistics are presented in Table 5.1. For all the four samples sequenced, 90bp paired-end reads were produced with the insert size of 200bp. A total read number of between 50 million and 71 million were produced per sample. These numbers correspond to 4.5-6.4Gbases. In all the samples, more than 98% of the bases have Phred quality score of 20 or above (i.e. ≤ 1% sequencing error rate). The evaluation of the nucleotide compositions (see Figure A5.1 of Appendix A5) shows fluctuation of nucleotides at the first few bases of the reads. This distribution might be because of the sequencing bias or adapter presence. To avoid any complication that might arise because of the nature of reads, the first 11bp of every read was trimmed off. As a result, between 3.9-5.6Gbases were used for transcriptome assembly.

5.3.2 *De novo* transcriptome assembly of akanezumi

To obtain a comprehensive set of akanezumi transcripts, the reads of the four samples were combined for *de novo* transcriptome assembly. Although the use of whole bodies and combination of tissues would ensure the extraction of more number of transcripts, complications can also arise due to microbial RNA contaminations from whole bodies and the presence of tissue-specific alternatively spliced transcripts across tissues (Merkin et al. 2012).

Table 5.1: Sequencing statistics of akanezumi transcriptomes

Sample	Insert size	Read length (bp)	Read number (mi)	Total base (Gbp)
Liver	200	90	61	5.5
3-day postnatal	200	90	50	4.5
5-day postnatal	200	90	55	4.9
7-day postnatal	200	90	71	6.3

These make *de novo* assembly of transcripts more complicated, especially with the inclusion of whole body samples with many heterogeneous tissues included. With the assumption that microbial contamination would be at a low expression level, and that many genes have major splice variants, I set the k-mer coverage depth of Trinity (Grabherr et al. 2011) to 3. This threshold also ensures the sequences of relatively better quality are available for evolutionary analyses. The results of Trinity assembly are presented in Table 5.2.

For transcript annotation, I attempted to use blast with mouse amino acid sequences as the query. However, GMAP annotation seemed to perform better. So I decided to focus my analyses based on homology gene extraction using GMAP. For GMAP extraction, I used the longest mouse coding sequences (CDS). This implication of the search is that only one transcript per gene would be reported. This gave 13,764 homologs of mouse genes predicted from akanezumi transcripts.

Table 5.2: Summary of akanezumi Trinity assembly

	Transcripts	**Components (or genes)***
Total number	227,530	128,853
GC content	52.62	52.62
Contig N10 (bp)	6,641	4,349
Contig N20 (bp)	5,095	2,988
Contig N30 (bp)	4,147	2,157
Contig N40 (bp)	3,427	1,490
Contig N50 (bp)	2,806	940
Average length (bp)	1,281	602

*Component is the terminology in Trinity which loosely represents genes. A component may have single or multiple transcripts.

5.3.3 Establishing akanezumi phylogenetic position

With more abundant akanezumi transcriptome data, the first thing that I established is the phylogenetic relationships between akanezumi and other species examined. While it is well established that akanezumi is a murine, and thus more closely related to mouse and rat compared to naked mole rat, the distance between akanezumi-mouse and akanezumi-rat distance had not been well established. The specific question is whether akanezumi is phylogenetically closer to mouse or rat. To answer this, I concatenated all the aligned amino acid sequences to give an overall evolutionary pattern. About 3.8 million gapless amino acid sites were used to establish the relationship. As shown in Figure 5.1A, akanezumi clusters with mouse with 100% bootstrap value.

Phylogenetic relationships with amino acid sequences are sometimes not reliable because of evolutionary forces. Because of the heterogeneity in the intensity of selection across lineages (e.g. Palmer et al. 2005; Burri et al. 2015; Zhang et al. 2015), phylogenetic relationship might be affected. Specifically, branch lengths might be misleading. Therefore, I did the phylogenetic analyses using more sophisticated General Time Reversible (GTR) model on mostly neutrally evolving third codon positions. The same result is found, though with shorter relative branch between rat and mouse-akanezumi common ancestor (Figure 5.1B). Using the second codon positions, the divergence time between mouse and *Apodemus* was estimated to be about 70% of the divergence time between mouse and rat. Accurate estimation of divergence time requires high quality genome data and transcriptome data may not be appropriate.

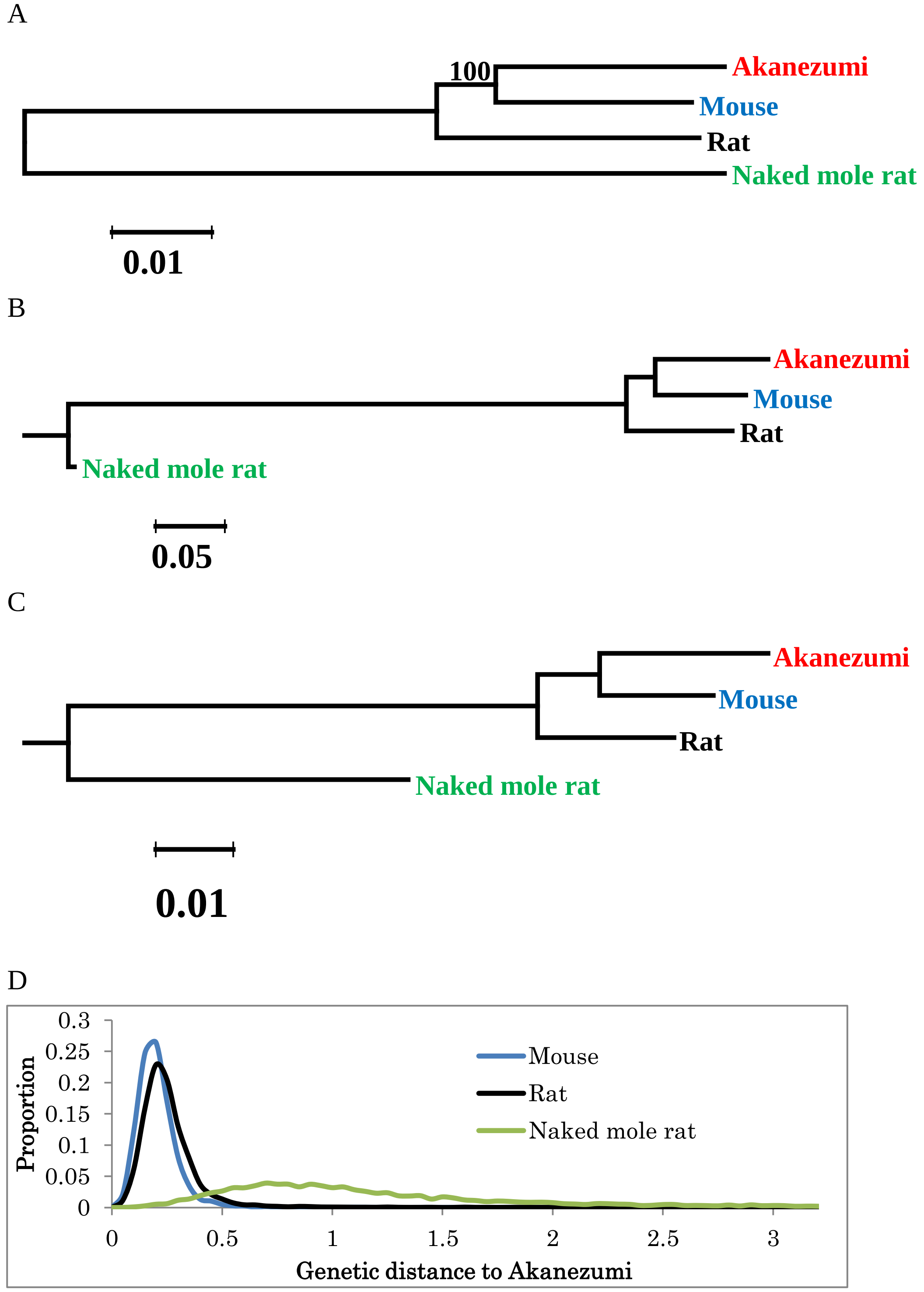


Figure 5.1: Establishing the phylogenetic position of akanezumi (A) NJ amino acid phylogenetic tree (Poisson model; Gamma = 5). Phylogenetic trees were constructed with third (B) and second (C) codon positions using GTR model of PAML. Note that for B and C, tree topology was supplied. Therefore, bootstrap values are not shown. (D) Genetic distances between akanezumi and other species using third codon sites.

The second codon positions are the most selected positions as they have the highest number of nonsynonymous substitutions. Even for second codon positions, with GTR model of substitution, mouse still clusters with akanezumi. The observation is not different with first codon positions (Figure A5.2 in Appendix 5). The observation of the same results using different types of data and different models demonstrates that the result is reliable.

A closer look at Figure 5.1B suggests that the splitting of rat, mouse and akanezumi occurred in rapid succession. The branch length suggests that mouse and akanezumi split soon after the split of their common ancestor and rat. Figure 5.1D gave a better description about the events. By looking at the distribution of the genetic distance of third codon positions of the genes, akanezumi-mouse distance, though significantly higher, is similar to akanezumi-rat distance. Taken together, the results suggest that rat, mouse and akanezumi splitting occurred in a short evolutionary time.

5.3.4 Analyses of expression data support the phylogenetic relationship

More closely related species should have more similar expression patterns because of more recent shared ancestry. Therefore, by examining the expression patterns among species, the phylogenetic relationship can be established. This observation of true representation of phylogenetic relationship specifically for the same tissue types has been shown by previous reports (e.g. Brawand et al. 2011; Harrison et al. 2015). However, considering the rapid splitting of the three species, would their expression patterns reveal the true phylogenetic relationship?

To investigate this, I first examined the correlations of expression between pairs of species using the expression patterns of all transcripts which are expressed in all the tissues. As expected, the Spearman’s correlation coefficient for akanezumi-mouse (0.80) is higher than 0.78 of mouse-rat and 0.76 of akanezumi-rat (Table A5.3). To compute the tree, the expression distance, estimated from the correlation, is used (see Material and methods for details). Figure 5.2A confirms that mouse is phylogenetically closer to akanezumi than rat. Surprisingly, the two individuals of naked mole rats have long branches. This may be because of the age difference. One individual was 4 years old while the other was 20 years old. This highlights the impact of age difference in gene expression dynamics. For example in Figure A3.4A in Appendix 3, gene enrichment of CNSs was different between embryonic and adult liver.

Another way to consider the expression dynamics is to represent the expression information on a principal component (PC) plot. PC analysis uses the same expression distance matrix. Figure 5.2B shows the plot for the species. The first PC separates the murines from naked mole rats. Notably, 74% of all variance is explained by PC1 (Figure A5.3 in Appendix 5). PC2 which explains about 11% of the total variation clearly separates rat from mouse and akanezumi. Although majority of the variance can be explained by the first and second PC, as would be predicted, third PC separates mouse and akanezumi (Figure A5.3B in Appendix 5). Unlike the observation in Figure 5.2A however, the two individuals of mole rat were located close to each other.

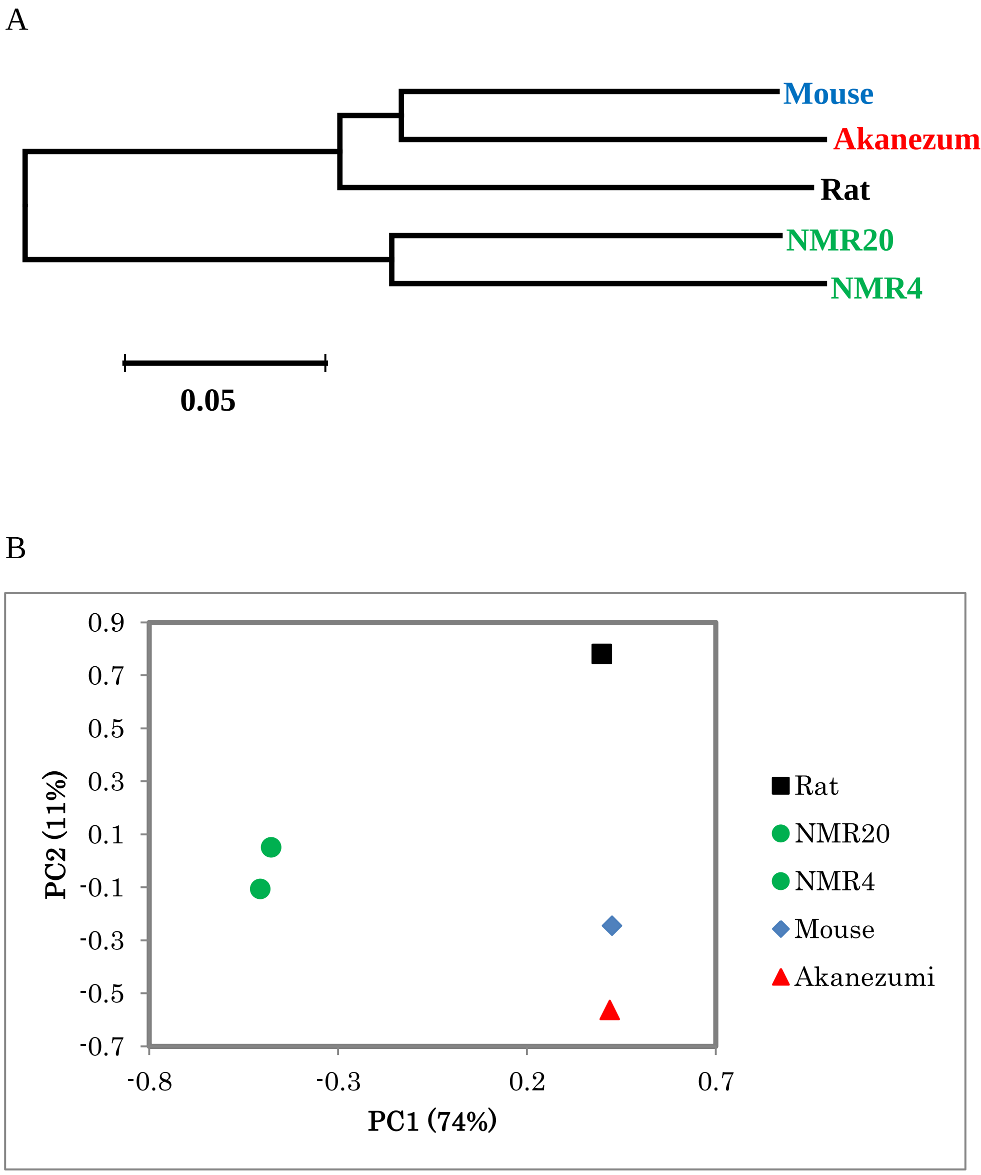


Figure 5.2: Phylogenetic relationship established from liver expression dynamics. (A) NJ phylogenetic tree was computed by using the expression distance. (B) PC2 separates rat from akanezumi and mouse. NMR20 and NMR4 are 20 and 4 years old naked mole rats, respectively.

5.3.5 Conserved noncoding sequences (CNSs) are not active in liver

It has been shown that gene expression dynamics can correctly reproduce phylogenetic relationship even among species that diverged rapidly. The next question was whether CNSs play important role in liver expression dynamics. It is important to note that genes expressed in liver tend to be underrepresented in CNSs (see Figure A3.4 of Appendix 3 in Chapter Three). First, I checked whether CNS-enriched GO terms (development, transcription and nervous system) have higher expression correlation, that is more conserved expression (see Chapter Three for more information), than CNS-depleted GO terms (Chapters Three and Four). The mouse coordinates of CNSs reported in *Chapter Three* (conserved among mouse, human, dog, cow and chicken) were used. However, Spearman’s correlation coefficients for each pair of species are similar (Figure A5.5 in Appendix 5). This is contrary to the expectation of higher expression conservation of CNS-associated genes in Chapter Three.

A further consideration of CNS association with liver expression confirmed the results in Chapter Three that CNS-associated genes are less expressed in liver. Focusing on akanezumi for example (Figure 5.3A), the median expression values of genes not associated with CNSs (5.04FPKM) is significantly higher than 4.22, 3.68 and 3.25 for genes with 1-3, 4-9 and >9 CNSs, respectively (p value < 0.001). Similar results are found in other examined species (Figures A5.5 in Appendix 5).

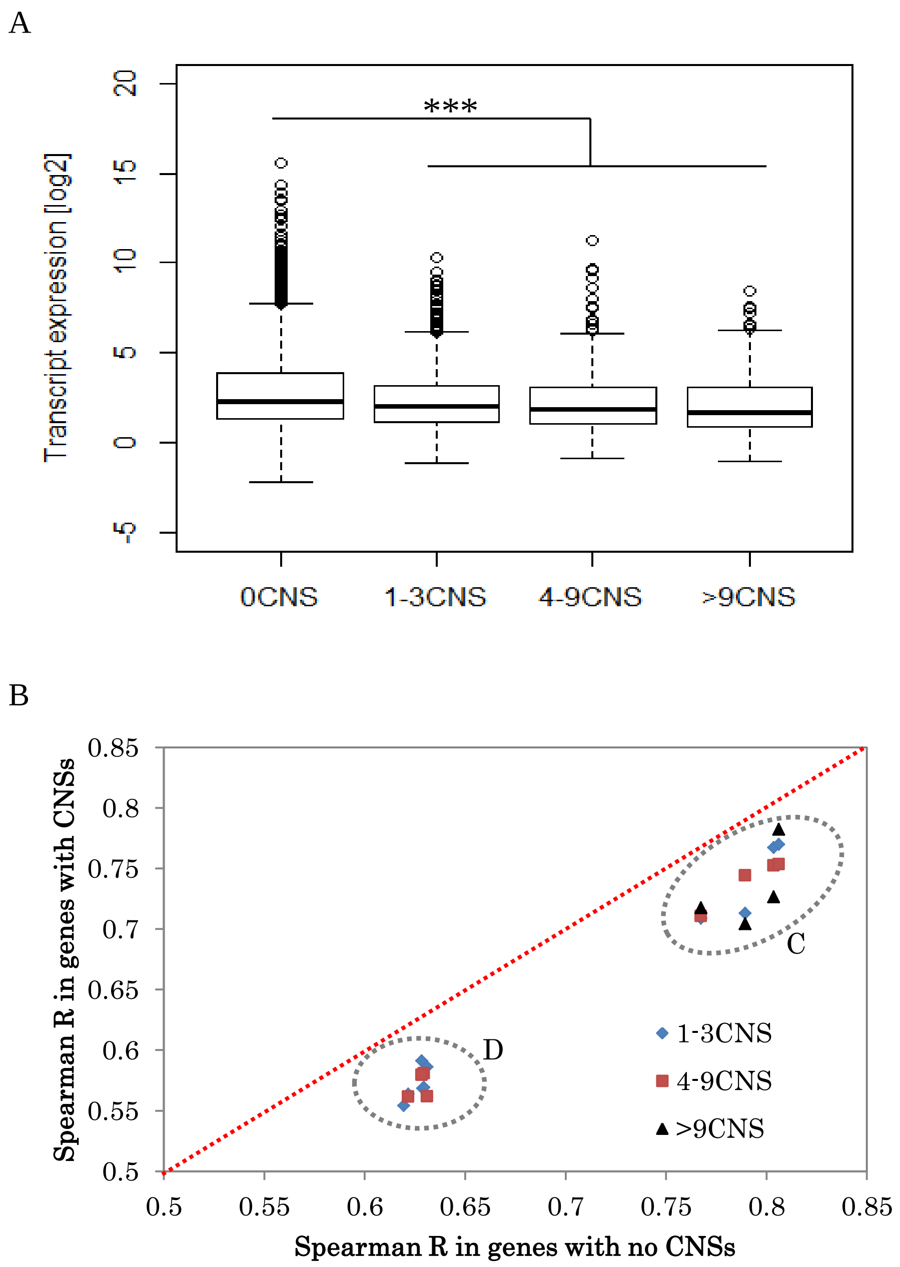


Figure 5.3: Conserved noncoding sequences are not active in liver. (A) The expression data of akanezumi liver shows that CNS-associated genes have lower expression (p value < 0.001; Mann Whitney U). (B) Genes with no CNSs tend to have higher expression correlation. Each point represents each pair of species. Spearman expression correlations for each class of genes (based on CNS association) were computed. While closely related species (e.g. two murines) have more similar expression (cluster C), more distantly related species (e.g. murine and naked mole rat) have less similar expression (cluster D).

Although the actual expression level for CNS-enriched genes in liver are lower, is the expression level more conserved? To answer this, I computed the Spearman's correlation for all the pairs of species using different classes of genes. Gene classification was based on the number of CNSs associated with the genes. Figure 5.3B shows that the correlation of expression is actually higher in genes with no CNSs than in those with various numbers of CNSs. This is contrary to the report in Chapter Three that CNS-associated genes tend to have higher expression. Therefore, this absence of basic property of CNS-associated gene suggests that CNSs are not active in liver.

5.4 Discussion

The transcriptome of akanezumi 3-, 5- and 7-day postnatal and adult liver samples were analyzed using RNA-Seq technology on Illumina Hiseq2000 platform. For each sample, more than 45 million reads were produced. In collaboration with other Japanese researchers, we are now assembling the complete genome of akanezumi. However, up to the time of writing this thesis, the genome assembly had not been finalized. That is the reason why *de novo* transcriptome assembly was used. For comprehensive evolutionary analyses, all the reads from the four samples were used for assembly. For proper evaluation of gene expression dynamics however, only liver samples were used for *de novo* assembly. Although the genomes of mouse, rat and naked mole rats have been published (Mouse Genome Sequencing Consortium 2002;

Rat Genome Sequencing Project Consortium 2004; Kim et al. 2011), to ensure consistency, *de novo* assembled transcripts of all the species were used.

The first question I resolved is the phylogenetic position of akanezumi, with respect to mouse and rat. The analyses of various sequence data show that akanezumi is more closely related to mouse than to rat. This is consistent with the reports of Liu et al. (2004), Steppan et al. (2005) and Suzuki et al. (2008). This phylogenetic relationship is supported with 100% bootstrap probability. However, the splitting of the three species happened in rapid succession such that akanezumi-mouse genetic distance is only slightly lower than akanezumi-rat genetic distance. Despite the rapid splitting of the three species, liver evolutionary dynamics correctly reveal the phylogenetic relationship. The first principal component (explaining about 74%) separates naked mole rat from murines. The second clearly separates rat from mouse and akanezumi. The third component then separates mouse and akanezumi. This observation shows that gene expression dynamics are changing so fast, and that phylogenetic relationship can be established by expression dynamics.

Having established the speed of gene expression dynamics, I decided to further investigate the role of CNS in liver expression. From the results in Chapter Three, CNS-associated genes tend to have more conserved expression. However, CNS activity was shown to be highly tissue-specific. I then decided to analyze the effect of gene-CNS association in liver expression dynamics. Because of the unavailability of akanezumi genome sequences, I did not extract CNSs from the same species. However, the coordinates obtained in Chapter

Three were used. These coordinates are conserved among mouse, human, cow, dog and chicken. Parsimoniously, there is a high probability that they are also found in akanezumi, rat and naked mole rat. The GO analyses show that the terms enriched with CNSs tend to have similar correlation expression than those depleted in CNSs. As reported in Chapter Three however, the expression level of CNS-associated genes is lower than for CNS-depleted genes.

I found that the expression correlations of genes with no CNSs are higher than for genes with CNSs. This is contrary to the results in Chapter Three that CNSs are associated with more conserved expression. There is a simple explanation for this observation. CNSs are not active in liver. Because CNSs are not active in liver, there is no need for expression conservation in liver. In fact, gene expression patterns clearly show that CNS-associated genes are less expressed in liver. It is important to note that CNSs are overrepresented in GO terms associated with development, transcription and nervous system (e.g. Takahashi and Saitou 2012, Babarinde and Saitou 2013). On the contrary, they are underrepresented in genes associated with immunity, defense and response to stimulus (Babarinde and Saitou 2013). Therefore, it is logical to expect that they would not be active in liver.

In conclusion, I have used the transcriptome data to establish the phylogenetic relationship among akanezumi, mouse and rat. The phylogenetic relationship was established from both sequence and expression data. The congruence of the expression and sequence evolutionary phylogenetic data, despite rapid splitting, demonstrates that the expression patterns of species evolve so fast. Using liver expression data, I have also established further proofs for

CNS activities and some hitherto vague CNS properties.

CHAPTER SIX

RODENT EVOLUTIONARY RATES: INSIGHTS FROM CAPYBARA GENOME SEQUENCING

6.0 Chapter summary

With the advancement of sequencing technology, it is now time to re-evaluate some of the long-standing hypotheses. This chapter investigates rodent evolutionary rate dynamics. For this analysis, the genome sequences of capybara (*Hydrochoerus hydrochaeris*) were determined to a depth of 13× on Illumina Miseq platform. The newly determined capybara genome was analyzed together with other previously published genomes. First, I show that though rodents have higher evolutionary rates than primates, there is heterogeneity among rodents, with mouse and rat being higher than capybara, guinea pig and naked mole rat. Three pairs of species that are phylogenetically closely related, but significantly different in body sizes were investigated for body size effect of evolutionary rate. The analyses of these species show limited support for the association of body size and evolutionary rate.

6.1 Introduction

Molecular clock (Zuckerkandl and Pauling 1965) assumes the constancy of evolutionary rates across lineages. This constancy has been widely used in determining the divergence time from nucleotide or amino acid sequences (e.g. Vawter 1980; Hasegawa 1985; Bromham and Penny 2003; Peterson 2004). Supporting the constancy of evolutionary rate, Kumar and Subramanian (2002) reported that the mutation rates are largely similar, not only across genes, but also across lineages. They therefore reported that molecular constancy can be assumed in calculating the divergence time. As the number of genome projects increases, the heterogeneity of nucleotide divergence among genes are now widely accepted (Zhu et al 2015; Dos Reis et al. 2015). Notably, genes associated with immunity and defense tend to have higher nucleotide divergence. On the other hand, genes associated with morphogenesis and development tend to have lower divergence. Although differences in intensity and direction of selection pressure contributes abundantly to the differences in nucleotide divergence, analyses using synonymous substitution patterns have suggested that difference in the substitution rates contributes to the nucleotide diversity observed among genes.

The heterogeneity is not only limited to genes. Several studies have reported that heterogeneity of evolutionary rates across lineages. An interesting example of different evolutionary rates in mammals is the difference between rodent and primate evolutionary rates (Li and Wu 1985; Wu and Li 1987; Li et al. 1996). Using the available mouse, rat, human and dog sequences, Wu and Li (1985) reported that rodent evolutionary rate is much higher than in primates. The observed differences were then attributed to the difference in generation interval which causes difference in the number of replications per year. The results imply that species with short generation time would have higher evolutionary rates. According to Li et al. (1996), if generation interval is used as the unit of time, the differences in evolutionary rates would not be found. Martin and Palumbi (1993) proposed the reconsideration of generation time hypothesis to include physiological processes. It is now widely accepted that molecular clock does not always hold. In fact, many studies have supported differences in mouse and human evolutionary rate, not only in protein-coding sequences (Zhu et al 2015; Dos Reis et al. 2015), but also in conserved noncoding sequence evolution (Hiller 2012, Takahashi and Saitou 2012, Babarinde and Saitou 2013).

A number of hypotheses have been proposed to explain evolutionary rate differences across lineages (Bromham et al. 1996; Bromham 2011). A notable hypothesis is the generation time hypothesis (Ohta 1993; Bromham et al. 1996). The hypothesis predicts that species with long generation interval would have lower replications per unit time. Since mutations arise from replication errors, more replications would imply higher substitution rate (see review by Bromham 2011). The generation time hypothesis has been found to apply to a broad range of species (Thomas et al. 2010). Using complete genomes of several mammalian species however, Huttley et al. (2007) reported that their observation is inconsistent with the generation time hypothesis. On the contrary, they reported that replication enzymology and shifts in nucleotide pools are the factors that contribute to the differences in mutation rates. They claimed that fidelity of replication is different across species and this contributes to evolutionary rate difference. Huttley et al. (2007) like Kumar and Subramanian (2002), challenged the generation time hypothesis. Another attribute that has been associated with substitution rate difference among species is the body size. As body size has been suggested to be correlated to the metabolic rates and generation interval, it is logical to conceive the relationship between body size and evolutionary rate. Indeed, a number of reports have shown results suggesting the relationship (e.g. Martin and Palumbi 1993; Speakman 2005).

The earlier studies (Wu and Li 1985; Li and Wu 1987; Li et al. 1996) that reported evolutionary rate difference between primates and rodents were carried out in pre-genomic era. At that time, it was not easy to analyze many species because of limited data. At the time of the study, mouse and rat were used as the representative of rodent species, while human was used as the representative primate species. Although the difference in evolutionary rates of human, mouse and rat on the other hand is widely supported, it is still unclear whether the evolutionary rates of these species reflect the overall respective order. Although mouse and rats were found to have higher evolutionary rates (Wu and Li 1985; Li and Wu 1987; Li et al. 1996; Hiller 2012, Takahashi and Saitou 2012, Babarinde and Saitou 2013), it is still difficult to generalize that rodents have higher evolutionary rates. This is because mouse and rats are phylogenetically very close, belonging to the same sub-family, Murinae (Musser and Carleton 1993). The high evolutionary rate observed in mouse and rat was reported as a general feature of rodent order. However, it is also possible that the high evolutionary rate is a unique attributes of Murinae.

Similarly, the reported primate lower evolutionary rate may be unique to human lineage and may not represent the general feature of primate lineage. Therefore, it is important to further investigate more phylogenetically and phenotypically strategic species before concrete conclusions could be drawn.

With the advancement of sequencing technology, it is now possible to investigate a larger number of species. Higher number of species and higher evolutionary rates, at least in some species, make rodents an interesting mammalian order for the study of evolutionary rate. A number of studies have investigated the phylogenetic relationships of rodent (e.g. Blanga-Kanfi et al. 2009; Huchon et al. 2002). It is important to note that the nucleotide-based and amino-based phylogenetic relationships accurately depict morphology-based taxonomic classification. The order Rodentia is divided into three major lineages namely; mouse-related, squirrel-related and Ctenohystrica (Blanga-Kanfi et al. 2009). Mouse-related lineage is probably the most genetically studied clade. This clade includes Muridae which represent about two thirds of the total known rodent species (Musser and Carleton 1993). Mouse and rat are members of the clade. Squirrel-related lineage contains all types of squirrels with unremittingly growing pair of incisors. The third clade, Ctenohytrica is very interesting for the investigation of evolutionary rates because of the heterogeneity of the member species. Notably, the order contains capybara (*Hydrochoerus hydrochaeris*), the largest extant rodent species and naked mole rat (*Heterocephalus glaber*), the rodent species with the longest lifespan. While capybara can weigh up to 81kg (Ferraz et al. 2005), naked mole rat can live up to 30 years. Interestingly, naked mole rat is relatively small-bodied while capybara, though big-bodied does not have an exceptionally long life span.

The genome of naked mole rat has been published (Kim et al. 2011). The ability of naked mole rat to live for a very long time, while still maintaining the full productive potential without showing the signs associated with old age is being genetically studied. This feature makes naked mole rat a good model species for the study of cancer genetics. However, unlike naked mole rat, the genome sequences of capybara have not been reported. Capybara is originally endogenous to South America, and the wild population is still largely restricted to the region. To investigate the evolutionary rate of rodent, the genome sequences of capybara were determined using next generation sequences. In this study, I present the first report of capybara genome sequences and the aspects of rodent evolution that has hitherto been unreported.

6.2 Materials and methods

6.2.1 Sample collection and DNA extraction

Masseter muscle of an adult male capybara which died of an undisclosed cause was donated by Izu Shaboten Park, Japan. For DNA extraction, the tissue sample was digested with Qiagen ATL reagent and protease, and then treated with RNase. DNA was then extracted and purified using phenol isomyl chloroform extraction method. The quantity and the quality of the extracted DNA were checked using NanoDrop Spectrophotometer, Qubit® 3.0 Fluorometer and Agilent 2100 Bioanalyzer.

6.2.2 Library preparation and sequencing

The library was prepared with Agilent SureSelect QXT library preparation kit. The starting material was 30ng genomic DNA sample reconstituted in 1ul of ultrapure water. The genomic DNA sample was then fragmented and adaptor-tagged. The adaptor-tagged library was purified before PCR amplification. The amplified library was then purified with AMPure XP beads after which library DNA quantity and quality were assessed using Qubit® 3.0 Fluorometer and Agilent 2100 Bioanalyzer and DNA 1000 Assay. The average fragment size was 750bp. Two library sets were prepared to minimize random effects due to fragmentation. After the library integrity had been ascertained, the two library sets were pooled by combining 2.5ul of 4nM of each library set. Three runs of sequencing were done for the pooled library set on Miseq sequencing platform using 600 cycle kits. The set sequencing length for read1 was 350bp while the length of 250bp was set for read2.

6.2.3 Sequencing output and quality assessment

After combining the output of the three sequencing runs, a total of ~158 million reads were produced. The quality of the sequencing reads were assessed using FastQC. Because the aim of the study is to evaluate evolutionary rate, sequencing error would have a serious impact on the results. Therefore, I decided to filter and trim the reads. To maximize the yield while ensuring the strict quality threshold, a python script was written. With the python script, all base positions in any kept read would not be lower than the threshold Phred quality score and length.

The python script therefore filters and/or trims the reads as the situation demands.

6.2.4 *De novo* assembly of capybara draft genome

CLC workbench and SOAP-denovo were used for the assembly of the reads. In order to have abundant data for better contig formation, I used the unfiltered reads for initial *de novo* assembly. First, the contig formation was done using CLC workbench available in Cell Innovation project (https://cell-innovation.nig.ac.jp). Thereafter, scaffolds were made and gaps were closed using SOAP-denovo (Li et al. 2010). For contig formation, scaffold making and gap closing, default settings of the respective software were used.

6.2.5 Extraction of high-quality genomic regions

Because of the limited amount of sequencing output and the purpose of this study, it became imperative to minimize the effect of the sequencing error. At very high coverage, such concerns would be irrelevant. However, in this study, I attempted to minimize the error rate to the bare minimum. I first attempted to analyze the sequencing depth. I used BWA-MEM 0.7.5a (Li and Durbin 2009) to align the reads to the assembled draft genome using the default settings. The depth per base position was determined using Samtools 0.1.19 (Li et al. 2009). To extract high-quality regions, I mapped filtered reads (minimum Phred quality score or all bases =20; minimum length = 50bp). The depth per position when high-quality reads were used was computed. The average depth was about ~10×. While low-depth regions might have some erroneous positions, too high depth may also have low integrity because of repetitive sequences. Therefore, I decided to use regions between 3-30×depths (between a third of the average and three times the average). Minimum Phred quality of 20 and depth of 3× imply that the sequencing error rate would be at most $10^{-2\times3}$ (i.e. 10^{-6}) per site. This value corresponds to one erroneously determined site per megabase pair region. Nucleotide on each position was examined using Samtools (Li et al. 2009) mpileup with –uf flag. I thereafter converted the files to vcf files to assess the variant sites. If the high-quality mapped reads have a variant nucleotide, the variant nucleotide is called for the position. In the case of heterozygous sites, the allele with the highest frequency was picked. In addition, I understand that mapping may be another source of errors and required that the mapping quality must be at least 30 for each position. Positions that do not meet these criteria were masked to "N" and treated as undetermined in the downstream analyses. Furthermore, I applied a stricter threshold of Q30 and 3-27× depth to validate the reliability of the Q20 threshold.

6.2.6 Gene extraction

Because of the sequencing depth and coverage, I reasoned that *de novo* gene prediction might be difficult. I therefore relied on the homology-based gene extraction. The availability of guinea pig gene annotation made the gene extraction feasible. GMAP (Wu and Watanabe 2005) was used to detect capybara genes using guinea pig coding sequences as the query. Because capybara and guinea pigs are not the same species, I used 'cross-species' option. The GMAP commands are as follow;

```
gmap_build -D . -d Capybara_20Q_50bp_d3D30 -k 8 -q 1

gmap -n 1 -t 100 --cross-species -D . -d Capybara_20Q_50bp_d3D30 transcript.fa -Q > Capybara_20Q_50bp_d3D30_prot.fa

gmap -n 1 -t 100 --cross-species -D . -d Capybara_20Q_50bp_d3D30 transcript.fa –E genomic > Capybara_20Q_50bp_d3D30_exons.fa
```

The extracted exons were joined to form a single coding sequence per homologous guinea pig gene.

6.2.7 Extraction of homologous gene sets

The protein-coding genes of the selected species were retrieved from Ensembl database. I then extracted the longest transcript per gene. Reciprocal BLASTP (Altschul et al. 1997) searches were conducted using guinea pig and every other species. For the blast searches, I set the threshold E-value to 0.00001. For each species pair, I extracted the reciprocal best hit using the bit scores of the alignments. Any alignment with the bit score of less than 100 was discarded. Thus, I retrieved the homologous gene pairs between guinea pig and every other species. For each guinea pig gene with identifiable homologs in all other species, I extracted the amino acid sequences. Each gene cluster was written into different file.

6.2.8 Multiple sequence alignment

The extracted multiple sequences were aligned using CLUSTALW2 (Larkin et al 2007). To make the codon alignment, I retrieved the coding sequences for all the genes from Ensembl database. The amino acid sequence alignments were then converted to coding sequence alignments. The aligned coding sequences of all gene clusters were concatenated to produce single multiple alignment sequences. From the multiple alignment sequences, the first, second and third codon positions were extracted independently. Any position with gap in any of the species used was discarded. Although this step leads to the underestimation of the actual evolutionary rates, the patterns across species are not likely to be affected.

6.2.9 Phylogenetic and distance computation

The phylogenetic relationship was first established using the gapless amino acid alignment. The Neighbor-Joining (NJ) tree (Saitou and Nei, 1987) was computed using MEGA6 (Tamura et al. 2013). For the initial NJ tree, the phylogeny was tested with bootstrap method. The nucleotide substitution model used was maximum composite likelihood. Gamma distribution (k =5) was used for rates among sites. Also, the patterns among lineages were set to be heterogeneous. Almost all phylogenetic relationships were supported with 100% bootstrap value. Using the phylogenetic relationship from the first NJ tree, I computed the nucleotide substitution using General Time Reversible (GTR) model implemented in BASEML of PAML 4.8 (Yang 2007).

To investigate the precision of the distances, I split the concatenated sequences into 500 fragments. The splitting was done in a step-wise manner such that there is not overrepresentation of any fragments in a region. For example, the first fragment contains the 1st, 501st, 1001st, 1501st ... sites of the original alignment. Likewise, the second fragment contains the 2nd, 502nd, 1002nd, 1502nd... sites of the original alignment. This procedure checks the precisions of the distances and reduces overestimation due to regional variations.

6.2.10 Estimation of divergence times using MCMCtree

The divergence times were estimated using Bayesian method of relaxed molecular clock implemented in MCMCtree of PAML 4.8 (Yang 2007). MCMCtree estimates species divergence times using fossil calibration by performing Bayesian estimation under various molecular clock models. The estimation of divergence times in MCMCtree involves three major steps. I used the steps similar to what were used by Inoue et al. (2010). First, overall substitution rate was estimated assuming molecular clock. Next, the gradient and Hessian are estimated using the estimated substitution rate. Finally, the actual MCMC analyses was carried out using the gradient and Hessian as inputs.

For the estimation of the overall substitution rate, BASEML of PAML 4.8 was used with the concatenated multiple alignment file of second codon positions as the input. In addition, phylogenetic relationship was indicated with human gorilla divergence time of 11 million years ago, MYA (Suwa et al. 2007). GTR model was used and strict molecular clock was assumed. After the estimation of the overall substitution rate, gradient and Hessian matrix were produced by running MCMCtree on the original alignment and tree files. For gradient and Hessian estimation mode, usedata=3 was used. GTR model of nucleotide substitution was used. In addition, relaxed clock model (clock=2) was used. Finally, the divergence times were calculated by running MCMCtree. This time, the gradient and Hessian matrix file from the previous run was used by setting usedata=2. The prior substitution rate (rgene_gamma) was set to 1 15.8 based on the overall substitution rate. The sigma2-gamma was set to 1 1.05 based on the reported divergence time of 105MYA for elephant and human (Hedges et al. 2015). Relaxed clock model (clock=2) was used. Root age was restricted to less than 120MYA.

6.3 Results

6.3.1 Capybara genome sequencing and assembly

The genome sequences of capybara were determined using Illumina Miseq platform. Agillent QXT library preparation kit was used to prepare 650bp fragment size library. The prepared library was sequenced in three Miseq runs. For the Miseq sequencing, paired-end reads of 350bp and 250bp were produced. In total, ~42Gbp from 157.6 million reads of sequence data were produced. In each of the runs, an average of 75% of the total yield has the minimum quality value of Q30. The details of the reads are presented in Table A6.1 in Appendix 6.

For the assembly, CLC workbench and SOAP-denovo2 were used. The estimated genome size from k-mer (19-mer) distribution is 2.62Gbp. The assembled fragments cover 2.49Gbp, representing about 95% of the estimated genome size. However, 2.47Gbp (~94% of the estimated genome size) has a determined nucleotide. The longest fragment from the assembly was 140.2kbp with N50 length of ~8kbp (Table 6.1). The average read coverage was 15× with the peak around 13× (Figure 6.1).

Table 6.1: Capybara genome and assembly statistics

Parameter	Value
Estimated genome size (Gbp)	2.62
Total positions (Gbp)*	2.48
Determined positions (Gbp)**	2.47
Percent GC content	40.2
Minimum scaffold (bp)	200
Longest scaffold (kbp)	140.24
N20 length (bp)	18,744
N50 length (kbp)	7,745
N80 length (kbp)	2,193

*"Total positions" are all positions in all scaffolds, including Ns

**"Determined positions" are all non-N positions in all scaffolds

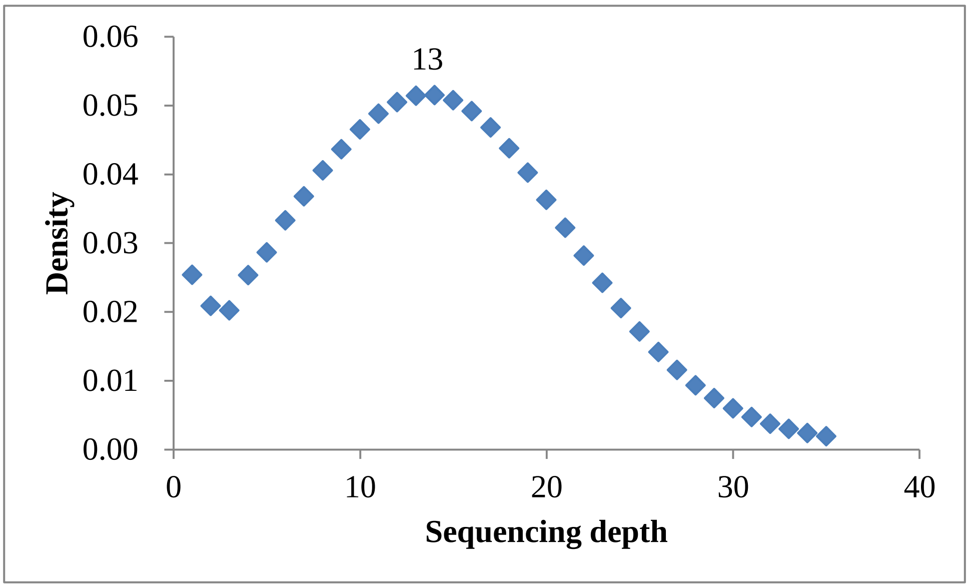


Figure 6.1: The sequencing depth from mapping of raw reads. The distribution of the sequencing depth per base position is shown

6.3.2 Extracting reliable genomic regions

Because the objective of the analyses is to investigate evolutionary rate dynamics, sequencing error should be minimized. To minimize the effects of sequencing errors, I decided to exclude positions that are likely to contain sequencing error. I employed the strategy that involved mapping of high-quality reads to the assembled genomes. The three basic steps include; (i) critical read filtering; (ii) mapping of filtered reads, and (iii) extraction of positions covered by the minimum depth. The detailed procedure of the filtering is presented in the Material and methods (Section 6.2.5). The quality of the filtered reads was examined using FastQC tool. Figure A6.1 in Appendix 6 shows the quality distribution of the filtered reads. After filtering, about 24Gbp (57% of the initial nucleotides) were available.

It is important to evaluate the effectiveness of the extracted high-quality regions. As described under Materials and Methods section, the maximum sequencing error rate per position was 10^{-6} (minimum quality and depth of Q20 and 3×, respectively). This value corresponds to 1 sequencing error in 1Mbp region. Also, I set the mapping quality threshold to 30. Theoretically, high-quality regions are used. To empirically evaluate the quality, I checked the nucleotide composition of 3rd codon positions (see Materials and Methods for detailed procedure). Sequencing error has been reported to be nucleotide-biased (Hansen et al. 2010; Schirmer et al. 2015). If there is sequencing error, huge differences would be expected in comparison to the closely related species. I therefore compared the nucleotide composition of the homologous third codon positions in capybara and guinea pig and found no significant difference (Figure 6.2B). Similar patterns were observed in first and second codon positions. The obvious effect of sequencing error would be in the genetic distance. If the threshold was not proper, using a stricter threshold would result in much lower genetic distance. To further evaluate the effectiveness of the quality, I applied stricter threshold (Q30 with 50bp). With this filtering, only ~17Gbp (or 40% of the initial bases) were used. I computed the synonymous substitution ration (dS) values for all homologous gene pairs using Kimura 2-parameter model in MEGA6. However, no obvious difference was found (Figure 6.2C). The slightly lower distance in Q30 thresholds can be explained by the alignment thresholds. Higher quality regions are fewer in length and number. Genes that are poorly conserved therefore would not be included. Taken together, the adopted threshold values (Q20 and 3X) seem to be appropriate.

6.3.3 Heterogeneity of rodent evolutionary rates

Having established the reliability of the extracted regions, I proceeded to investigate the evolutionary rate dynamics among rodents. Using GMAP (Wu and Watanabe 2005), I extracted the amino and nucleotide coding sequences from high-quality regions of capybara sequences using guinea pig coding sequences as the query. I then extracted the reciprocal best hits from the pairwise BLAST (Altschul et al. 1997) searches of capybara, naked mole rat, mouse and rat, using guinea pig as reference. I used human as outgroup species due to the quality of genome sequences.

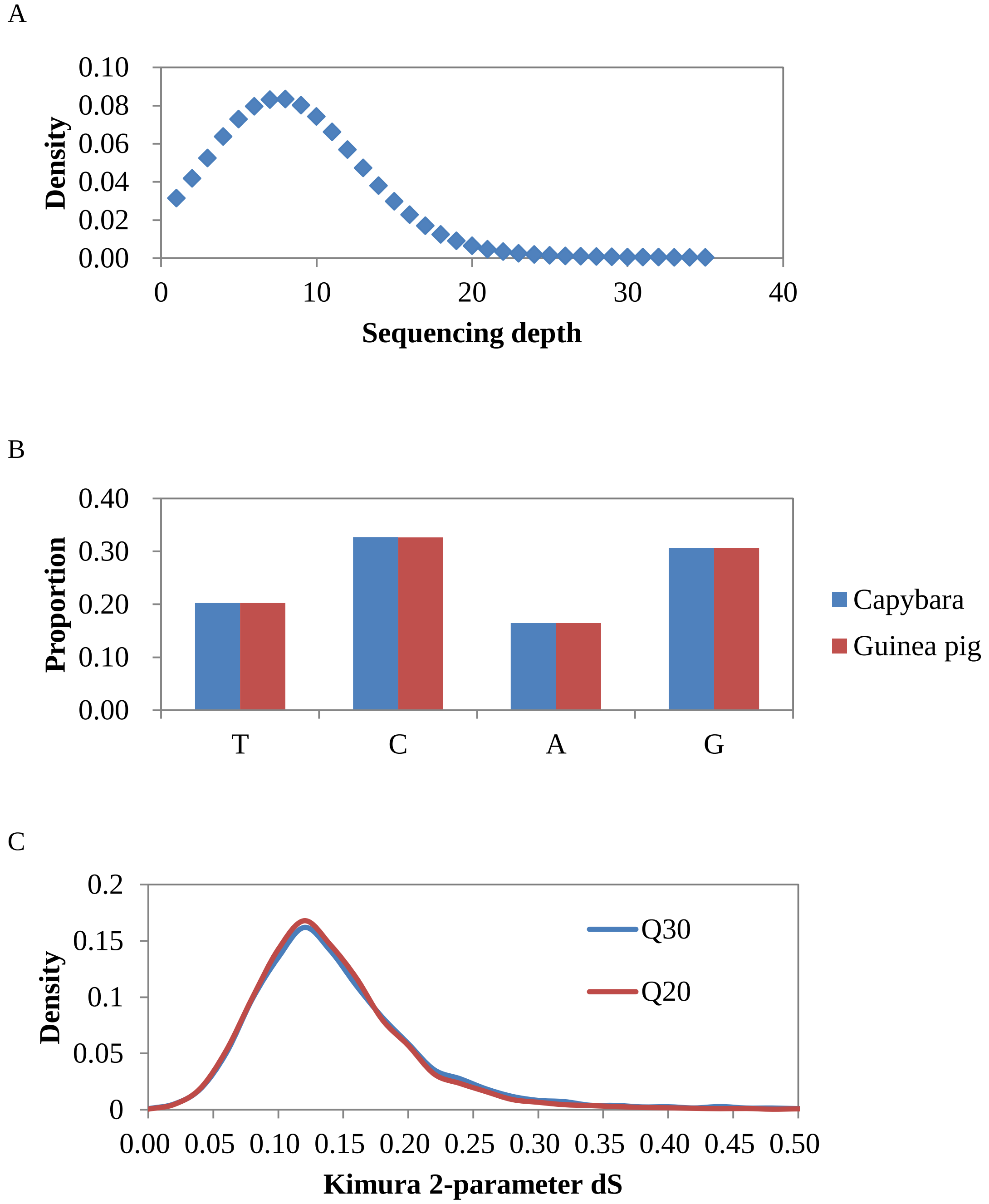


Figure 6.2: Assessing the effectiveness of quality filtering. (A) The distribution of per base depth of the quality-filtered reads. Positions with <3×or >30× coverage were masked. (B) The nucleotide composition of the third codon positions of capybara and guinea pig are not statistically different, demonstrating that nucleotide frequency shifting sequencing errors are negligible. (C)The distribution of genetic distance between capybara and guinea pig using Q20 and more stringent Q30 are similar.

After extracting amino acid sequences from the species, I conducted multiple sequence alignment using CLSUTALW (Thompson et al. 1994). In total, 9,815 homologous genes were used for the alignment. The phylogenetic tree of species, plotted from concatenated amino acid sequences, is shown in Figure A6.2 in Appendix 6. The tree correctly represents the expected phylogeny. After the alignment, the corresponding codons of the aligned amino acids were retrieved to get aligned codons.

I checked the assumption of molecular clock using the likelihood ratio test independently for the three codon positions using MEGA7-CC (Kumar et al. 2012). For the test, GTR model of nucleotide substitution and gamma distribution per site (k=5) were used. In the three codon positions, the molecular clock was violated (p value < 0.00001). The second codon positions have the highest number of nonsynonymous substitutions, and thus are with the highest purifying selection. To evaluate the branch lengths, the phylogenetic tree was first computed from the second codon positions. Figure 6.3A shows that the rates are heterogeneous among rodents. Another way of considering this is by evaluating the distance between each rodent species and the outgroup species (human). Figure 6.3B shows the distribution with distances plotted from 500 fragments (see Section 6.2.9 of Material and methods). Unexpectedly, capybara which is the largest rodent species has the longest branch. Similar results are found in first codon positions (Figures A6.3A and B in Appendix 6). The discussion about the differences requires considerations of genetic drift, population size, selection pressure and other complicated models.

For simplicity, I decided to focus on third codon positions because majority of the substitutions on third-codon positons are neutral. In third codon positions, substitution rate is very close to evolutionary rate. Pairwise distances were computed using the GTR model in PAML (Yang 2007). The Neighbor-Joining tree computed from the pairwise distances of third codon positions is shown in Figure 6.4A. The genetic distances between human and the rodent species examined are shown in Figure 6.4B. Figures 6.4A and B show that mouse and rat (representatives of Myomorpha) have significantly higher genetic distances than capybara, guinea pig and naked mole rat (representatives of Ctenohystrica). This result demonstrates the heterogeneity of rodent substitution and evolutionary rates.

6.3.4 Comparing rodent and primate evolutionary rates

To evaluate the previously reported higher evolutionary rates in rodents compared to primates, I included four primate species and African elephant (as outgroup). The total number of homologous genes that could be used was 8,346. The phylogenetic relationships were confirmed using amino acid (Figure A6.4 in Appendix 6), first codon position (Figure A6.5A in Appendix 6) and second codon positions (Figure A6.5B in Appendix 6). Focusing on mostly neutrally evolving third codon positions, Figure 6.5A shows the phylogenetic tree of the species.

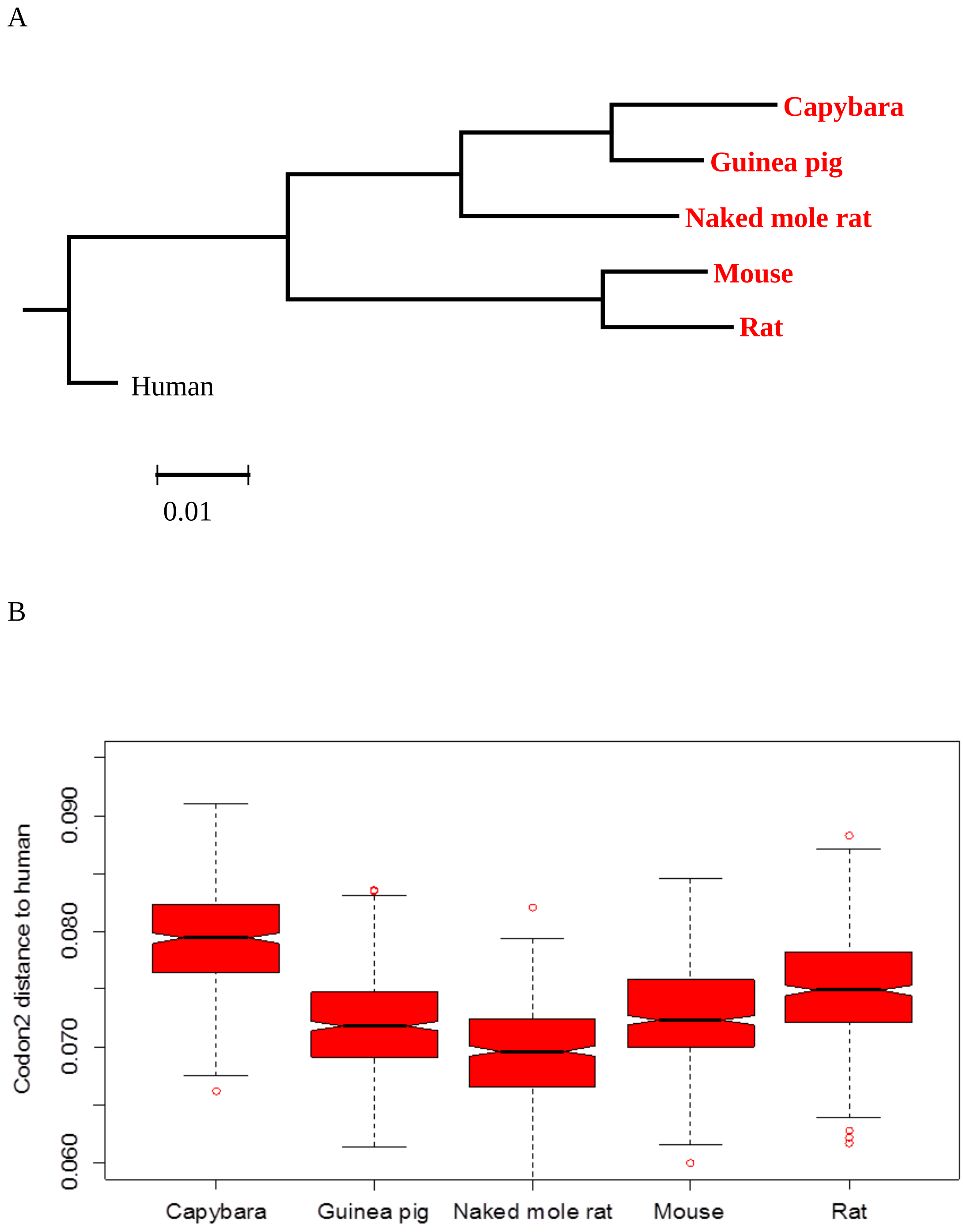


Figure 6.3: The evolutionary distance dynamics of rodent second codon positions. Selection is strongest on the second codon positions. (A) Phylogenetic tree computed from second codon positions. (B) Second codon position distance to human.

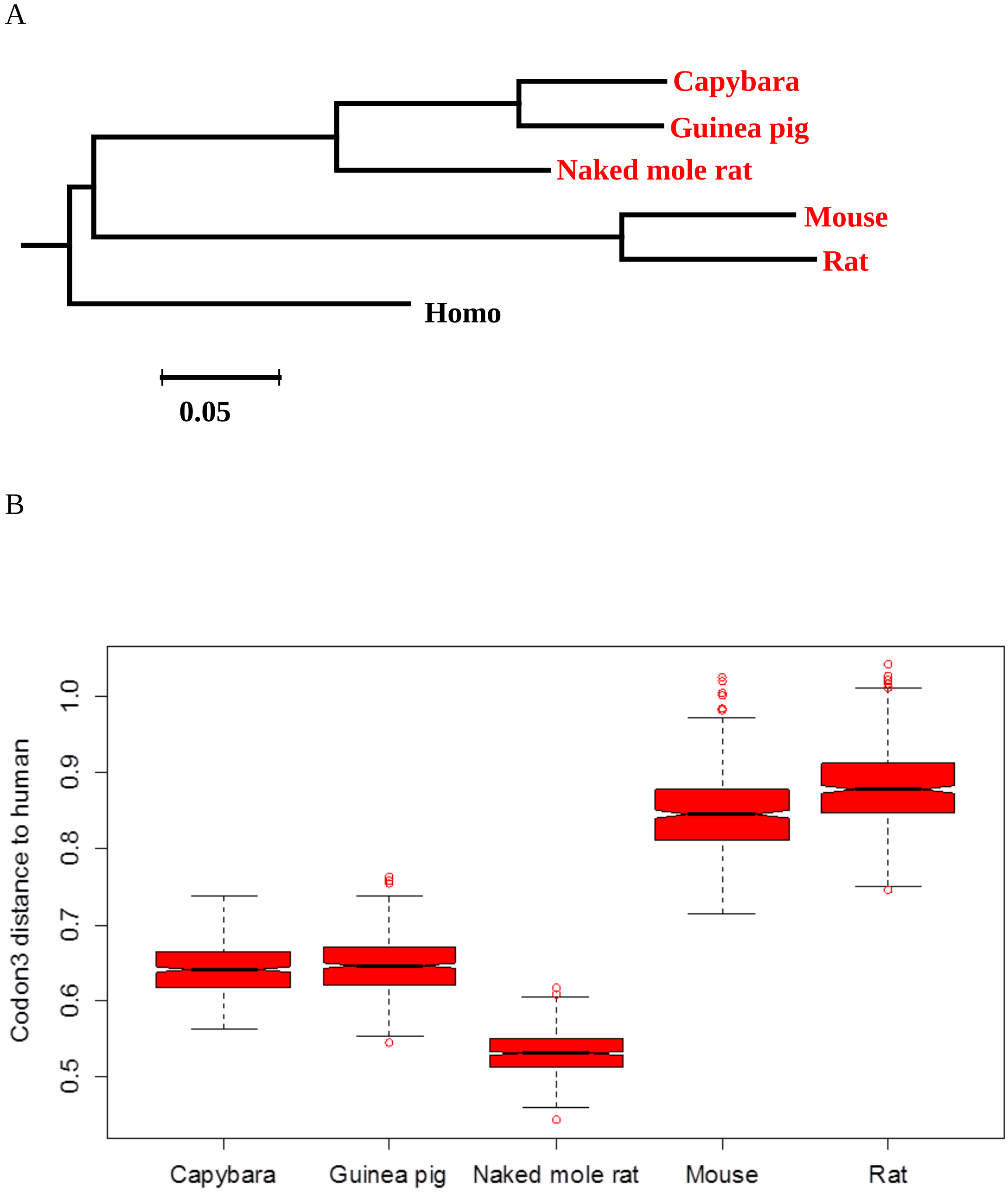


Figure 6.4: The evolutionary distance dynamics of rodent third codon positions. Majority of third codon positions are neutrally evolving. (A) Phylogenetic tree computed from third codon positions. (B) Third codon distance to human.

Although there is heterogeneity among rodents, primate species tend to have lower genetic distances as reported by previous studies (Wu and Li 1985; Li and Wu 1987; Takahashi and Saitou 2012; Babarinde and Saitou 2013). The genetic distances between elephant and the species are shown in Figure 6.5B. Considering the rodent evolution in primates' context, the heterogeneity of rodents is clearly presented. From Figure 6.5B, the species included can be roughly classified into three groups based on the genetic distance to elephant. The first group contains the Myomorpha, the second group contains the Ctenohystrica, while the third contains the Primates.

For the results so far, I selected a few species based on the quality of the genome sequences and the body size. The important question is whether the species truly represent the lineages. The main challenge in including more samples is in the quality of genome sequences. Because of this limitation, it would be difficult to give a detailed report using many publicly available genome sequences. Nevertheless, inclusion of more species would give broader overview. Therefore, I downloaded amino acid and coding sequences of primate and rodent species in Ensembl database. For better comparisons, I also downloaded Largomorpha, carnivore and cetartiodactyl species. The phylogenetic relationship is shown in Figure A6.6 in Appendix 6. All phylogenetic relationship is supported with 100% bootstrap values, except for the cluster between Sciuridae and mouse-related rodent lineage. Indeed, the analysis of rodent lineages in Chapter 2 shows that this branching is wrong.

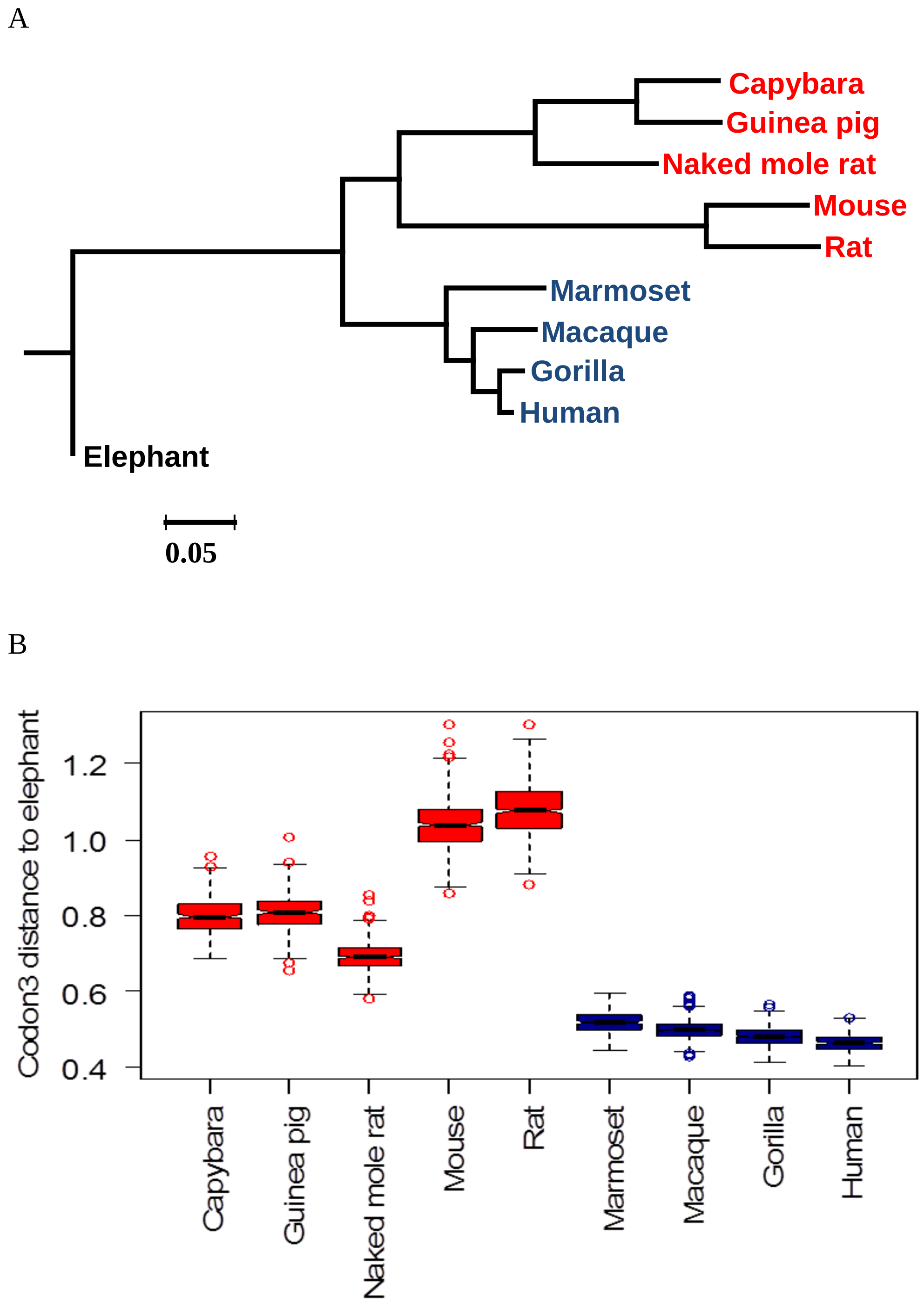
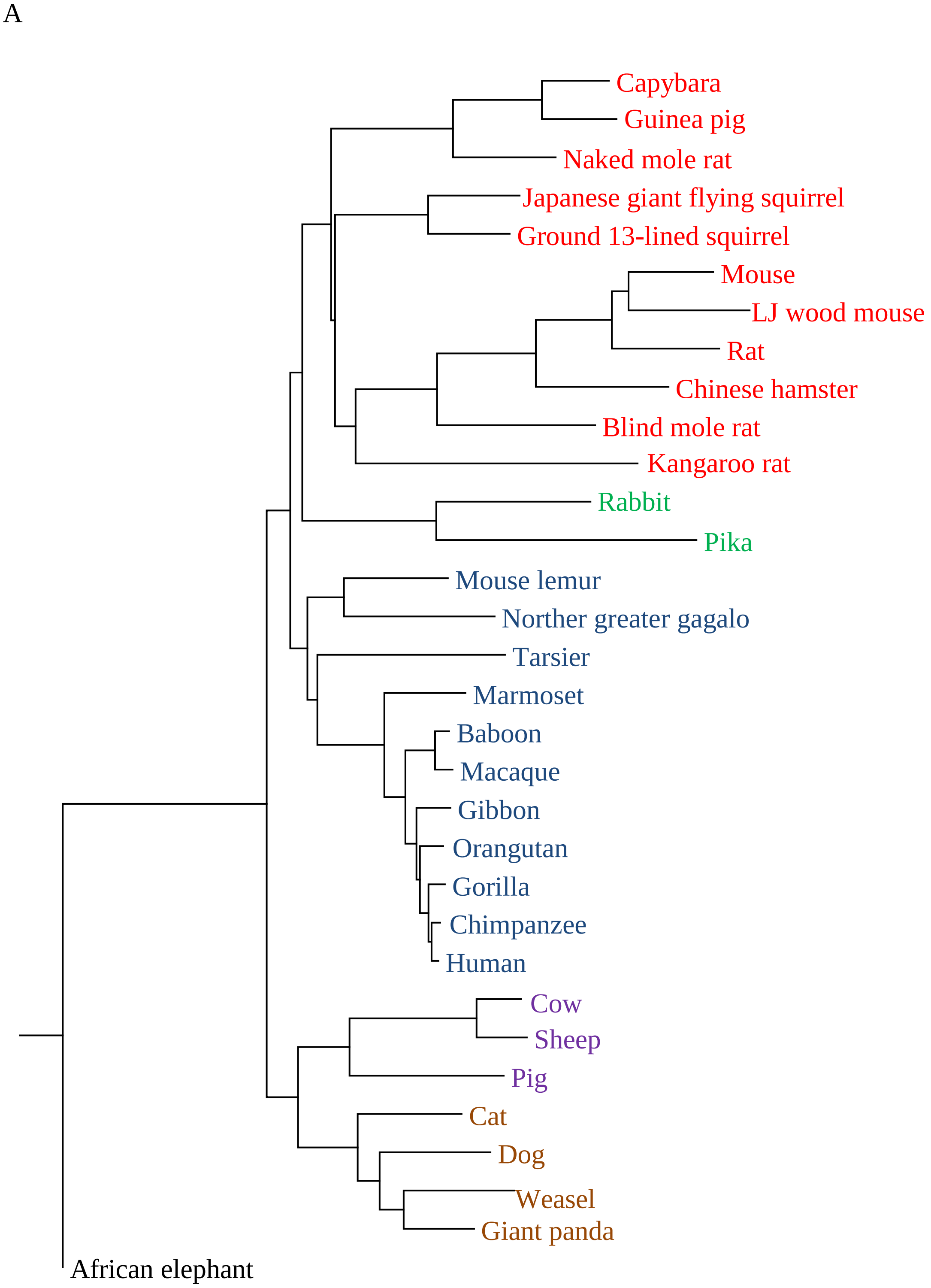


Figure 6.5: Evolutionary distance difference between primates and rodents. (A) Phylogenetic tree computed from third codon positions. (B) Third codon distance to elephant.

Figure 6.6A shows the phylogenetic relationships of the 32 species computed from third codon positions. Overall, rodents tend to have higher evolutionary rates than the other lineages. The observation is more obvious when the distance to African elephant is considered (Figure 6.6B). The average of rodent distance to elephant is significantly higher than primate, cetartiodactyl and carnivore distances. Largomorpha is phylogenetically the closest lineage to rodents, and they also have high rates. However, only two species were used, and the quality of pika genome is not excellent.

Another striking uniqueness of rodent lineage is the heterogeneity of evolutionary rates. Compared to other lineages, rodents have the highest range. Because of the limitation in the availability of genome data, equal numbers of species were not used. However, in both primates and rodents, 11 species were used in each lineage. Whereas the rodent range of distances to elephant is 0.4504, the range in primate is just 0.1100. It is important to note that similar results were found in first codon (Figure A6.7A in Appendix 6) and second codon (Figure A6.7B in Appendix 6) positions.


A
Capybara
Guinea pig
Naked mole rat
Japanese giant flying squirrel
Ground 13-lined squirrel
Mouse
LJ wood mouse
Rat
Chinese hamster
Blind mole rat
Kangaroo rat
Rabbit
Pika
Mouse lemur
Norther greater gagalo
Tarsier
Marmoset
Baboon
Macaque
Gibbon
Orangutan
Gorilla
Chimpanzee
Human
Cow
Sheep
Pig
Cat
Dog
Weasel
Giant panda
African elephant

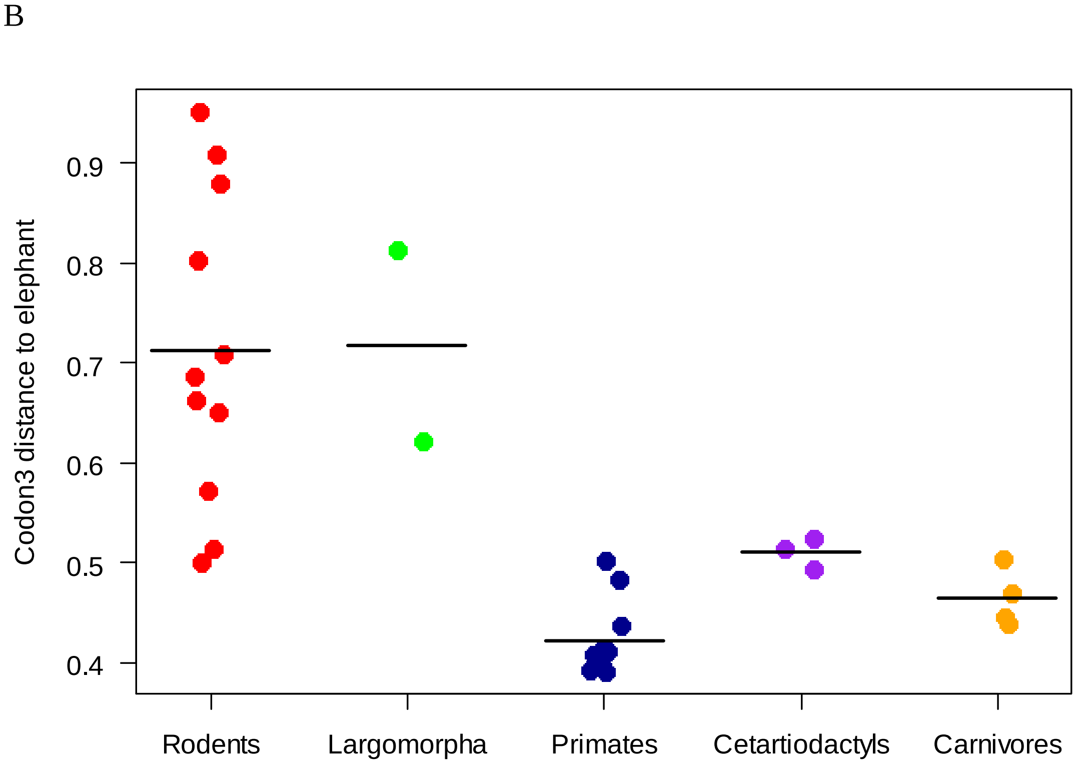


Figure 6.6: Evolutionary distance differences among more orders and species. (A) Phylogenetic tree computed from third codon positions. (B) Third codon position distance to elephant. The horizontal lines are the mean values for each order. Red, green, blue, purple and orange colors were used for Rodentia, Largomorpha, Primates, Cetartiodactyla and Carnivora, respectively.

6.3.5 Estimation of divergence times

Genetic distance between species comprises an evolutionary history and may not necessarily imply the true picture of what is happening in the extant species. A more representative picture would be to examine the evolutionary rate dynamics over phylogenetic timescales. The computation of actual rate from the genetic distance requires the accurate divergence time estimation. In order to obtain accurate results, only the initial 10 species of trustable quality were used. To decide which data set to use, I computed coefficient of variation (CV) of the distances between elephant and all the species from the three codon positions. The CV value of 0.119 in second codon position is lower than 0.145 and 0.335 in first and third codon positions, respectively. Selection forces acting on second codon positions seem to trim down the differences in evolutionary rates across species. Therefore, I decided to use second codon positions to estimate divergence time.

For the computation of divergence times, 10-12 MYA gorilla-human split (Suwa et al. 2007) was used for calibration. This value is higher than most reports of human-gorilla split. The median divergence time in Time Tree is 8.3MYA. In addition, the root was placed to below 120 MYA. The results of major splits are presented in Table 6.2. More comprehensive results showing the divergence times and 95% confidence intervals are presented in Figure A6.8. For comparison, divergence times reported by Hedges et al. (2015) in Tree of Life are also presented in Table 6.2. For most splits, the values computed were higher than the reported by Hedges et al. (2015). For example, the estimated 99.6MYA of human-mouse divergence is higher than reported 90MYA.

6.3.6 Violation of body-size hypothesis

Using the divergence time estimation, the evolutionary rates per branch was determined from third codon substitution. Figure 6.7 shows the rate heterogeneity across evolutionary timescales. Specifically, evolutionary rates in rodent lineage tend to be significantly higher than those in primate lineage. The highest rate (3.4×10^{-9} substitution/site/year) was observed in the common ancestor of mouse and rat (murines). Coincidentally, murines have the highest number of species. The lowest evolutionary rate (0.8×10^{-9} substitution/site/year) on the other hand was observed in human lineage in agreement previous report (Yi 2013).

Having established the heterogeneity of evolutionary rates across phylogenetic timescales, I asked if the body size accurately predicts the evolutionary rates. The rate heterogeneity suggests that the comparison of the evolutionary rates of distantly related species in relation to some particular features of extant species may not be appropriate. For example, slow down of evolutionary rates was observed in human and naked mole rat lineage. Therefore, focusing on closely related species with drastically different features would be a better approach to investigate factors that affect evolutionary rates. For the investigation of body size effect, I focused on three pairs of closely species with significantly different body sizes. The first pair is capybara (62kg) and guinea pig (1kg) from Ctenohystrica. The second pair is mouse (30g) and rat (500g) from Myomorpha. The last pair is human (62kg) and gorilla (180kg) from Primates.

Table 6.2: Estimation of divergence times

		Divergence time in million years ago (MYA)			
Species 1	**Species 2**	**MCMC tree**	**Min 95% CI**	**Max 95% CI**	**Tree of life***
Human	Gorilla	11.1	10[c]	12[c]	8.9
Human	Macaque	31	18.1	52.9	29.1
Human	Marmoset	42.9	29	64.3	43.1
Human	Elephant	108.9	99.9	119.8	105
Human	Mouse	99.6	81.4	115.9	90.9
Guinea pig	Capybara	28.6	16.7	43	24.6
Guinea pig	NMR**	53.8	36.6	71.7	46.1
Guinea pig	Mouse	82.9	56	102.8	77.2
Mouse	Rat	24.1	14.5	35	22.6

* Values published in Hedges et al. 2015

** NMR = Naked mole rat

[c] Calibration point

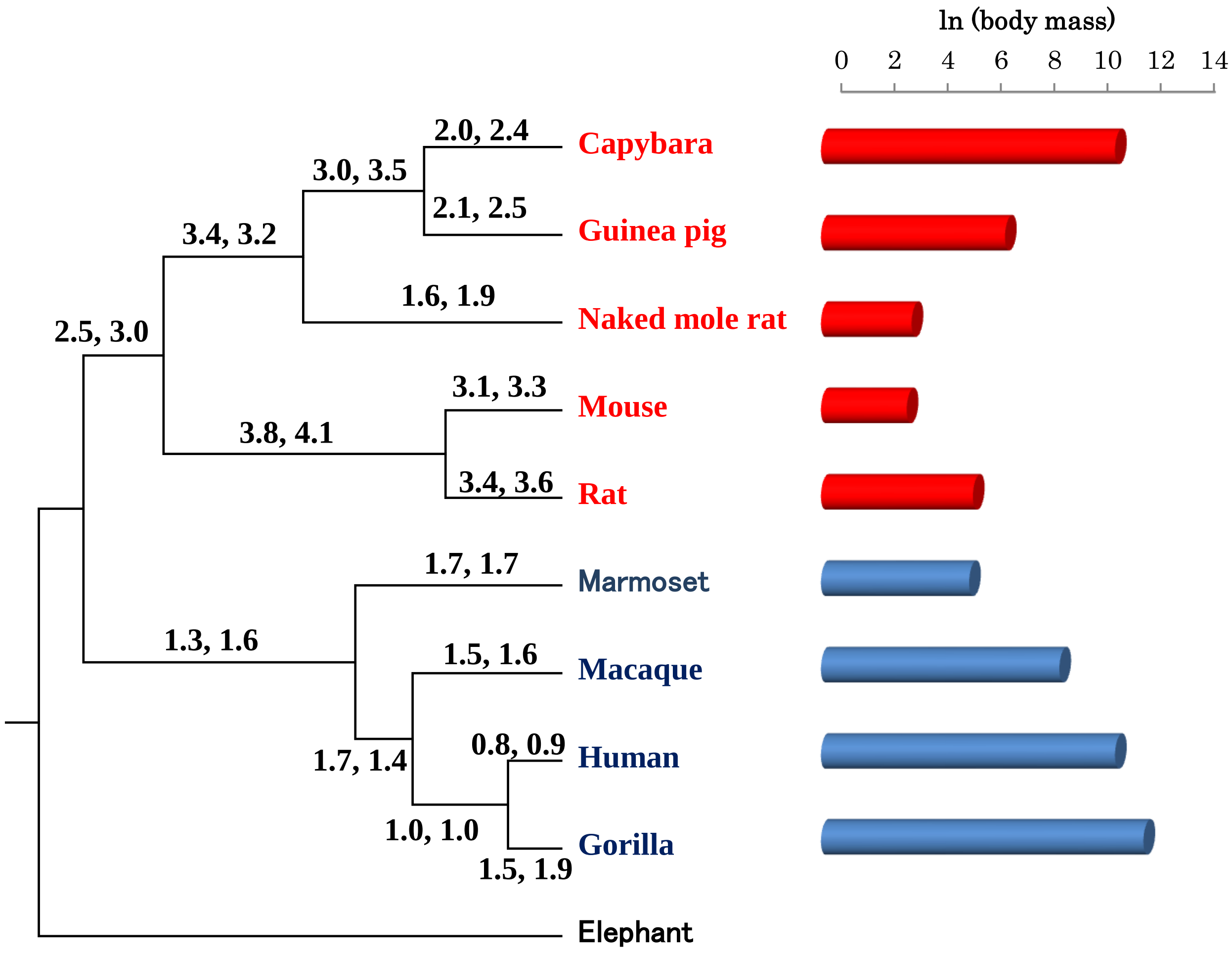


Figure 6.7: Evolutionary rate dynamics and body size effect. The left panel shows the phylogenetic relationship among species. The values on the branches are the evolutionary rates in ×10^{-9} substitutions per site per year. The first and the second values were computed from MCMCtree and Tree of Life divergence times, respectively. The right panel shows the body sizes in natural log scale of gram.

As would be predicted from body-size effect, capybara substitution rate (2.0×10^{-9} substitution/site/year) is lower than guinea pig substitution rate (2.1×10^{-9} substitution/site/year), although the difference does not accurately match the body size differences. Interestingly, naked mole rat, which is the smallest of the three Ctenohystrica species examined, has the lower substitution rate (1.6×10^{-9} substitution/site/year). In the other two pairs of species considered, the observation is different from what would be expected from body size effect. Mouse's 3.1×10^{-9} substitution/site/year is lower than rat's 3.4×10^{-9} substitution/site/year, despite the latter being bigger in size (Figure 6.7). Similar results were found in human and gorilla. These observations challenge the generality of body-size effect on evolutionary rate differences.

6.4 Discussion

In the study, I have reported the first draft genome sequence data of capybara, the largest extant rodent species. This report is from three runs of Illumina Miseq platform. The nature of the sequence library prepared restricted the proper assembly formation. Consequently, in-depth analyses of capybara genome structure could not be properly described. However, a large proportion of the predicted genome size was covered in this study, albeit in short fragments. Also, I have shown that the estimated genome size and GC content are comparable to what is found in other rodent species. In addition, I have shown that the homologs of majority of the annotated guinea pig genes could be found in capybara genomes, at least in partial fragments.

To conduct a reliable evolutionary analysis, I have adopted a procedure that minimizes sequencing errors. This procedure ensures a reliable result from limited genome sequencing depth. I have shown that this procedure minimizes sequencing error to an acceptable level. Unlike majority of previously published studies with limited data, I have focused on whole genome analyses consisting of millions of nucleotide positions. This analysis gave a comprehensive understanding of the overall evolutionary rates among rodent species with reference to primate species. The huge number of sites also minimizes the effect of the variation among genes.

The three codon positions have different selection pressures. For example, the only redundancy in second codon positions is a two-fold redundancy in stop codon. There are only three two-fold redundancies in first codon positions. On the contrary, third codon substitutions are mostly synonymous. In fact, half of the total possible combinations of first and second positions have four-fold redundancies at the third codon positions. Therefore, while second and first codon positions reflect the impact of selection to a large extent, third codon positions are mostly neutrally evolving. This difference explains some differences in distance patterns. Because mutation rate is approximately equal to substitution rate on third codon positions, rate comparisons were based on third codon positions.

Using the genome data of more rodent species, I have shown that rodent evolutionary rates are highly dynamic. Specifically, I have shown that the substitution rates in capybara, guinea pig and naked mole rat (members of Ctenohystrica) are lower than those of mouse and rat (representatives of Myomorpha). Whether these features are actually the representatives of the respective clades requires further analyses involving more species. Of particular interest is the fact that the two Myomorpha species used in this analysis belong to the family Murinae. This family has the highest number of species among mammals. Strikingly, upon the incorporation of more species, murine species (mouse, large Japanese wood mouse and rat) showed higher distance than other Myomorpha species (Figure 6.7). The high evolutionary rate in Murinae might be related to higher number of species found in this family. This association is in line with Mutation-driven hypothesis of evolution (Nei 2013). Mutation-driven hypothesis postulates that species with higher mutation rates would have higher morphological diversity and chance to adapt to new environment. Therefore, higher mutation rates in murines might have contributed to higher number of species in the family.

The availability of high-quality primate genome data makes it possible to re-evaluate the long-standing report of higher evolutionary rates in rodents. The earlier reports (Wu and Li 1985; Li and Wu 1987; Li et al. 1996) used few genes. In this study, I have used genome data involving large number of genes. Although there is heterogeneity among rodents, all examined rodents tend to have higher evolutionary rates than primates. The rates in carnivores and cetartiodactyls are not as high as in rodents (Figure 6.6). This suggests that the accelerated evolutionary rate started in rodent common ancestor. The phylogenetic analyses of evolutionary rates indeed showed that the substitution rates per site per million year is comparatively high in rodent common ancestor, becoming higher in Murinae common ancestor. This observation cautions the assumption of molecular clock across linages and evolutionary times.

Several factors have been proposed for the reason for the differences in evolutionary rates across species. Capybara genome allows me to scrutinize the body-size effect. However, the results show that body-size does not always properly predict evolutionary rates. Whether this lack of association are because the species evaluated are exceptions, or there is actually no relationship between body size and evolutionary rates cannot be concluded. Unfortunately, low number of good-quality genomes limits the extensive evaluation of the body-size effect. In any case, my analyses show that differences in body size do not generally predict difference in evolutionary rates. Interestingly, Huttley et al. (2007) came to a similar conclusion while examining the relationship between generation interval and evolutionary rates. The implied relationship between evolutionary rates and generation interval, body size, metabolic rate and reported features assume the constancy of replication error rate across lineages. As reported in naked mole rat (Kim et al. 2011), the replication error rates vary across species. In conclusion, my analyses show the heterogeneity of evolutionary rates across species, lineages and phylogenetic timescales. This heterogeneity cannot be properly explained by body size. It would be interesting to evaluate other factors such as the aquatic adaptation.

CHAPTER SEVEN

GENERAL CONCLUSION

7.1 Overall Summary

The large number of species, high phenotypic diversity and ecological success across various habitats make rodents excellent for the understanding of mammalian evolution. Adequate understanding of rodent evolution implies automatic understanding of about half of all mammalian species. And because of shared evolutionary mechanisms, the knowledge gained from rodent evolution can be employed in further understanding of mammals as a taxonomic Class. Although the evolutionary study of rodents is very promising for the understanding of mammalian evolution, genomic studies have been restricted to very few species, mouse and rat being the most beneficiaries. In this study, I have presented the *de novo* genome sequences of two phylogenetically strategic species. One of the species, capybara (presented in *Chapter Six*), is the largest rodent species and is endemic to South America. The second species, Japanese giant flying squirrel (musasabi), has the gliding ability and is endemic to Japan. The Sequencing report was presented in *Chapter Two*. In addition, I also determined the transcriptome data of large Japanese wood mouse (akanezumi). The results were presented in (*Chapter Five*).

Together with some Japanese scientists, we are now sequencing the whole genome of large Japanese wood mouse. The sequencing project is already at an advanced stage.

The main focus of this study is to unravel genomic changes associated with rodent evolution. For these analyses, regulatory evolution cannot be overlooked. My main approach is to analyze conserved noncoding evolution as signatures for regulatory evolution. In *Chapter Three*, I combined evolutionary and computational approaches to analyses sequence, expression and ChIP-Seq data. My analyses show that CNSs are indeed associated with more conserved expression. Specifically, CNS-associated genes have higher expression correlation between mouse and human. I also report nonrandom distribution of CNSs. CNSs tend to be overrepresented around genes associated with developments, transcription and nervous system, and are underrepresented around house-keeping genes and genes associated with response to stimulus, defense and immunity. Closer investigation of genomic locations of CNSs revealed interesting observations. Of particular interest is the observation that CNS-gene physical distance tends to be conserved over evolutionary time and that CNS flanking genes are often the target genes. This suggests that the appropriate location of CNS is important for its gene regulatory activity.

Having established the relevance of CNS in gene expression regulation, the understanding of the dynamics across lineage is important. Particularly, is rodent CNS evolution different from other mammals? *Chapter Four* answers this question. By comparing CNS evolution in rodents and three other mammalian orders including primates, carnivores and cetartiodactyls, I showed CNS turnover rate in rodents is very high. Rodents lost more ancient CNSs and gained less. Focusing on the evolution of CNSs across evolutionary timescales, I reported that older CNSs tend to have stronger constraints, suggesting the involvement of CNSs in recent evolutionary novelty. The evaluation of rodent evolutionary dynamics in Chapter Two shows that the rate of loss of ancestral CNSs vary across rodent lineages. Specifically, mouse-related lineage lost the highest number of ancestral CNSs while squirrel-related lineage lost the least.

In *Chapter Five*, I report the transcriptome analyses of large Japanese wood mouse. Using sequence data, I established the phylogenetic relationship of large Japanese wood mouse, laboratory mouse and rat. Although the splitting of the species appeared to be rapid, the expression dynamics could reproduce the correct phylogeny. This shows how rapid the gene expression dynamics evolve among closely related species of rodents. Further, CNS-associated genes tend to have lower expression level and expression conservation in liver samples of rodents species, suggesting that CNSs are not likely to be involved in liver gene expression regulation.

The final aspect of the study was to investigate the evolutionary rate dynamics among rodents. Using newly determined capybara genome data, I showed that the previously reported high evolutionary rate in mouse and rat is also found in Ctenohystrica, though to a lower extent. This shows a strong heterogeneity among rodents. My analyses also show that the negative correlation between body size and substitution rate is not always true. There are some big-bodied species with faster substitution rate, and some small-bodied species with slower substitution rate.

Overall, I have shown rodent CNSs are evolving very fast. The fast evolution of CNSs implies a high turnover of regulatory elements. This high regulatory turnover might have contributed immensely to the rodent phenotypic dynamics. High evolutionary rates in rodents might have contributed to this regulatory evolution, as murines with very high rates lost the highest number of ancestral CNSs. It is important to note that loss of CNSs might not necessarily imply loss of regulatory elements. It can also imply that the regulatory elements, though functional, are not conserved. In a similar reasoning, fewer ancestral CNSs do not necessarily imply fewer ancestral regulatory elements. Interestingly, murines have the highest number of species, not only among rodents, but also among mammals. Also, I show that gene expression evolve so rapidly among species that diverged in rapid succession. The violation of body-size hypothesis suggests that rodent high evolutionary rate, and number of species might not just be due to the smaller body size.

7.2 Biological and evolutionary implications

My analyses show series of evidence suggesting that CNSs are very important in gene expression dynamics. The high CNS turnover rate in rodents suggests that regulatory elements in rodents evolve faster than in other mammalian lineages. This high regulatory evolution in rodents might have contributed to rodent phenotypic diversity and ecological success (as hypothesized by Carroll 2008). The evolutionary changes indeed happen at nucleotide level. The high regulatory evolution is because of high rodent substitution rate. This study therefore supports the Mutation-driven hypothesis (Nei 2013) that species with higher mutation rates are more evolutionarily successful. The success might be more related to ecological success than to actual number of species. Too high mutation rates usually come with cost (Kimura 1960). Unfortunately, the actual estimation of population sizes is still difficult.

In reconciling the evolutionary benefits of high mutation rate and the associated cost, it may be more informative to discuss the concepts of environment and niche. Higher mutation rates imply higher number of ecological success. The species with higher mutation rates would have more raw materials that can be used to colonize new niche (Nei 2013). After niche colonization, the reproductive isolation leads to speciation (Rice 1987; Palumbi 1993). In the new niche, only few or restricted mutations are favorable. With strong purifying selection, unfavorable mutations would be removed. Consequently, the evolutionary rates, though not necessarily mutation rates, would be reduced.

The story of rodent ecological success cannot be complete without discussing migration. Even with “abundant resources” to establish in a new environment, the species would still have to migrate to the new location. This becomes a major problem for migration over water as many rodent species cannot swim or fly over long distance. Small-bodied rodent species would therefore be more successful in terms of migration, because they can “hitchhike” with human. To establish in a new environment however, species with higher evolutionary rates are more advantageous.

7.3 Contributions to knowledge

The *de novo* genome sequences and transcriptome data are invaluable resources for future evolutionary and genetic studies. From the evolutionary point of view, the association of CNS evolution to phenotypic diversity and ecological success is an important discovery. Although the regulatory evolution has been suspected to contribute to morphological evolution (King and Wilson 1975; Carroll 2008), CNS evolution had not been considered in light of phenotypic diversity. In this study, I have shown an interesting association. On the CNS activity, I report that CNS-gene physical distances are evolutionarily conserved, suggesting the importance of the actual genomic location of regulatory elements in proper gene regulation. This knowledge is important in genome editing experiments. Finally, the analyses shed more light to factors affecting evolutionary rate differences across species. Specifically, cases in which body size poorly predicts evolutionary rates are shown.

7.4 Future direction

Having investigated rodent evolution at the genomic level, the next line of analyses would be to focus on some interesting phenotypes. It would be important to establish which genomic regions are responsible for which phenotypes. Having published the first draft genome of capybara, the largest living rodent species, the first phenotype that I would like to focus will be body size. Also, the investigation of the genomic regions associated with the gliding ability of Japanese giant flying squirrel would be interesting. Using computational analyses, I would like to identify the candidate genomic regions responsible for the phenotypes.

Whereas computational analyses can identify the candidate regions, they are not enough to establish causal relationship. To do this, molecular experiments would be very important. Fortunately, there has been a great advancement in genome editing technique. Although the species are not model organisms, the identified candidate regions can be tested using model species. Personally, I have no experience in this kind of detailed molecular experiments. With the collaborations with other scientists, I hope some causal relationships can be established.

Various hypotheses have been proposed to explain evolution of species. My study particularly supports mutation-driven hypothesis. I would want to investigate other hypotheses in details. Particularly, niche-filling hypothesis and key innovation would be investigated. About the mutation-driven and niche-filling hypothesis, it would be important to further evaluate the impacts on population size dynamics. Recent advances in population genomics have made it possible to evaluate population size history even from a single genome (Li and Durbin 2011). Incorporating the same principles, I would like to investigate the population size dynamics of highly and slowly evolving lineages.

References

Ahituv N, et al. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol 5:9.

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF, et al. 2005. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336: 736-739.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Altschul SF, et al. 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Ando M, Shiraishi S. 1993. Gliding flight in the Japanese giant flying squirrel *Petaurista leucogenys*. J. Mammal. Soc. Jpn 18:19?32.

Antoniv TT, De VS, Wells D, Denton CP, Rabe C, de Crombrugghe B, Ramirez F, Bou-Gharios G, et al. 2001. Characterization of an evolutionarily conserved far-upstream enhancer in the human alpha 2(I) collagen (COL1A2) gene. J Biol Chem. 276:21754-21764.

Babarinde IA and Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biol Evol. 5:2330-2343.

Beck RM, Bininda-Emonds OR, Cardillo M, Liu FR, Purvis A. 2006. A higher-level MRP supertree of plancetal mammals. BMC Evol Biol. 6:93.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004.

Ultraconserved Elements in the Human Genome. Science 304:1321-1325.

Bejerano G, et al. 2004. Ultraconserved elements in the human genome. Science 304:1321-1325.

Bhatia S, Monahan J, Ravi V, Gautier P, Murdoch E, Brenner S, van Heyningen V, Venkatesh B, Kleinjan DA. 2014. A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. Dev Biol. 387:214-228.

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol 9:71.

Blanga-Kanfi S, et al. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol. 9:71.

Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res. 15:98?110.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343-8.

Bromham L, Penny D. 2003. The modern molecular clock. Nat Rev Genet. 4:216-24.

Bromham L, Rambaut A, Harvey PH. 1996. Determinants of rate variation in mammalian DNA sequence evolution. J Mol Evol 43:610-621.

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. Philos Trans R Soc Lond B Biol Sci. 366: 2503-2513.

Buffenstein R, Yahav S. 1991. Is the naked mole-rat Heterocephalus glaber an endothermic yet poikilothermic mammal? J Therm Biol 16:227–232.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bures S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. doi: 10.1101/gr.196485.115.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics 187:1225-1234.

Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res. 15:1061-1072.

Carnaval AC, Waltari E, Rodrigues MT, Rosauer D, VanDerWal J, Damasceno R, Prates I, Strangas M, Spanos Z, Rivera D, Pie MR, Firkowski CR, Bornschein MR, Ribeiro LF, Moritz C. 2014. Prediction of phylogeographic endemism in an environmentally complex biome. Proc Biol Sci. 281:1792.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25-36.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415-418.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jonsson B, Schluter D, Bell MA, Kingsley DM. 2009. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327(5963):302-5.

Chen HP, et al. 2008. Screening reveals conserved and nonconserved transcriptional regulatory elements including an E3/E4 allele-dependent APOE coding region enhancer. Genomics 92:292-300.

Ciochon RL, Dolores RP, Robert GT. 1990. Opal phytoliths found on the teeth of the extinct ape *Gigantopithecus blacki*: Implications for paleodietary studies. Proc Natl Acad Sci U S A. 87: 8120-8124.

Crawford GE, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 16:123-131.

Cretekos CJ, et al. 2008. Regulatory divergence modifies limb length between mammals. Genes Dev. 22:141-151.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 107:21931-21936.

De S, Teichmann SA, Babu MM. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. Genome Res. 19: 785-794.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing Chromosome Conformation. Science 295:1306-1311.

Dickerson JE, Zhu A, Robertson DL, Hentges KE. 2011. Defining the role of essential genes in human disease. PLoS One. 6(11): e27368.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci. 279:3491-500.

Dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PC, Yang Z. 2015. Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. Curr Biol. S0960-9822(15)01177-X.

Dostie J, Dekker J. 2007. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2:988-1002.

Drake JA et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38:223-227.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38(2):223-7.

Duret L, Dorkeld F, Gautier C. 1993. Strong conservation of noncoding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res. 21:2315-2322.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19:362-365.

El-Kasti MM, Wells T, Carter DA. 2012. A novel long range enhancer regulates postnatal expression of Zeb2: implications for Mowat-Wilson syndrome phenotypes. Hum Mol Genet. 21:5429-5442.

Elgar G. 2009. Pan-vertebrate conserved noncoding sequences associated with developmental regulation. Brief Funct Genomic Proteomics 8:256-265.

Ferraz, KMPMB, Bonach K, Verdade LM. 2005. Relationship between body mass and body length in capybaras (*Hydrochoerus hydrochaeris*). Biota Neotropica 5:BN03405012005.

Frankel N, et al. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466:490?493.

Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M,, et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. Nat Biotechnol. 18:181-186.

Graur D, et al. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 5:578-590.

Gupta S, et al. 2008. Predicting human nucleosome occupancy from primary sequence. PLoS Comput Biol 4:e1000134.

Halder G, Callaerts P, Gehring WJ. 1995. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. Science 267(5205):1788-92.

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 38, e131.

Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genetics 16:369-372.

Harrison PW, Wright AE, Zimmer F2, Dean R, Montgomery SH, Pointer MA, Mank JE. 2015. Sexual selection drives evolution and rapid turnover of male gene expression. Proc Natl Acad Sci U S A. 112:4393-8.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22:160-74.

Hemberg M, et al. 2012. Integrated genome analysis suggests that most conserved noncoding sequences are regulatory factor binding sites. Nucleic Acids Res. 40:7858-7869.

Hiller M, Schaar BT, Bejerano G. 2012. Hundreds of conserved noncoding genomic regions are independently lost in mammals. Nucleic Acids Res. 40:11463-11476.

Hirota T, Hirohata T, Mashima H, Satoh T, Obara Y. 2004. Population structure of the large Japanese field mouse, *Apodemus speciosus* (Rodentia: Muridae), in suburban landscape, based on mitochondrial D-loop sequences. Mol Ecol. 13:3275-82.

Huchon D, Madsen O, Sibbald MJ, Ament K, Stanhope MJ, Catzeflis F, de Jong WW, Douzery EJ. 2002. Mol Biol Evol 19:1053-65.

Inoue JG, Miya M, Lam K, Tay BH, Danks JA, Bell J, Walker TI, Venkatesh B. 2010. Evolutionary origin and phylogeny of the modern holocephalans (chondrichthyes:

chimaeriformes): a mitogenomic perspective. Mol Biol Evol 27: 2576-2586.

Irimia M, Tena JJ, Alexis MS, Fernandez-Minan A, Maeso I, Bogdanovic O, de la Calle-Mustienes E, Roy SW, Gomez-Skarmeta JL, Fraser HB. 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 22:2356-2367.

Ishii N, Kaneko Y. 2008. *Petaurista leucogenys*. The IUCN Red List of Threatened Species 2008: e.T16720A6314320.

Janes DE, et al. 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. Genome Biol Evol. 3:102?113.

Jones E et al. 2011. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org.

Karro JE, Peifer M, Hardison RC, Kollmann M, von Grunberg HH. 2008. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. Mol Biol Evol. 25:362-374

Katzman S, et al. 2007. Human genome ultraconserved elements are ultraselected. Science 317:915.

Kawamichi T. 1997. Seasonal changes in the diet of Japanese giant flying squirrels in relation to reproduction. J. Mammal. 78:204?212.

Kay EH, Hoekstra HE. 2008. "Rodents". Current Biology 18: R406?R410

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223-7.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182-187.

Kimura M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. J Genet 57:21-34.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. Science 188: 107?116.

King MC, Wilson AC. 1975. Evolution at two levels in human and chimpazees. Science 188:107-116.

Kumar S, Stecher G, Peterson D, and Tamura K (2012) MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis. Bioinformatics 28:2685-2686.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Larkin MA et al. 2007. “ClustalW and ClustalX version2”. Bioinformatics 23:2947-2948.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.

Lee AP, Kerk1 SY, Tan1 YY, Brenner S, Venkatesh B. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol. 28:1205-1215.

Lettice LA et al. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12:1725-1735.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12: 1725-1735.

Levy S, Hannenhalli S, Workman C. 2001. Enrichment of regulatory signals in conserved noncoding genomic sequence. Bioinformatics 17:871-877.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform. 12:323.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493-6.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265-72.

Li W, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation?time effect hypothesis. Mol Phylogenet Evol. 5:182-187.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol. 5:182-7.

Li WH, Wu CI. 1987. Rates of nucleotide substitution are evidently higher in rodents than in man. Mol Biol Evol. 4:74-82.

Liao BY, Weng MP, Zhang J. 2010. Contrasting genetic paths to morphological and physiological evolution. Proc Natl Acad Sci U S A. 107:7353-8.

Liu GE, Matukumalli LK, Sonstegard TS, Shade LL, Van Tassell CP. 2006. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. BMC Genomics 7:140.

Liu X, Wei F, Li M, Jiang X, Feng Z, Hu J. 2004. Molecular phylogeny and taxonomy of wood mice (genus *Apodemus* Kaup, 1829) based on complete mtDNA cytochrome b sequences, with emphasis on Chinese species. Mol Phylogenet Evol. 33:1-15.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288, 136?140.

Matsunami M, Saitou N. 2013. Vertebrate paralogous conserved noncoding sequences may be related to gene expression in brain. Genome Biol Evol. 5:140-150.

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Greg Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 5(12).

McGaughey DM, Stine ZE, Huynh JL, Vinton RM, McCallion AS. 2009. Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend. BMC Genomics 10:8.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 20:1335-1343.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338:1593-9.

Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny or vertebrates. Ann. Rev. Ecol. Evol. Syst. 34:311-338

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci U S A. 77(12):7328-32.

Morris JR, Geyer PK, Wu C. 1999. Core promoter elements can regulate transcription on a separate chromosome in trans. Genes Dev. 13:253-258.

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-62.

Musser GG, Carleton MD. 1993. Family Muridae. In: WilsonDE, ReederDM, eds. Mammal species of the world. Washington D.C.: Smithsonian Institution Press, 510?755.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635-640.

Nei M. 2013. Mutation-driven evolution. New York: Oxford University Press

Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. Megabase deletions of gene deserts result in viable mice. Nature 431:988-93.

Nowak RE., Walker's Mammals of the World. The Johns Hopkins Press (1999), ISBN 978-0-8018-5789-8.

Ohta T. 1993. An examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci U S A 90:10676-10680.

Okano T, Onuma M, Ishiniwa H, Azuma N, Tamaoki M, Nakajima N, Shindo J, Yokohata Y. 2015. Classification of the spermatogenic cycle, seasonal changes of seminiferous tubule morphology and estimation of the breeding season of the large Japanese field mouse (*Apodemus speciosus*) in Toyama and Aomori prefectures, Japan. J Vet Med Sci. 2015 77:799-807.

Ong CT, Corces VG. 2009. Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. J. Biol 8:73.

Osada N, Hettiarachchi N, Babarinde IA, Saitou N, Blancher A. 2015. Whole-genome sequencing of six Mauritian cynomolgus macaques (*Macaca fascicularis*) reveals a genome-wide pattern of polymorphisms under extreme population bottleneck. Genome Biol Evol 7: 821-830.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res. 15:137-145.

Ovcharenko I, et al. 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res. 15:137?145.

Palmer CA, Watts RA, Gregg RG, McCall MA, Houck LD, Highton R, Arnold SJ. Lineage-specific differences in evolutionary mode in a salamander courtship pheromone. Mol Biol Evol. 22:2243-56.

Palumbi SR. 1994. Genetic Divergence, Reproductive Isolation, and Marine Speciation. Annu. Rev. Ecol. Evol. Syst. 25:547-572.

Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA. 2004. Proc Natl Acad Sci U S A. 101:6536-41.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Rao YS, Wang ZF, Chai XW, Wu GZ, Zhou M, Nie QH, Zhang XQ. 2010. Selection for the compactness of highly expressed genes in Gallus gallus. Biol Direct 5:35.

Rice WR. 1987. Speciation via habitat specialization: the evolution of reproductive isolation as a correlated character. Evol. Ecol. 1:301-314.

Rinderknecht A, Blanco RE. 2008. The largest fossil rodent. Proc Biol Sci. 275:923–928.

Sagai T, Amano T, Tamura M, Mizushina Y, Sumiyama K, Shiroishi T. 2009. A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. Development 136, 1665-1674.

Sagai T, Masuya H, Tamura M, Shimizu K, Yada Y, Wakana S, Gondo Y, Noda T, Shiroishi T. 2004. Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh). Mamm Genome 15:23-34.

Saitou N. 2014. Introduction to evolutionary genomics.

Saitou N, Nei M. 1987. The Neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406-425.

Sakamoto SH, Suzuki SN, Degawa Y, Koshimoto C, Suzuki RO. 2012. Seasonal Habitat Partitioning between Sympatric Terrestrial and Semi-Arboreal Japanese Wood Mice, *Apodemus* speciosus and A. argenteus in Spatially Heterogeneous Environment. Mammal Study 37:261-272.

Sandelin A, et al. 2004. Arrays of ultraconserved noncoding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99.

Sato JJ, Kawakami T, Tasaka Y, Tamenishi M, Yamaguchi Y. 2014. A Few Decades of Habitat Fragmentation has Reduced Population Genetic Diversity: A case study of landscape genetics of the large Japanese field mouse, *Apodemus* speciosus. Mammal Study 39:1-10.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43:e37.

Schmidt D, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor bindingdominic. Science 328:1036-1040.

Serizawa K, Suzuki H, Tsuchiya K. 2000. A phylogenetic view on species radiation in *Apodemus* inferred from variation of nuclear and mitochondrial genes. Biochem Genet. 38:27-40.

Shen Y, et al. 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488: 116?120.

Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034-1050.

Speakman JR. 2005. Body size, energy metabolism and lifespan. J Exp Biol. 9:1717-30.

Steppan SJ, Adkins RM, Spinks PQ, Hale C. 2005. Multigene phylogeny of the Old World mice, Murinae, reveals distinct geographic lineages and the declining utility of mitochondrial genes compared to nuclear genes. Mol Phylogenet Evol. 37:370-88.

Stern DL. 2000. Evolutionary developmental biology and the problem of variation. Evolution 54:1079-1091.

Sucena E, Delon I, Jones I, Payre F, Stern DL. 2003. Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. Nature 424:935-938.

Sumiyama K, Irvine SQ, Stock DW, Weiss KM, Kawasaki K, Shimizu N, Shashikant CS, Miller W, Ruddle FH. 2002. Genomic structure and functional control of the Dlx3-7 bigene cluster. Proc Natl Acad Sci USA 99:780-785.

Sumiyama K, Ruddle FH. 2003. Regulation of Dlx3 gene expression in visceral arches by evolutionarily conserved enhancer elements. Proc Natl Acad Sci U S A. 100:4030-4034.

Suwa G, Kono RT, Katoh S, Asfaw B, Beyene Y. 2007. A new species of great ape from the late Miocene epoch in Ethiopia. Nature 448, 921-924.

Suzuki H, Filippucci MG, Chelomina GN, Sato JJ, Serizawa K, Nevo E. 2008. A biogeographic view of *Apodemus* in Asia and Europe inferred from nuclear and mitochondrial gene sequences. Biochem Genet. 46:329-46.

Suzuki H, Yasuda SP, Sakaizumi M, Wakana S, Motokawa M, Tsuchiya K. 2004. Differential geographic patterns of mitochondrial DNA variation in two sympatric species of Japanese wood mice, *Apodemus speciosus* and *A. argenteus*. Genes Genet Syst. 79:165-76.

Takahashi M, Saitou N. 2012. Identification and Characterization of Lineage-Specific Highly Conserved Noncoding Sequences in Mammalian Genomes. Genome Biol Evol. 4:641-657.

Takahasi M, Saitou N. 2012. Identification and characterization of lineage-specific highly conserved noncoding sequences in mammalian genomes. Genome Biol Evol. 4:641-657.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30: 2725?2729.

Tamura K, et al. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol. 28:2731-2739.

Tamuraa K, Battistuzzib FU, Billing-Rossb P, Murillob O, Filipskib A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proc Natl Acad Sci U S A. 109:19333-19338.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Thomas JA, Welch JJ, Lanfear R, Bromham L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. Mol Biol Evol. 27:1173-80.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 13:2129-2141.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-80.

Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. 2005. Defining a genomic radius for long-range enhancer action: duplicated conserved noncoding elements hold the key. Trends

Genet. 22:5-10.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved noncoding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. 8:R15.

Vawter AT, Rosenblatt R, Gorman GC. 1980. Genetic divergence among fishes of the eastern pacific and the Caribbean: support for the molecular clock. Evolution 34:705-711.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327-335.

Vilella AJ, et al. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327-35.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854-858.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser-a database of tissue-specific human enhancers. Nucleic Acids Res 35:D88-92.

Visel A, et al. 2009. A high-resolution enhancer atlas of the developing telencephalon. Cell 152:895?908.

Wilson DE, Reeder DM. 2005. Mammal species of the world: a taxonomic and geographic reference. 3rd edition. Baltimore, MD, Johns Hopkins University Press.

Wilusz JE, Sunwoo H, Spector DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 23(13):1494-504.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature 13:59-69.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3, e7.

Woolfe A, et al. 2004. Highly conserved noncoding sequences are associated with vertebrate development. PLoS Biol. 3:e7.

Wray GA. 2003. Transcriptional regulation and the evolution of development. Int J Dev Biol. 47:675-84.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. 8:206-216.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A. 82:1741-5.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

Yi SV. 2013. Morris Goodman's hominoid rate slowdown: the importance of being neutral. Mol Phylogenet Evol. 66:569-74.

Zhang F, Broughton RE. 2015. Heterogeneous natural selection on oxidative phosphorylation genes among fishes with extreme high and low aerobic performance. BMC Evol Biol 15:173.

Zhang F, Peterson T. 2006 Gene Conversion Between Direct Noncoding Repeats Promotes Genetic and Phenotypic Diversity at a Regulatory Locus of Zea mays (L.). Genetics 174:2 753-762.

Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 38, 1341-1347.

Zhu T, Dos Reis M, Yang Z. 2015. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. Syst Biol. 64:267-80.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. P. 97-166.

Appendix 1

Figure A1.1: Japanese giant flying squirrel (*Petaurista leucogenys*). The genome sequencing of Japanese giant flying squirrel (musasabi) is presented in Chapter Two. [Image Source: www.soapinterop.org]

Figure A1.2: Capybara (*Hydrochoerus hydrochaeris*). The genome sequences of capybara are reported in Chapter Six. [Image Source: animal.sandiegozoo.org]

Figure A1.3: Large Japanese wood mouse (*Apodemus speciosus*). Transcriptome analyses of large Japanese wood mouse (akanezumi) are presented in Chapter Five. [Image Source: Wikipedia]

Appendix 2

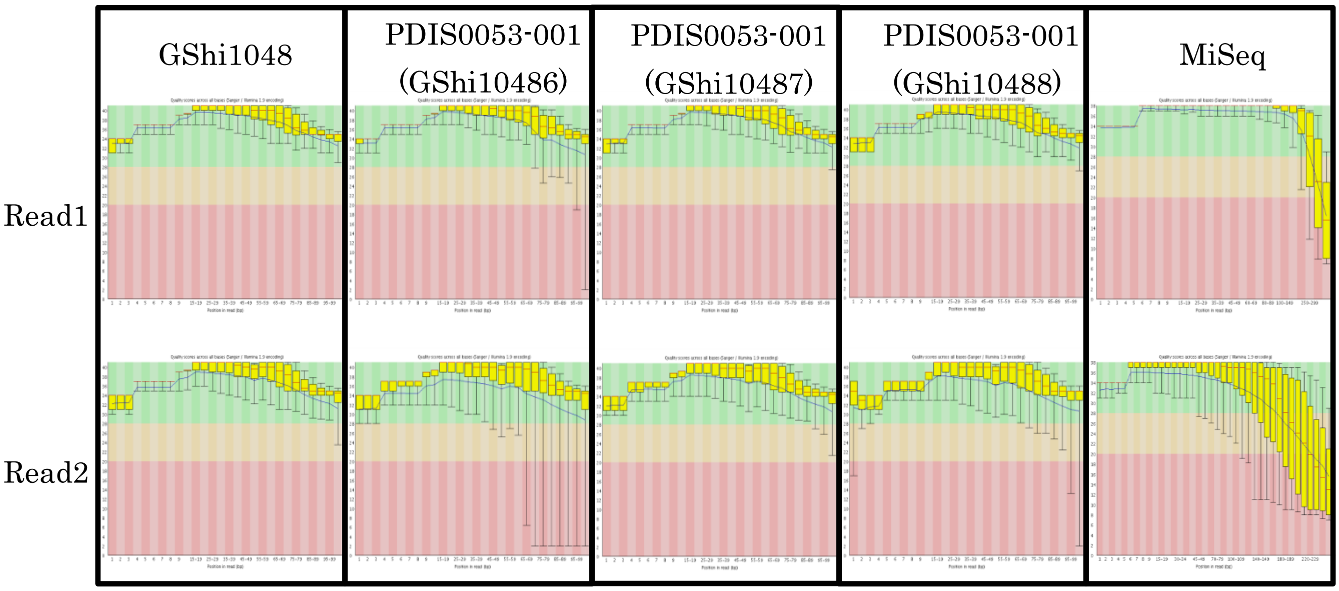


Figure A2.1: The read quality distribution of musasabi

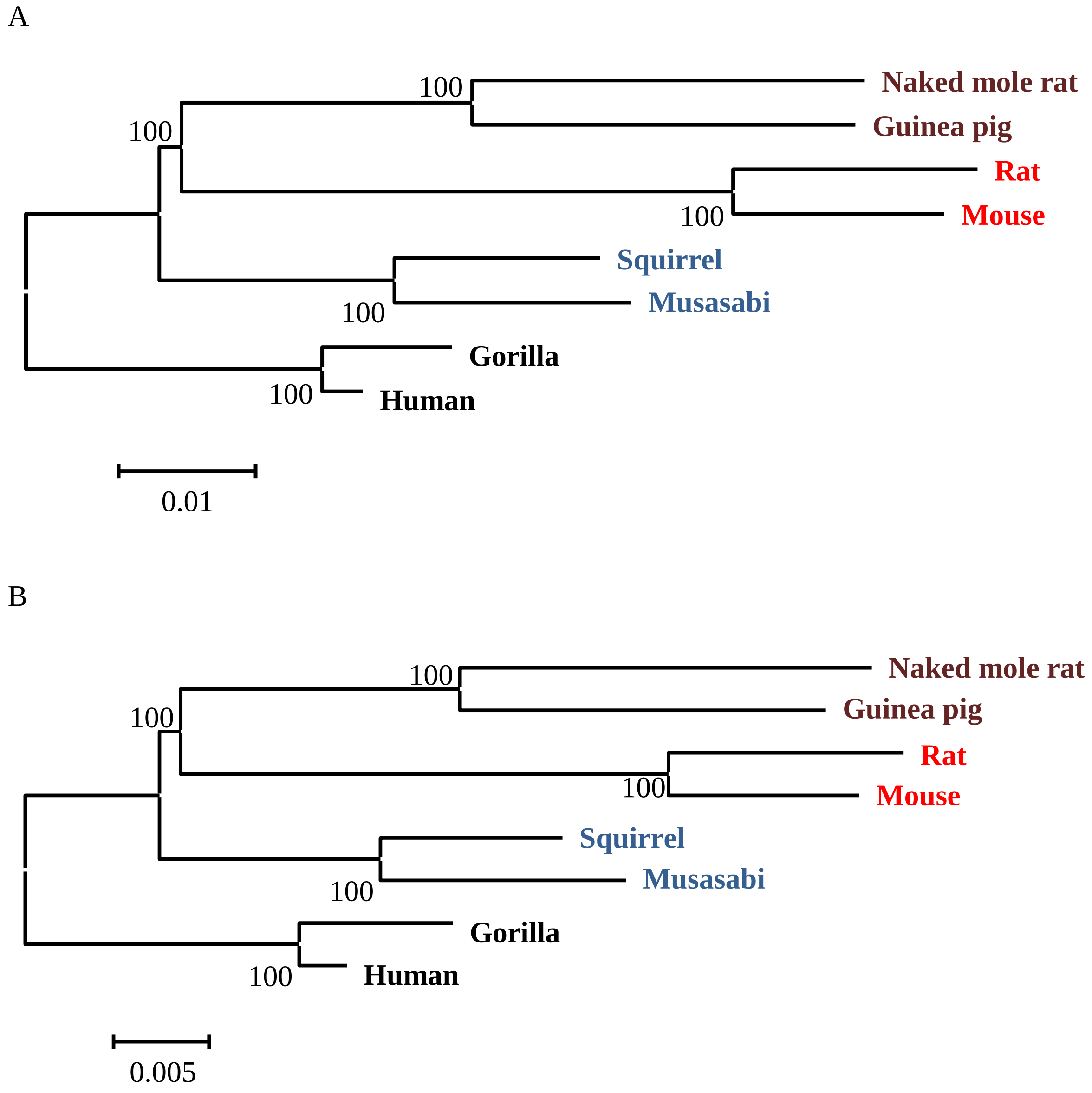


Figure A2.2: Phylogenetic tree of three rodent lineages constructed from first (A) and second (B) codon positions. About 2.9 million sites were used to construct the tree.

Table A2.1: Retention of ancestral CNSs in rodents

	Common	**Union**
Total	28,952*	46,816**
Musasabi	28,041	42,760
Squirrel	27,441	41,483
Naked mole rat	27,330	40,824
Guinea pig	26,530	38,947
Mouse	24,975	35,575
Rat	24,517	34,760

* The total number of CNSs conserved in chicken, human, cow and dog

** The total number of CNSs conserved in chicken and at least one of human, cow and dog

Table A2.2: The distance matrix of third codon positions.

	Squirrel	*NMR	Gorilla	Rat	Guinea pig	Musasabi	Human	Mouse
Squirrel		0.000409	0.00037	0.000508	0.000437	0.000222	0.000363	0.000488
*NMR	0.319969		0.000429	0.000558	0.000305	0.000404	0.000422	0.000556
Gorilla	0.285867	0.327201		0.000521	0.000449	0.000366	8.03E-05	0.000513
Rat	0.416873	0.453138	0.4313		0.000559	0.0005	0.000514	0.000248
Guinea pig	0.3473	0.208877	0.358423	0.474488		0.000425	0.000435	0.000567
Musasabi	0.126673	0.314194	0.278525	0.412227	0.341988		0.000357	0.000486
Human	0.276011	0.318004	0.024763	0.420595	0.34817	0.26952		0.000496
Mouse	0.408076	0.446627	0.423056	0.150285	0.4678	0.403627	0.412276	

*NMR = naked mole rat

The distances were estimated using GTR model (k = 5) of MEGA6. The lower left values are the distance estimates while the upper right values are the standard error estimates for the distances.

Appendix 3

Table A3.1. Number of CNSs retrieved in this study

	Total	**Intragenic**	**Intergenic**
Chicken	21,584	7,710	13,874
Human	21,191	10,120	11,071
Mouse	21,026	9,482	11,544
Dog	21,385	6,882	14,503
Cow	21,155	5,543	15,612

Table A3.2. Expression data of the tissues used retrieved from Necsulea et al. (2014)

Tissue	Availability in Human	Availability in Mouse
Brain	Yes	Yes
Cerebellum	Yes	Yes
CorticalPlate	Yes	Yes
EmbryonicBrain	Yes	Yes
EmbryonicLiver	Yes	No
EmbryonicKidney	No	Yes
EmbryonicStemCells	Yes	Yes
Heart	Yes	Yes
InnerSubventricularZone	Yes	No
Kidney	Yes	Yes
Liver	Yes	Yes
NeonatalBrain	No	Yes
OuterSubventricularZone	Yes	No
Ovary	Yes	Yes
Placenta	Yes	Yes
Testis	Yes	No
SubventricularZone	No	Yes
VentricularZone	Yes	Yes

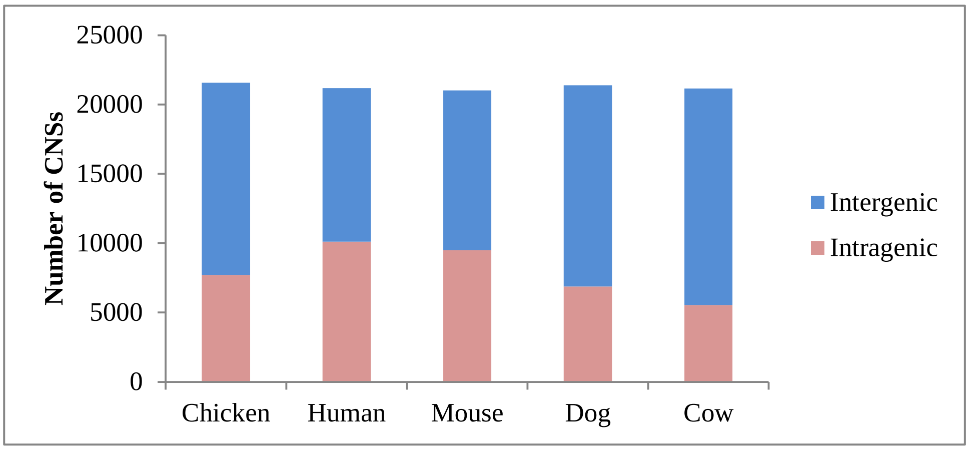
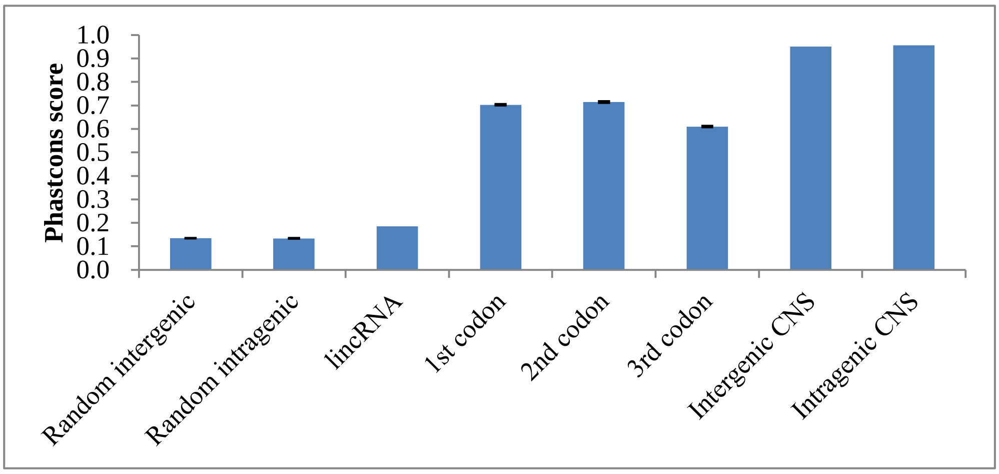


Figure A3.1: The overlap of the retrieved CNSs with protein-coding genes.

(A)



(B)

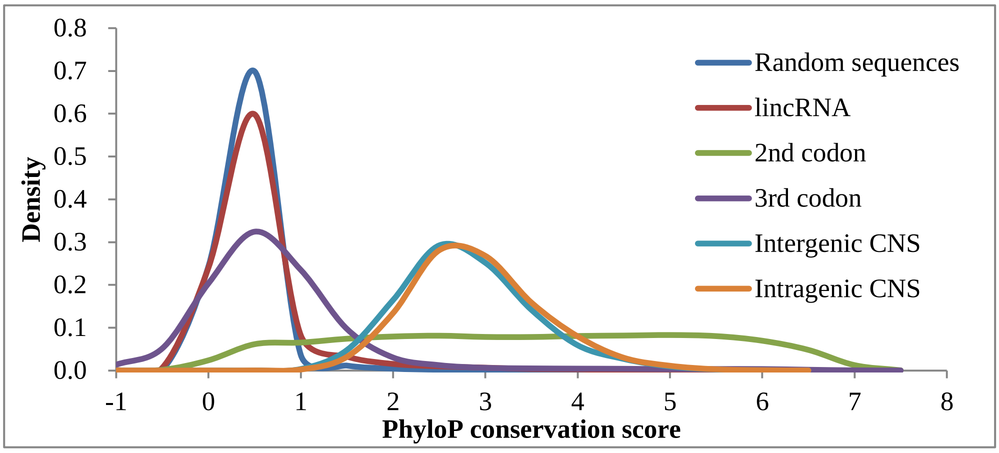
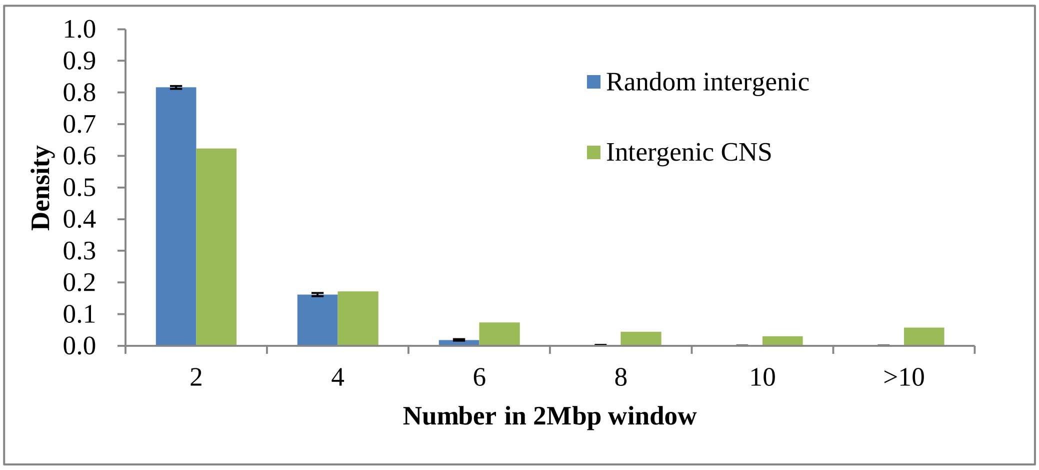


Figure A3.2: CNSs have stronger constraints in mouse, similar to Figure 3.1. (A) CNSs have the highest phastcons score. (B) CNSs have higher phyloP conservation score than random sequences.

(A)



(B)

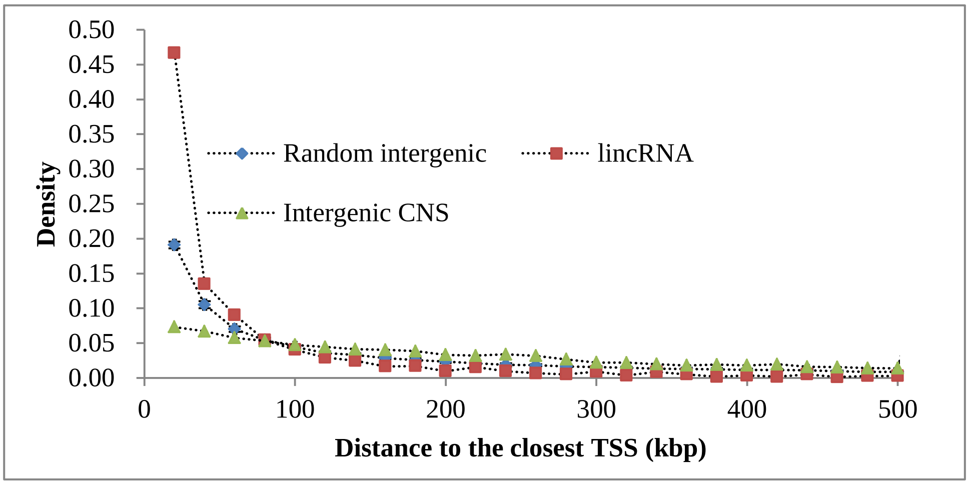


Figure A3.3: Nonrandom genomic location of CNSs in mouse, similar to fig. 2. (A) Compared to random sequences and lincRNAs, CNSs tend to exist in genomic clusters. (B) Intergenic CNSs tend to be located far away from protein-coding genes (Chi square p value <0.001).

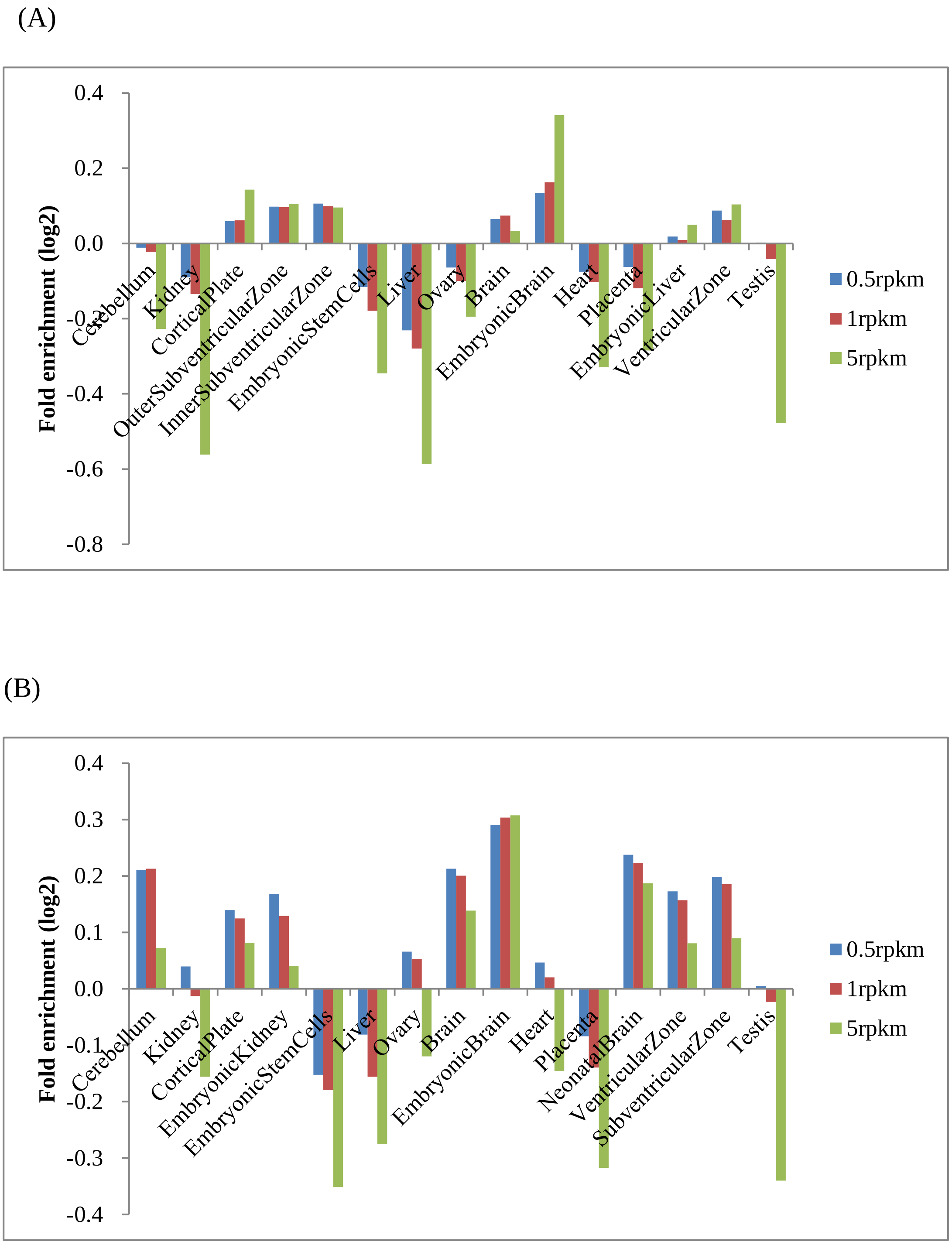
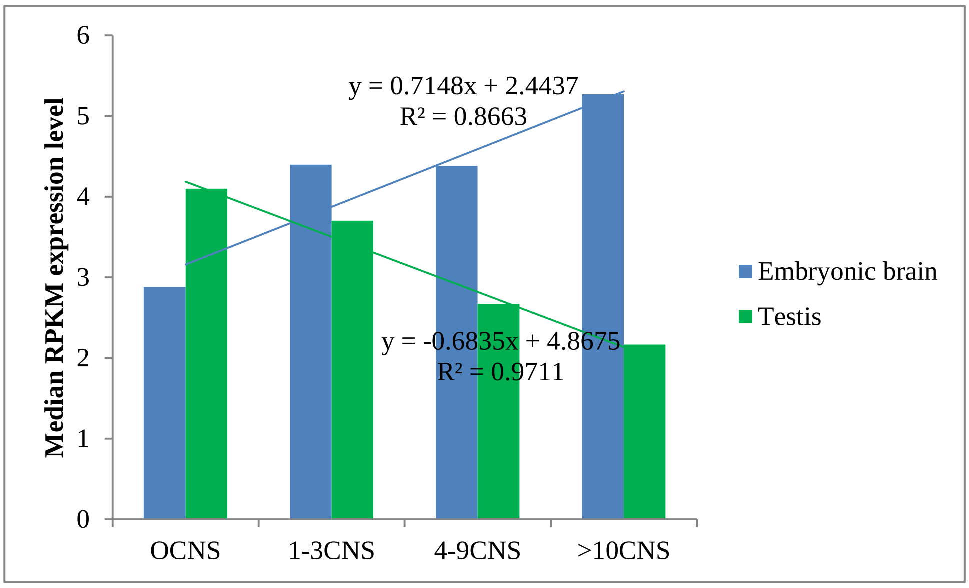


Figure A3.4: The enrichment of tissue expression at various expression levels in human (A) and in mouse (B).

(A)



(B)

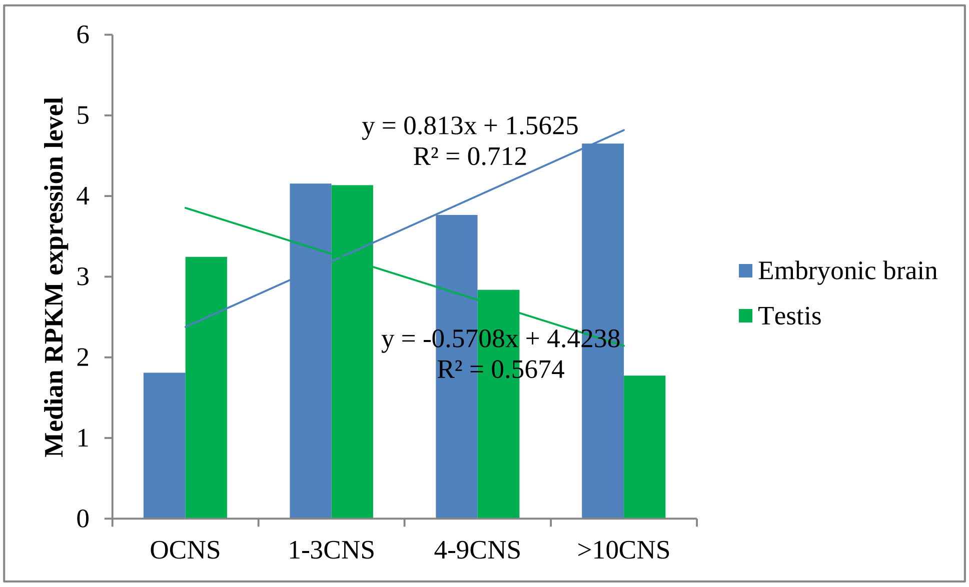


Figure A3.5: Genes with more CNSs tend to have higher embryonic brain expression and lower expression in testis. Similar results are found in human (A) and mouse (B).

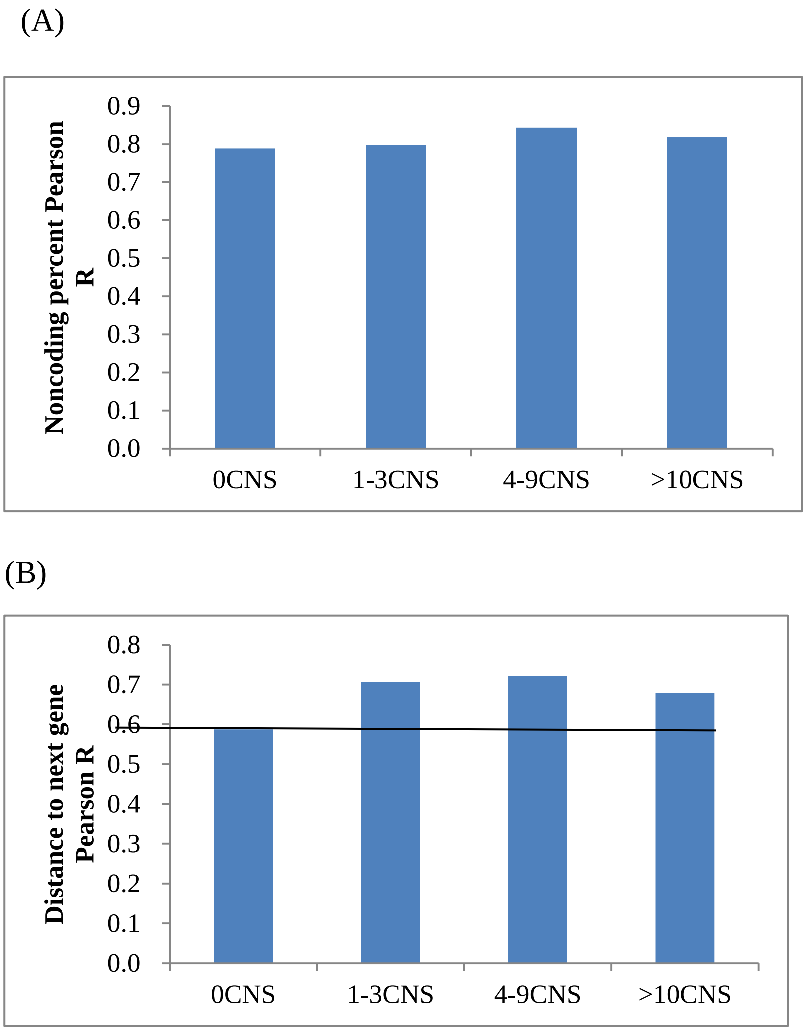


Figure A3.6: Conservation gene features. (A) Conservation of noncoding percent. (B) Conservation of distance to the next gene. Pearson's correlation coefficient was calculated for genes with one-to-one orthology between human and mouse. The black horizontal line shows the average Pearson's correlation coefficients for genes with no CNS.

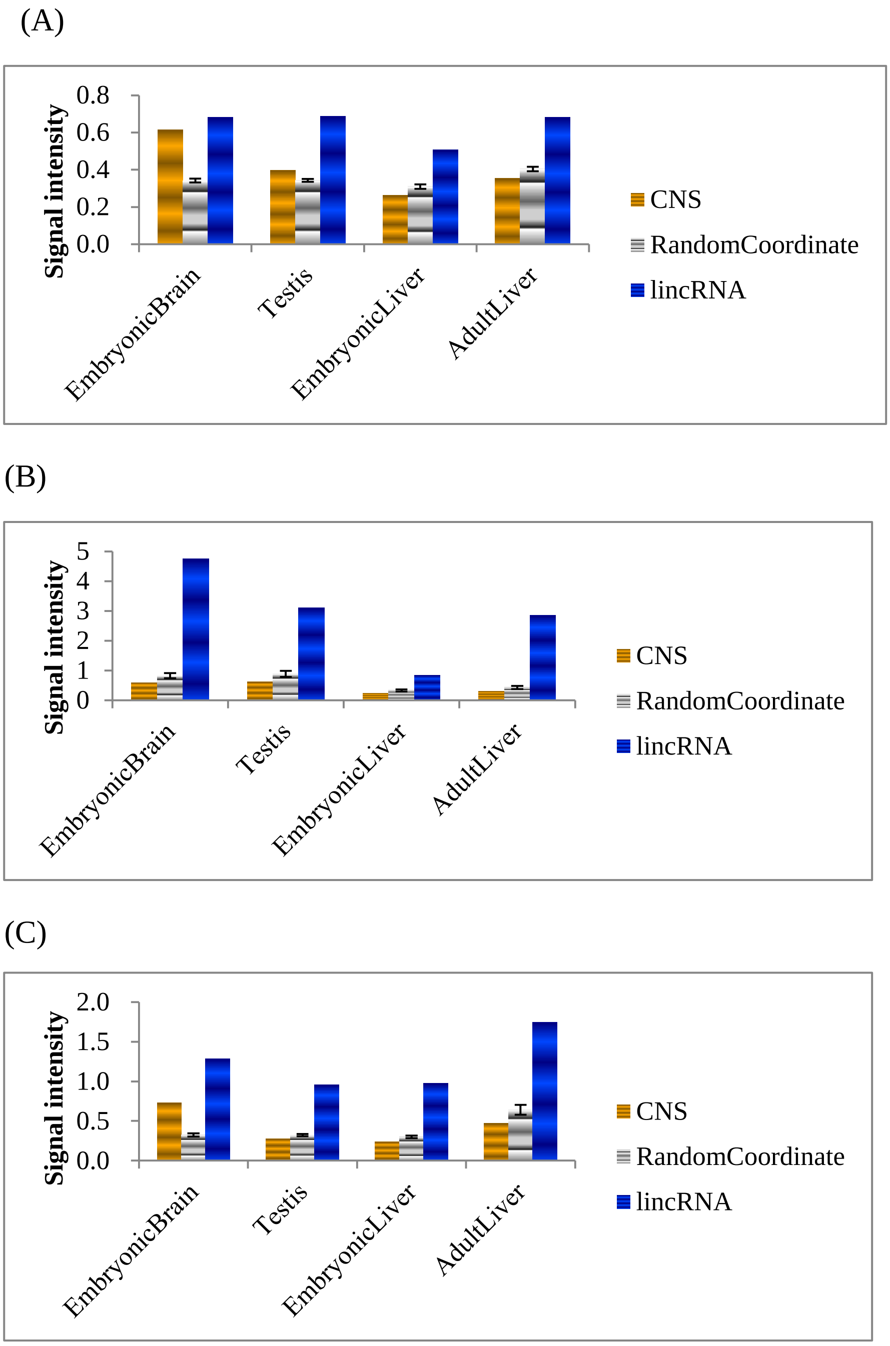


Figure A3.7. Histone modification marks differentiate CNSs from random sequences and lincRNA exons. H3k4me1 (A), H3k4me3 (B) and H3k27ac (C) for CNSs, random coordinates and lincRNA genes are shown.

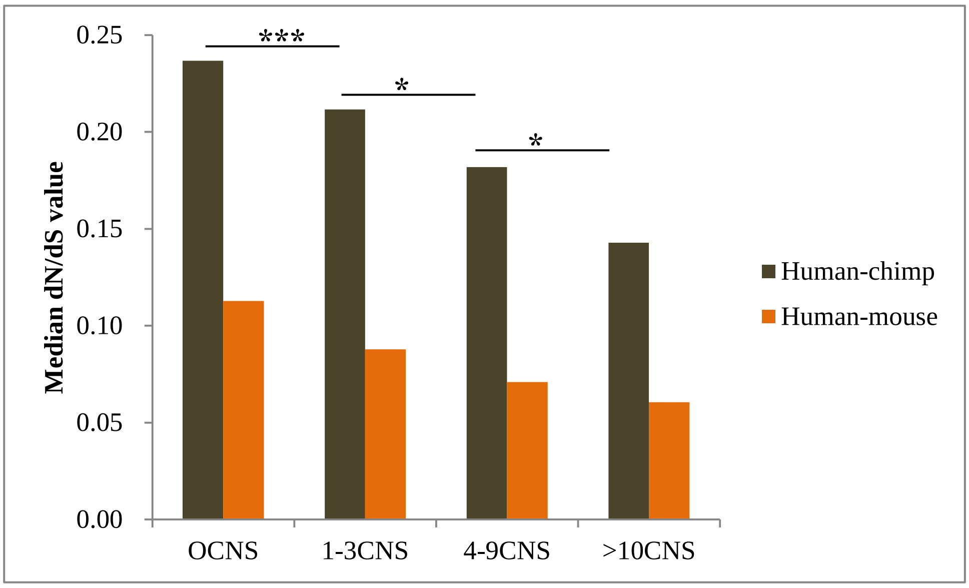


Figure A3.8: Genes with more CNSs tend to have stronger purifying selection. The data for dN and dS for human-chimpanzee and human-mouse were downloaded from *Ensembl biomart* (*** p value < 0.001; * p value <0.05; MannwhineyU test).

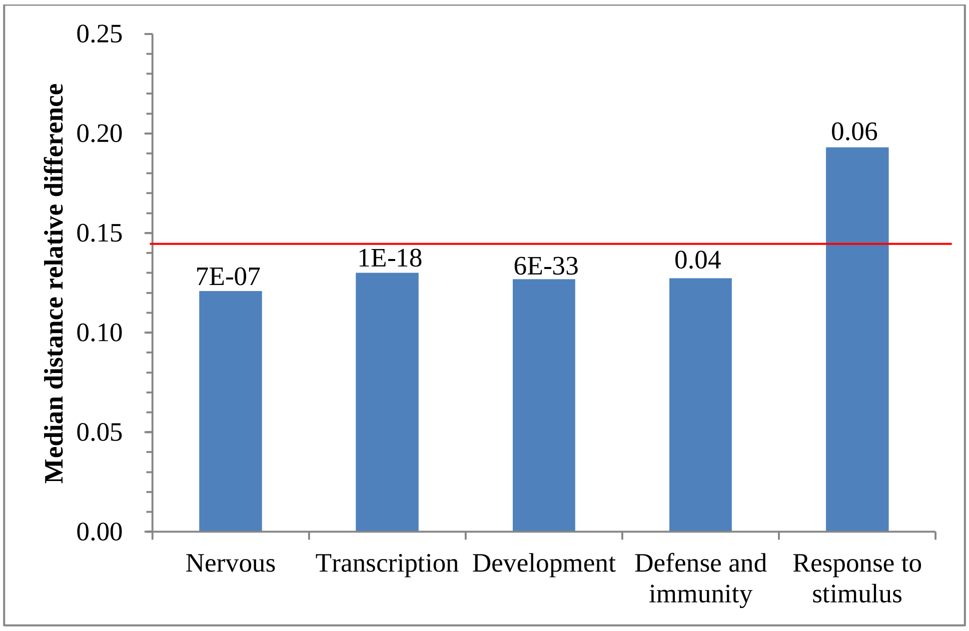


Figure A3.9: Gene ontology terms with more CNSs tend to have more significantly lower distance relative difference. The gene ontology terms shown to be overrepresented or underrepresented in CNSs in Figure 3.4 were tested whether they have significantly higher or lower distance relative difference. Each CNS-gene gene pair used for Figure 3.3 was used for gene ontology classification. The red horizontal line represents the overall median value for all CNS-gene pairs shown in Figure 3.2. The value above each bar represents the uncorrected MannwhineyU p value for the statistical test of difference between CNS-gene pairs with the GO term and those without the GO term.

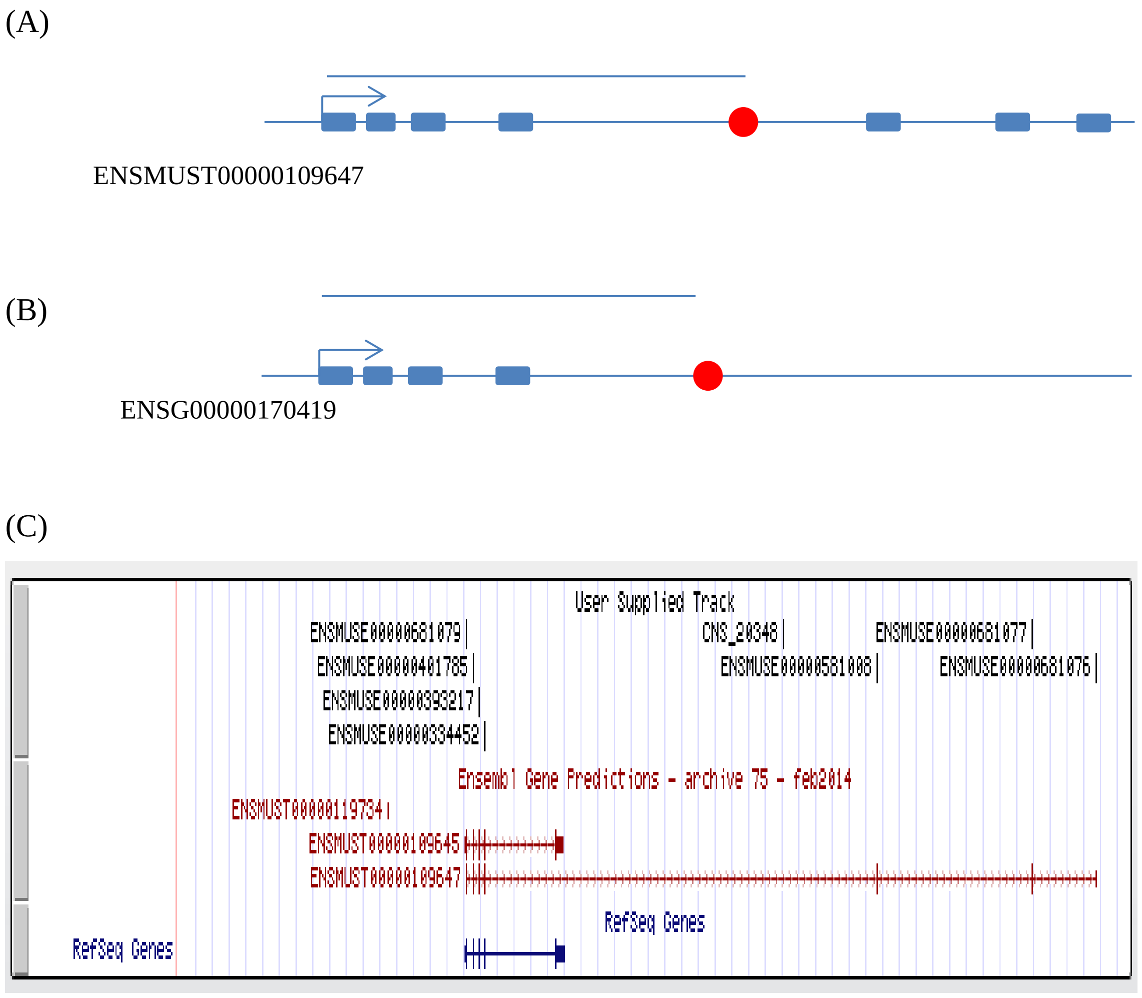


Figure A3.10: *STM2A* exon structures are different between human and mouse genomes. The exons are represented in blue rectangle while CNS is represented in red circle. (A) CNS_28348 is found in the intron of a transcript (ENSMUST00000109647) of *STM2A* gene in the mouse genome (upper panel). It is important to note that mouse has another shorter transcript, but Babarinde and Saitou (2013) used the longest transcript. (B) The CNS is located in the intergenic region flanking the human gene, ENSG00000170419. Although exon structures are different, the distance between the CNS and TSS are approximately the same (73kbp and 85kbp for human and mouse, respectively). (C) USCS browser view of the mouse gene and the exon structure.

Appendix 4

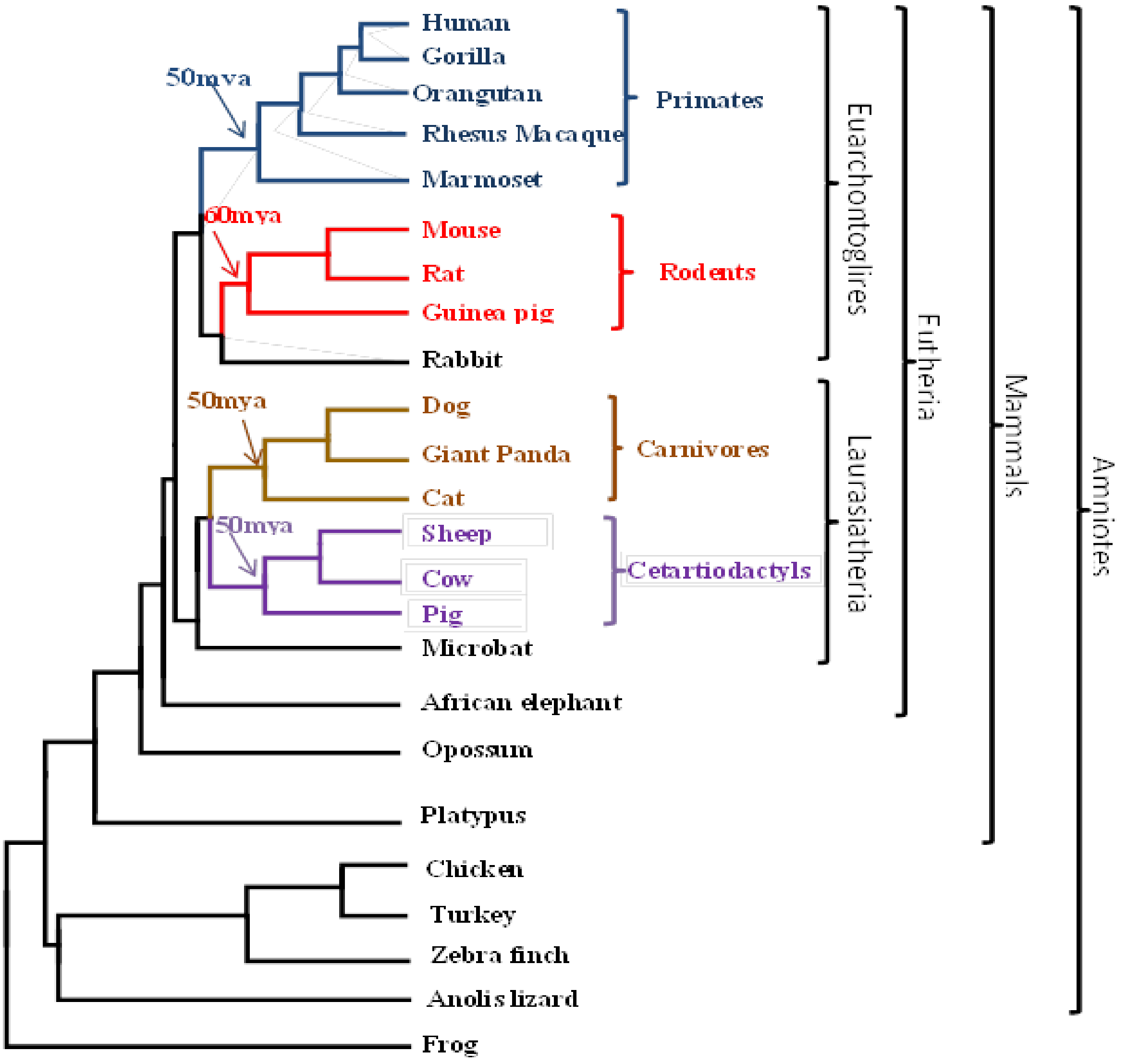


Figure A4.1: The phylogenetic and taxonomic classification of the species used in this study. The most diverged species in each of the four mammalian lineages used in this study is around 50~60mya. The tree is not scaled.

Table A4.1: Genome sequences used for the analyses

Species	Database	Build	Order
Human	*Ensembl*	Homo_sapiens.GRCh37.66.dna_rm.toplevel.fa	Primate
Gorilla	*Ensembl*	Gorilla_gorilla.gorGor3.1.66.dna_rm.toplevel.fa	Primate
Orangutan	*Ensembl*	Pongo_abelii.PPYG2.66.dna_rm.toplevel.fa	Primate
Rhesus macaque	*Ensembl*	Macaca_mulatta.MMUL_1.66.dna_rm.toplevel.fa	Primate
Marmoset	*Ensembl*	Callithrix_jacchus.C_jacchus3.2.1.66.dna_rm.toplevel.fa	Primate
Mouse	*Ensembl*	Mus_musculus.NCBIM37.66.dna_rm.toplevel.fa	Rodentia
Rat	*Ensembl*	Rattus_norvegicus.RGSC3.4.66.dna_rm.toplevel.fa	Rodentia
Guinea pig	*Ensembl*	Cavia_porcellus.cavPor3.66.dna_rm.toplevel.fa	Rodentia
Rabbit	*Ensembl*	Oryctolagus_cuniculus.oryCun2.66.dna_rm.toplevel.fa.	Lagomorpha
Dog	*Ensembl*	Canis_familiaris.CanFam3.1.69.dna_rm.toplevel.fa	Carnivora
Panda	*Ensembl*	Ailuropoda_melanoleuca.ailMel1.70.dna_rm.toplevel.fa	Carnivora
Cat	*Ensembl*	Felis_catus.Felis_catus_6.2.70.dna_rm.toplevel.fa	Carnivora
Cow	*Ensembl*	Bos_taurus.UMD3.1.66.dna_rm.toplevel.fa.	Cetartiodactyla
Sheep	UCSC genome	oviAri1.fa.masked	Cetartiodactyla
Pig	*Ensembl*	Sus_scrofa.Sscrofa9.66.dna_rm.toplevel.fa	Cetartiodactyla
Microbat	*Ensembl*	Myotis_lucifugus.Myoluc2.0.66.dna_rm.toplevel.fa	Chiroptera
African elephant	*Ensembl*	Loxodonta_africana.loxAfr3.66.dna_rm.toplevel.fa	Proboscidea
Opossum	*Ensembl*	Monodelphis_domestica.BROADO5.66.dna_rm.toplevel.fa	Didelphimorpha
Platypus	*Ensembl*	Ornithorhynchus_anatinus.OANA5.66.dna_rm.toplevel.fa	Monotremata
Chicken	*Ensembl*	Gallus_gallus.WASHUC2.66.dna_rm.toplevel.fa	Galliformes
Turkey	*Ensembl*	Meleagris_gallopavo.UMD2.66.dna_rm.toplevel.fa	Galliformes
Zebra finch	*Ensembl*	Taeniopygia_guttata.taeGut3.2.4.66.dna_rm.toplevel.fa	Passeriformes
Anolis lizard	*Ensembl*	Anolis_carolinensis.AnoCar2.0.66.dna_rm.toplevel.fa	Squamata
Frog	*Ensembl*	Xenopus_tropicalis.JGI_4.2.66.dna_rm.toplevel.fa	Anura

Table A4.2: Setting of divergence threshold for each lineage

	Synonymous (S)	Nonsynonymous (N)	Genomic Noncoding	Coding (C)	Mean Divergence Proportion (P)
Primates	0.2224 (1.81)	0.0441 (0.20)	0.1218	0.0575 (0.03)	0.0603
Carnivores	0.3868 (2.06)	0.0572 (0.15)	0.1861	0.0817 (0.04)	0.0633
Cetartiodactyls	0.5201 (2.87)	0.0687 (0.24)	0.2143	0.0979 (0.05)	0.0561
Rodents	0.9601 (2.99)	0.1101 (0.25)	0.239	0.1656 (0.06)	0.0578

The mean values of the divergences are given while the values in parentheses are the standard deviations. Coding divergence, which is significantly lower than synonymous and genomic noncoding divergences, but higher than nonsynonymous divergence, was used as threshold. Mean divergence proportion (P) is given by (C-N)/S. The e-value threshold in homology search reduces the standard deviations of C because poorly conserved alignments are not included.

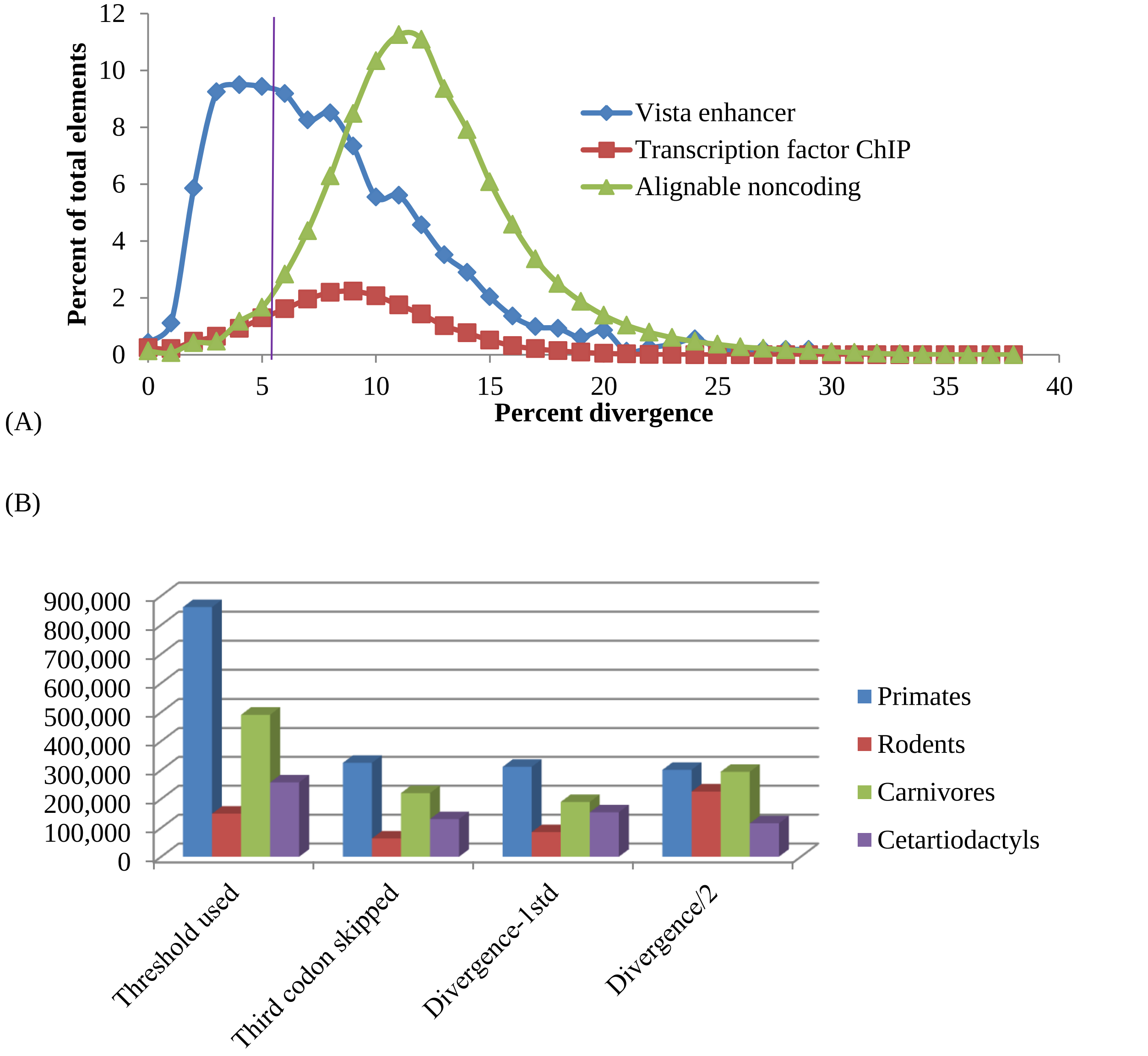


Figure A4.2: Assessing the suitability of the thresholds. (A) Human experimentally verified Vista enhancer elements and transcription factor ChIP sequences downloaded from UCSC table were used. The threshold used (shown in purple line) is shown on the distribution of alignable human and marmoset noncoding sequences. (B) Numbers of CNSs decrease with more stringent thresholds, but the pattern remains essentially the same. For third codon skipped, we set the thresholds after removing bases of third codon positions, which are mostly synonymous substitutions (3.8, 9.5, 4.8 and 6.1% divergence for primates, rodents, carnivore and cetartiodactyls, respectively). I further checked the numbers of CNSs conserved between the reference genomes and the most diverged species of each lineage using threshold divergence-1std and half or threshold divergence.

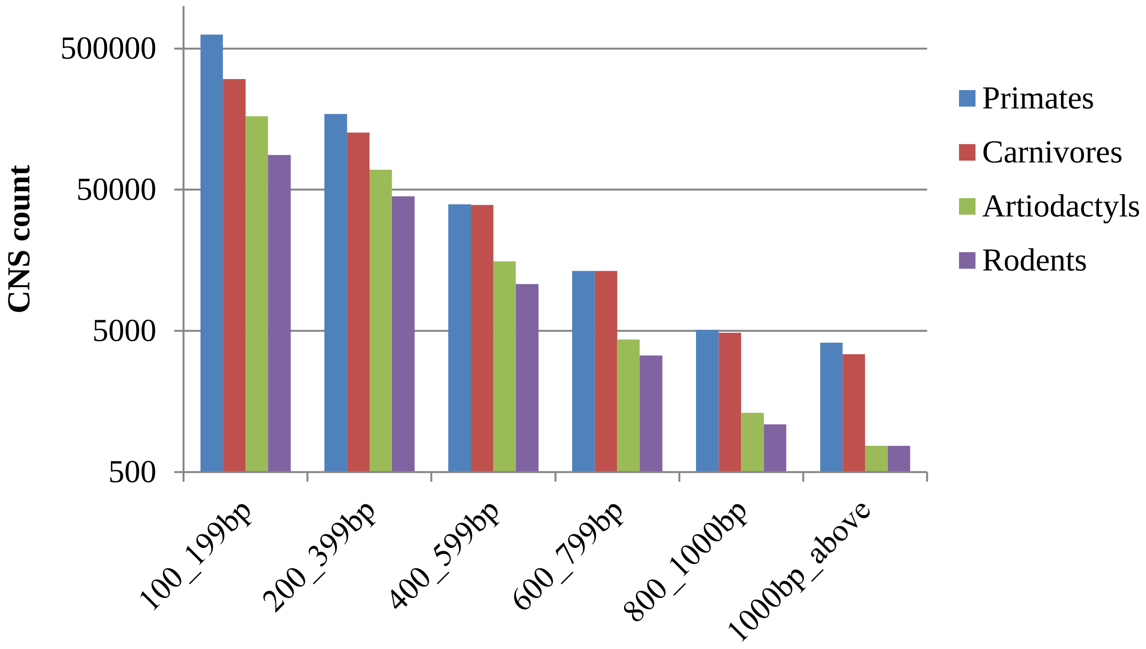


Figure A4.3: The length distribution of lineage common CNSs. In every length category, primate common CNSs are the most abundant while rodent common CNSs are the least abundant.

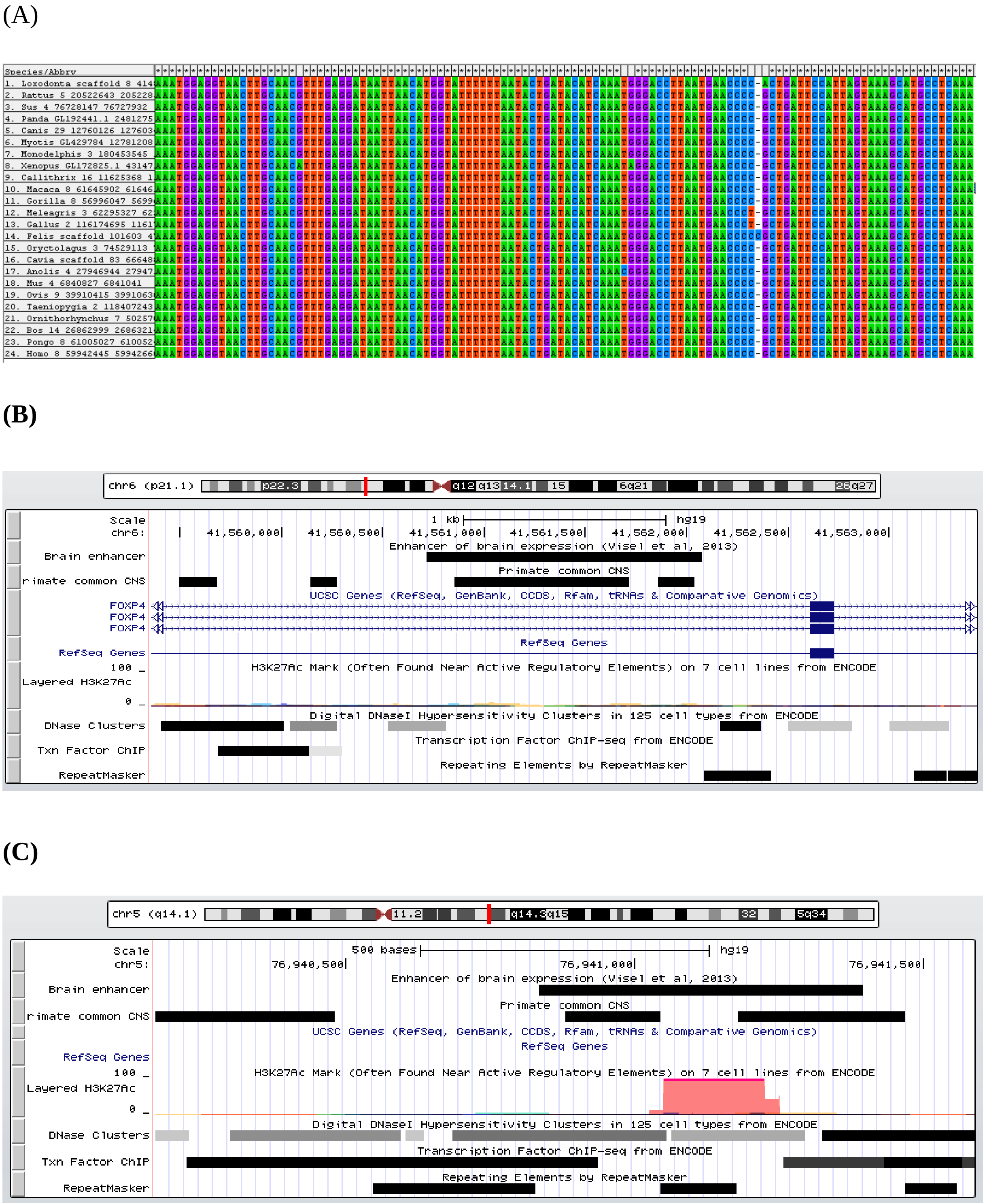

(A)
Species/Abbrv
1. Loxodonta scaffold 8 414
2. Rattus 5 20522643 205228
3. Sus 4 76728147 76727932
4. Panda GL192441.1 2481275
5. Canis 29 12760126 127603
6. Myotis GL429784 12781208
7. Monodelphis 3 180453545
8. Xenopus GL172825.1 43147
9. Callithrix 16 11625368 1
10. Macaca 8 61645902 61646
11. Gorilla 8 56996047 5699
12. Meleagris 3 62295327 62
13. Gallus 2 116174695 1161
14. Felis scaffold 101603 4
15. Oryctolagus 3 74529113
16. Cavia scaffold 83 66648
17. Anolis 4 27946944 27947
18. Mus 4 6840827 6841041
19. Ovis 9 39910415 3991063
20. Taeniopygia 2 118407243
21. Ornithorhynchus 7 50257
22. Bos 14 26862999 2686321
23. Pongo 8 61005027 610052
24. Homo 8 59942445 5994266
(B)
chr6 (p21.1)
p22.3
q12
q13
14.1
15
6q21
26
q27
Scale
chr6:
1 kb
hg19
41,560,000|
41,560,500|
41,561,000|
41,561,500|
41,562,000|
41,562,500|
41,563,000|
Enhancer of brain expression (Visel et al, 2013)
Brain enhancer
Primate common CNS
rimate common CNS
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
FOXP4
FOXP4
FOXP4
RefSeq Genes
RefSeq Genes
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Repeating Elements by RepeatMasker
RepeatMasker
(C)
chr5 (q14.1)
11.2
q14.3
q15
32
5q34
Scale
chr5:
500 bases
hg19
76,940,500|
76,941,000|
76,941,500|
Enhancer of brain expression (Visel et al, 2013)
Brain enhancer
Primate common CNS
rimate common CNS
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
RefSeq Genes
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Repeating Elements by RepeatMasker
RepeatMasker

Figure A4.4: Examples of identified CNSs. (A) An example of the multiple alignement of CNS showing high sequence conservation. (B) Four syntenic primate common CNSs with two overlapping brain expression, two overlapping DNase clusters and one overlapping transcription factor. (C) Three identified primate common CNSs located in intergenic regions.

Table A4.3: Number of ancestral CNSs

Query genome	Subject genome	Number of CNSs
Chicken	Human	45,532
	Mouse	33,865
	Dog	45,554
	Cow	43,858
	Microbat	40,079
	African elephant	43,848
	Opossum	41,027
	Platypus	41,038
African elephant	Human	310,912
	Mouse	155,828
	Dog	276,661
	Cow	142,485

Table A4.4: Maximum Composite Likelihood Estimate of the Pattern of Nucleotide Substitution

	A	T	C	G
A	-	*4.57(4.54)*	*2.76(3.43)*	**13.38(14.65)**
T	*4.55(4.53)*	-	**13.32(14.69)**	*2.78(3.44)*
C	*4.55(4.53)*	**22.09(19.46)**	-	*2.78(3.44)*
G	**21.9(19.33)**	*4.57(4.54)*	*2.76(3.43)*	-

Each entry shows the probability of substitution (r) from one base (row) to another base (column) for tetrapod common CNSs and primate unique CNSs (in parentheses). For simplicity, the sum of r values is made equal to 100. Rates of different transitional substitutions are shown in **bold** and those of transversional substitutions are shown in *italics*. The analysis involved concatenated CNSs of five primates species used. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA5. GC→AT substitutions (53.11% and 47.86% for tetrapod common and primate unique CNSs, respectively) are higher than AT→GC substitutions (32.24% and 36.21% for tetrapod common and primate unique CNSs, respectively).

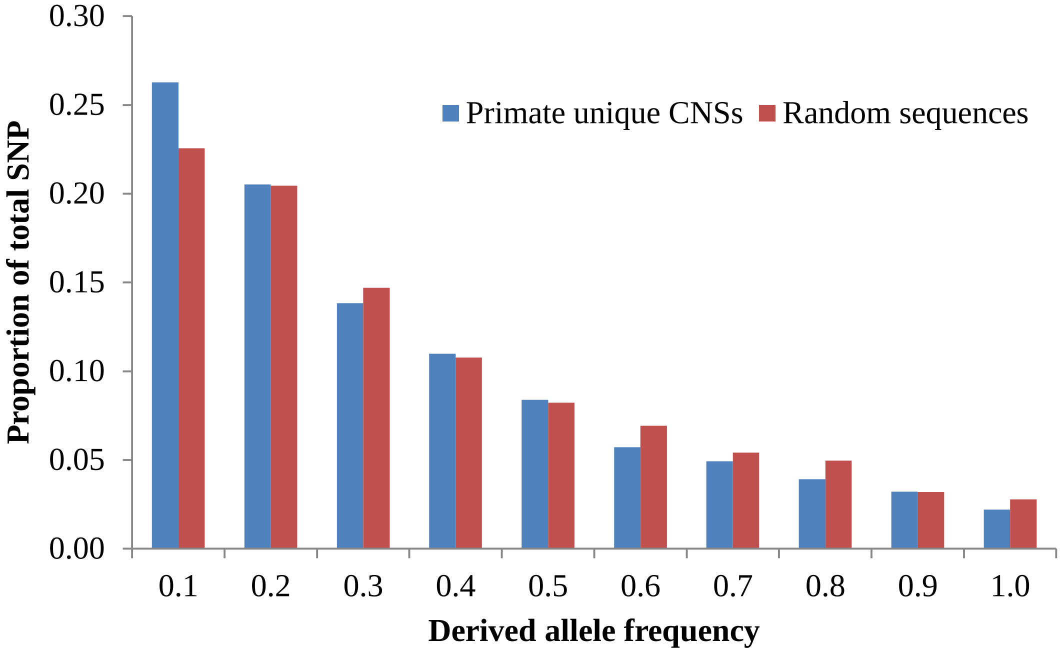
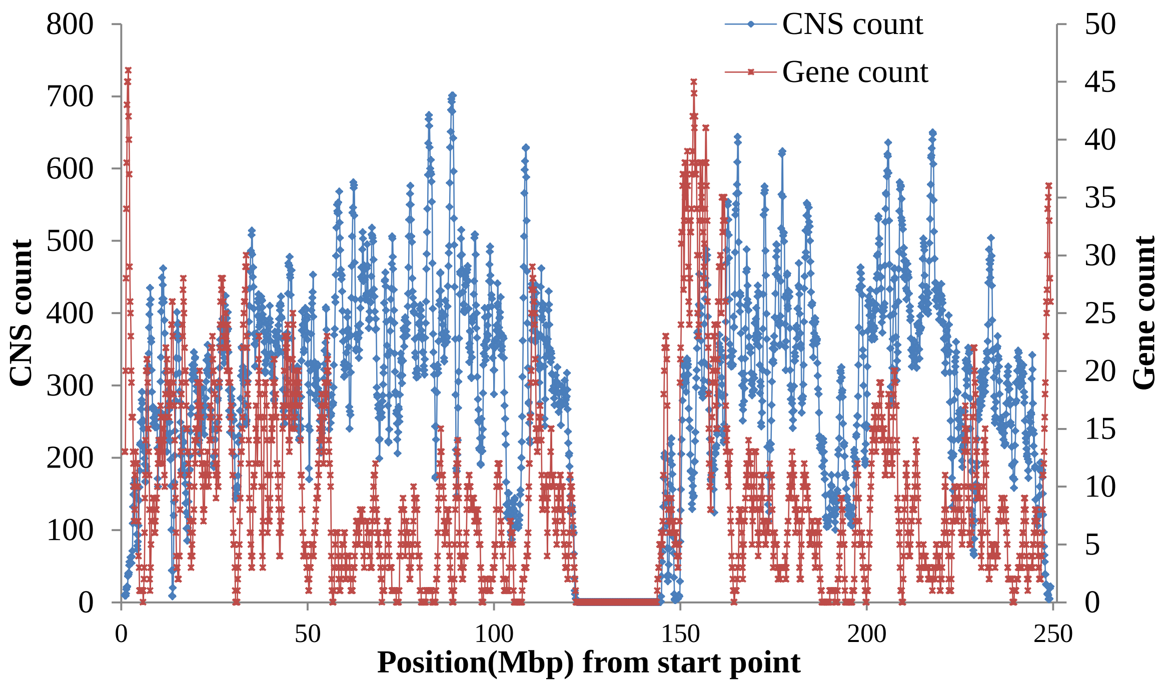


Figure A4.5: Derived allele frequency analysis for primate unique CNSs using Yoruba population of Hapmap Project III. Values on the horizontal axis is the upper limit of frequency. Compared to the random sequences, most of the SNPs in primate unique CNSs are of lower frequency. At derived allele frequency of at most 0.1, CNSs have significantly higher prortion of SNPs compared to the random sequences. This suggests that SNPs on CNSs do not spread in the population.

(B)



(A)

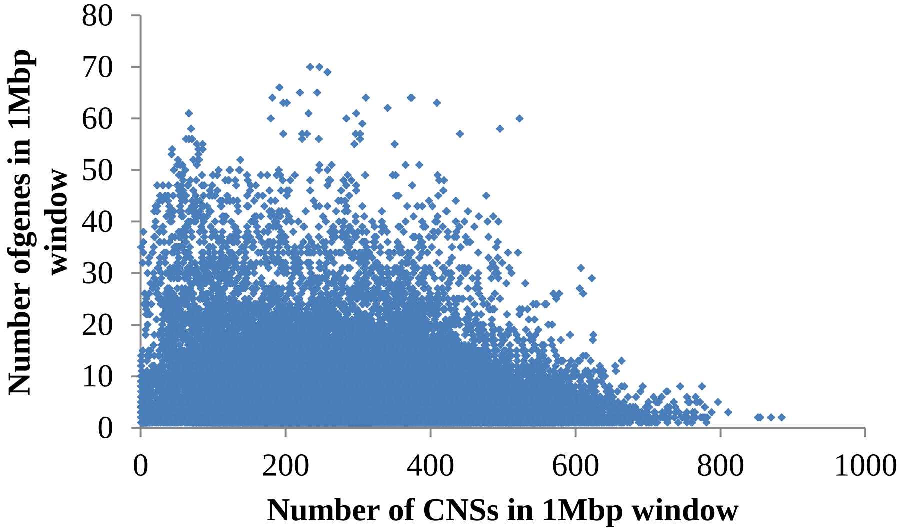


Figure A4.6: Distribution of CNSs and genes. (A) Distribution on human chromosome 1. CNSs and genes are not found around the centromeres and there distribution on the chromosome is not uniform. Each point represents 1Mbp window and the window was moved at 100kbp per step. (B) Distribution of all windows with at least a CNS and a gene. Pearson's r =-0.2234 (P-value<8.069×10^{-299}).

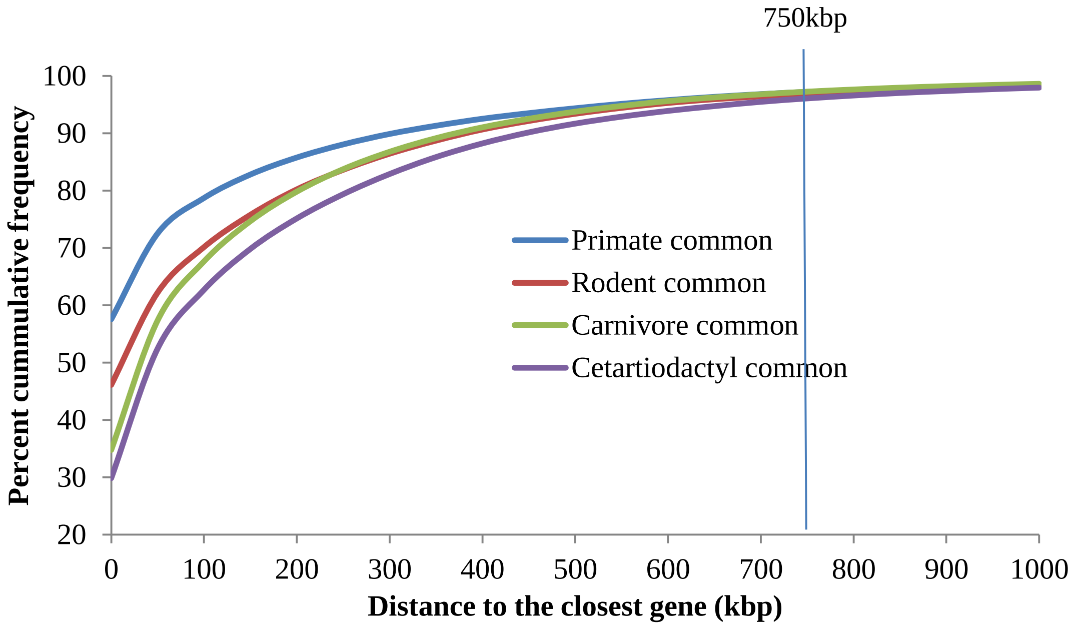


Figure A4.7: The proximity of the CNSs to genes. The horizontal axis represents the distance of the CNS to the closest protein coding gene.

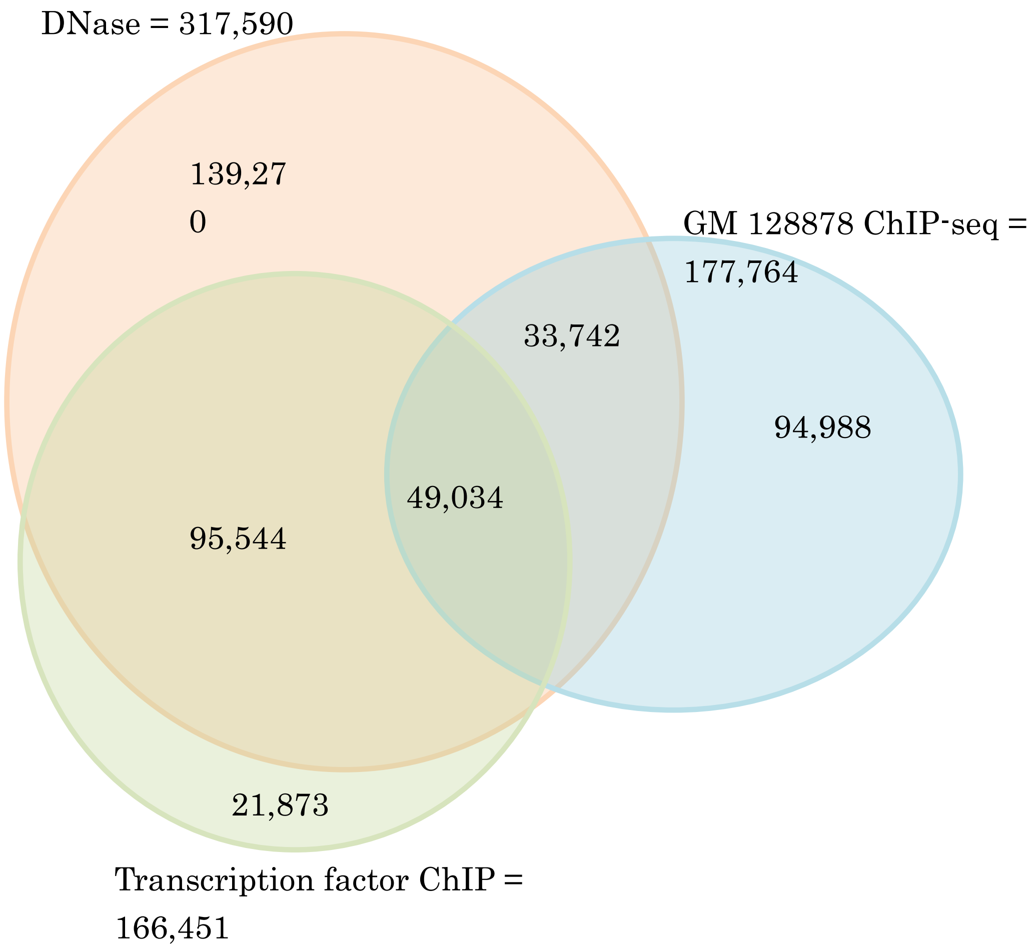


Figure A4.8: Overlap of primate common CNSs with selected regulatory regulation. The numbers of primate common CNSs that overlap the regulatory signatures are shown. In total, 434451 CNSs overlap with at least one of the three regulatory signatures. GM128878 ChIP-seq are the data of GM128878 cell line from ENCODE project while the transcription factor ChIP are the transcription factor binding site clusters.

Table A4.5: The genomic locations of the orthologous eutherian common CNSs between pairs of species

Species		**Total**												
Mouse	**Inter**	**18342**	**16582**	1551	39	170								
	Intr	**15209**	410	**14467**	12	320								
	Prom	**197**	32	32	**92**	41								
	UTR	**2158**	101	154	21	**1882**								
Dog	**Inter**	**22785**	**16861**	4910	71	943	**17895**	3926	86	878				
	Intr	**11885**	201	**11172**	11	501	352	**11166**	17	350				
	Prom	**259**	15	70	**59**	115	20	71	**66**	102				
	UTR	**977**	48	52	23	**854**	75	46	28	**828**				
Cow	**Inter**	**24807**	**17068**	6390	85	1264	**18189**	5353	90	1175	**21408**	2803	64	532
	Intr	**10122**	44	**9712**	2	364	131	**9750**	4	237	1108	**8938**	33	43
	Prom	**292**	7	87	**76**	122	11	83	**99**	99	43	62	**123**	64
	UTR	**685**	6	15	1	**663**	11	23	4	**647**	226	82	39	**338**
		Total	**17125**	**16204**	**164**	**2413**	**18342**	**15209**	**197**	**2158**	**22785**	**11885**	**259**	**977**
			Inter	**Intr**	**Prom**	**UTR**	**Inter**	**Intr**	**Prom**	**UTR**	**Inter**	**Intr**	**Prom**	**UTR**
		Species	**Human**				**Mouse**				**Dog**			

Inter - intergenic; Intr - intronic; Prom - promoter; UTR - untraslated region

The numbers in grey background represent the values of orthologous CNSs located on the same genomic locations genomic locations between the pair of species. Although for all pairs considered majority of the CNSs are located in homologous regions, some are located in different regions.

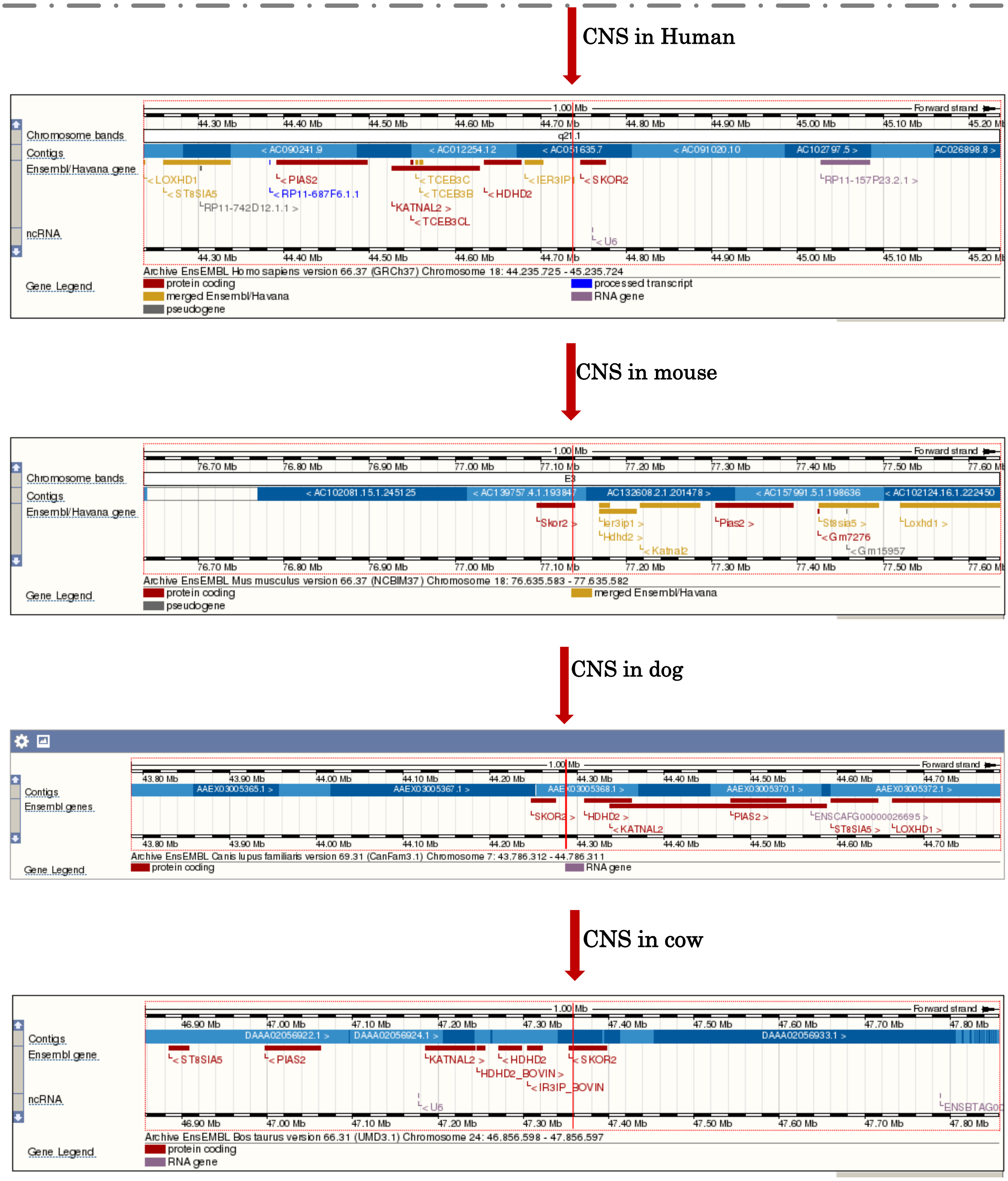


Figure A4.9: An example of CNS located on different genomic region. The position of the CNS is represented by the red vertical line at the center of each representation. In cow and mouse, the CNS is located in the intron of SKOR2 gene but in human and dog, it is found in the intergenic region outside SKORE2 gene.

Table A4.6: The gene ontology analysis using binomial test as described by PANTHER

Biological process	Tetrapod common	Primate unique	Rodent unique	Carnivore unique	Cetartiodactyl unique	Mouse-lost
Transcription	7.44E-108	0.046277	1.13E-27	0.414611	0.010728	9.37E-244
Development	3.18E-119	0	9.95E-20	3.10E-08	3.44E-05	0
Nervous system	3.58E-48	2.57E-300	3.77E-12	8.05E-06	0.405432	5.59E-134
Response to stimulus	3.64E-07	4.51E-180	0.024622	0.0811	0.406716	1.32E-44
Immune and defense	1.76E-13	3.48E-83	0.215629	0.698506	0.00055	7.23E-45

The values in gray shade are significant (Binomial P-value<0.001). Transcription, development and nervous system related genes are overrepresented while response to stimulus, immune and defense related genes are underrepresented. Genes under negative selection are more associated with CNSs while genes under positive selection are less associated with selective constraint.

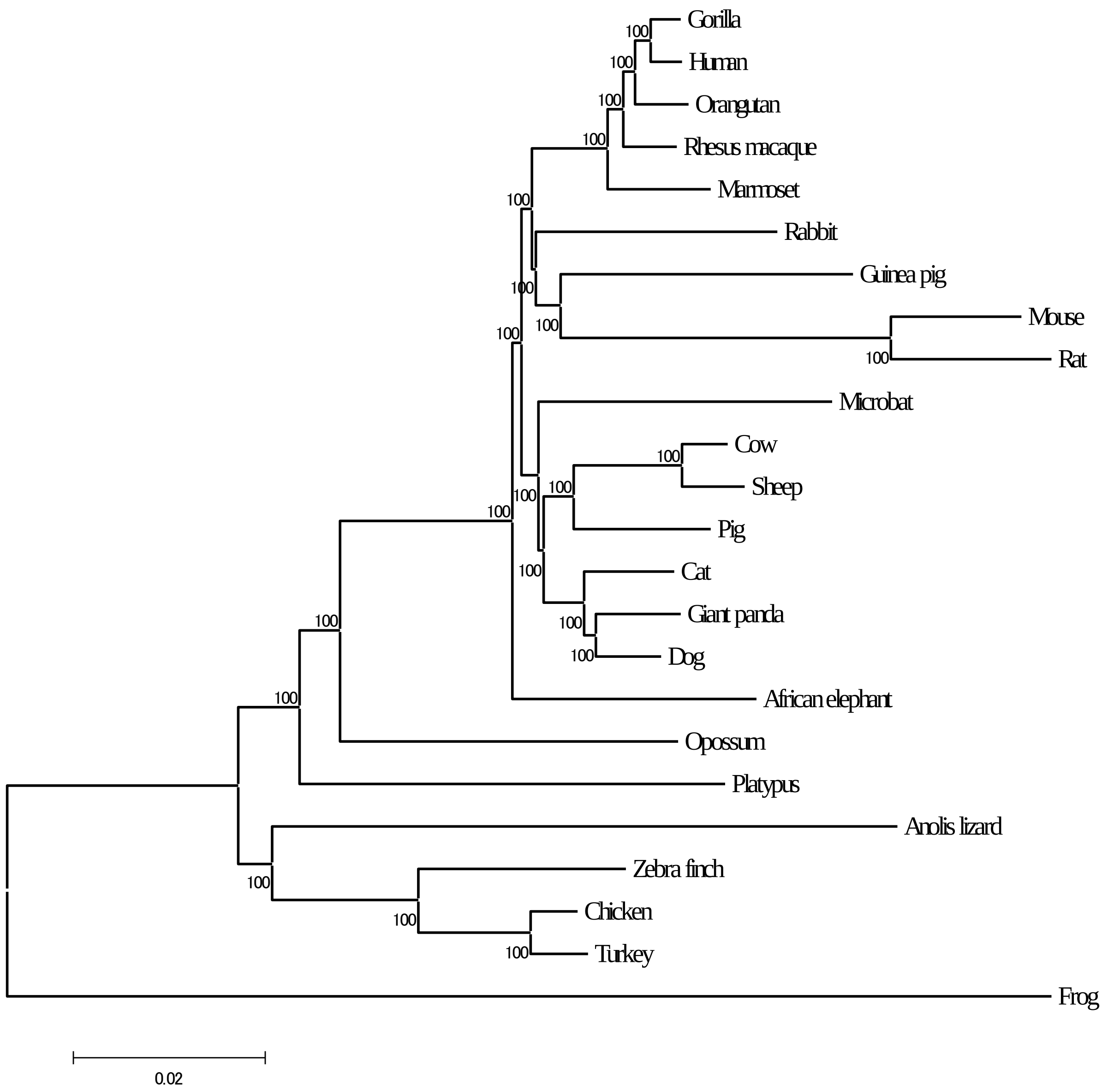


Figure A4.10: The phylogenetic tree of the CNSs. The concantenated tetrapod common CNSs were used to construct the phylogenetic tree using NJ method. All branches were supported with 100% boolstrap values.

Appendix 5

Table A5.1: Sources of data for species used for evolutionary analyses

Species	Gene sequences	Database
Mouse	Available	Ensembl
Rat	Available	Ensembl
Naked mole rat	GMAP-predicted*	UCSC
Akanezumi	GMAP-predicted**	This study

*Augustus-predicted gene sequences available on UCSC database gave fewer alignments. I then decided to extract the genes using GMAP with guinea pig genes as reference.

**For the prediction, Trinity-assembled transcripts from the combination of all sample reads were used to produce coding sequences. GMAP was used with mouse coding genes as query.

Table A5.2: Retrieval of liver expression data

Species	Tissue type	Data source	Short read ID	GEO Accession
Mouse	Adult liver	GEO	SRR594397	GSE41637
Rat	Adult liver	GEO	SRR594432	GSE41637
Naked mole rat	4yo liver	GEO	SRR306395	GSE30337
Naked mole rat	20yo liver	GEO	SRR306396	GSE30337
Akanezumi	Adult liver	This study	Not available	Not available

***yo = year old

Figure A5.1: Quality of reads

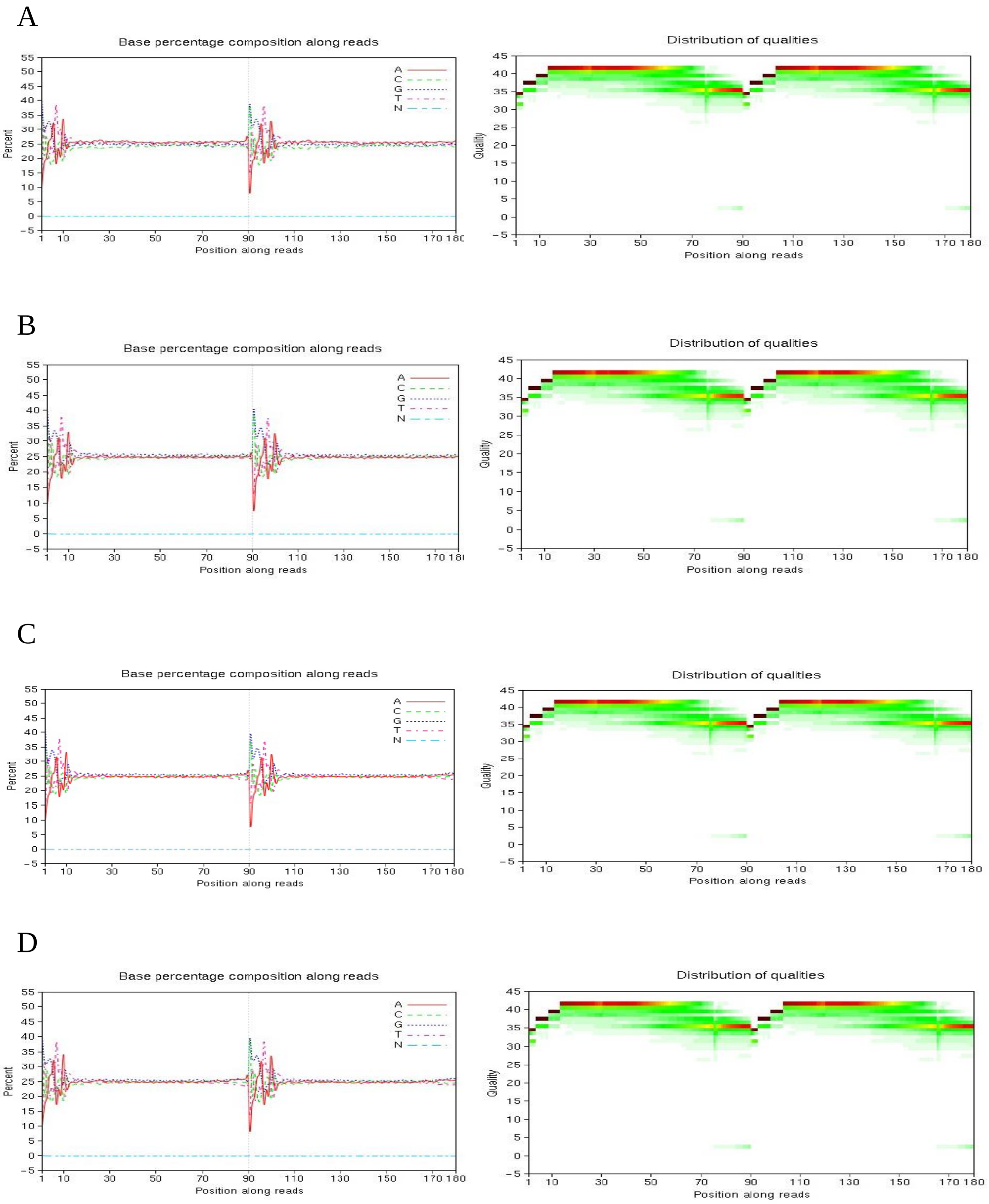


Figure A5.1: Base composition and quality distribution of akanezumi transcriptome data. The base composition (left panel) and quality distribution (right panel) are shown for adult liver sample (A), 3- (B), 5- (C) and 7-days postnatal (D) transcriptomes.

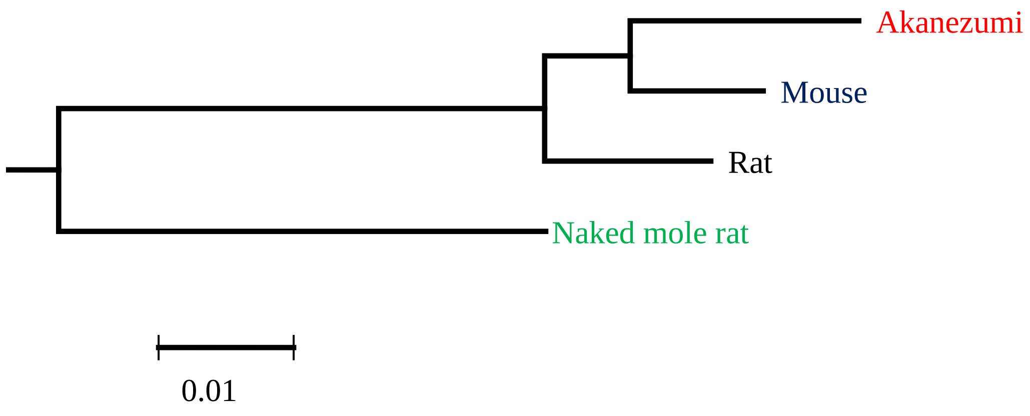


Figure A5.2: Phylogenetic tress from first codon positions.

Table A5.3: Spearman correlation matrix for all transcripts expressed in all species

	Mouse	Rat	Akanezumi	Molerat20yo	Molerat4yo
Mouse	1	0.776999	0.800422	0.617523	0.610981
Rat	0.776999	1	0.757272	0.618166	0.600955
Akanezumi	0.800422	0.757272	1	0.615158	0.605428
Molerat20yo	0.617523	0.618166	0.615158	1	0.794985
Molerat4yo	0.610981	0.600955	0.605428	0.794985	1

Molerat20yo = 20 years old naked mole rat

Molerat4yo = 4 years old naked mole rat

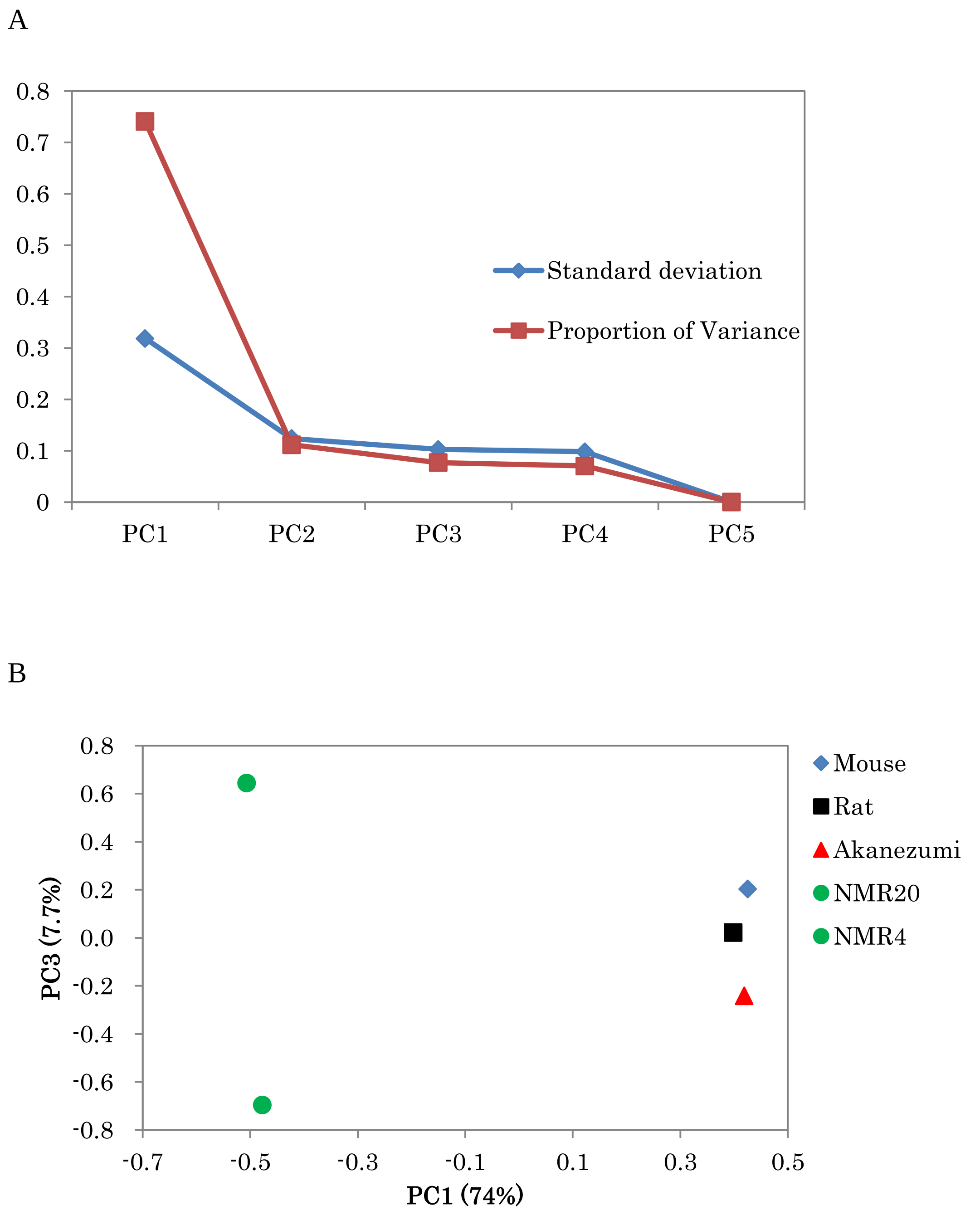


Figure A5.3: Principal component analyses of expression dynamics (A) Standard deviation and proportion of variance explained by PCA in Figure 5.2B. (B) PC plot for first and third components. NMR20 and NMR4 are 20 and 4 years old naked mole rats, respectively.

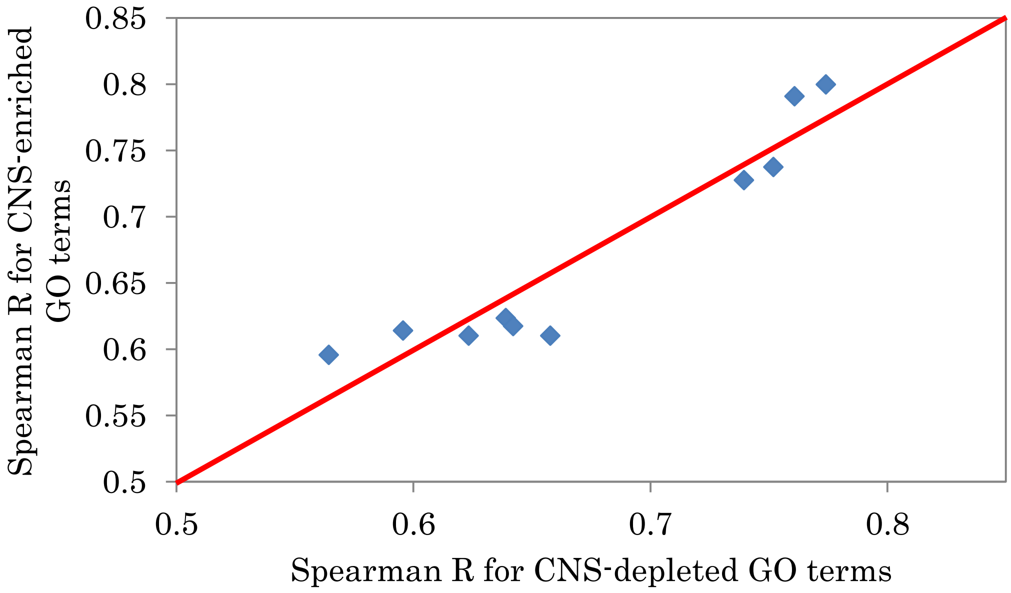


Figure A5.4: The expression correlations in liver does not reflect CNS GO enrichment. Each point represents each pair of species. The top most point represents the two naked mole rat, while the lowest point represents mouse and 20 year old mole rat.

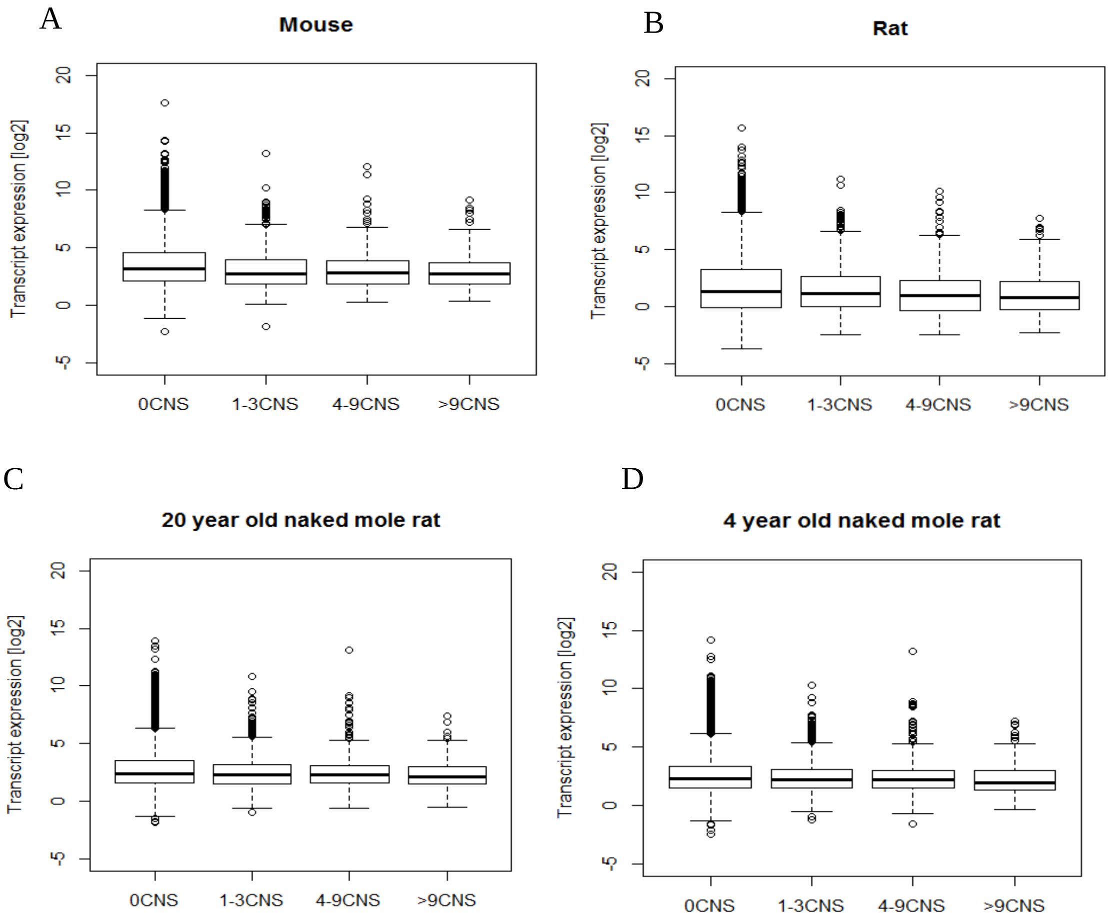


Figure A5.5: CNS-associated genes have relatively lower expression in liver. The boxplots for mouse (A), rat (B), 20 year old naked mole rat (C) and 4 year old naked mole rat (D) are shown.

Appendix 6

Table A6.1: Capybara sequencing statistics

Parameters	Forward reads	Reverse reads
Number of reads	78,804,042	7,8804,042
Total bases	23,460,242,033	18,607,187,591
Determined bases (ACTG)	23,445,468,372	18,573,695,574
GC content	41.285%	41.714%
Length (longest)	351	251

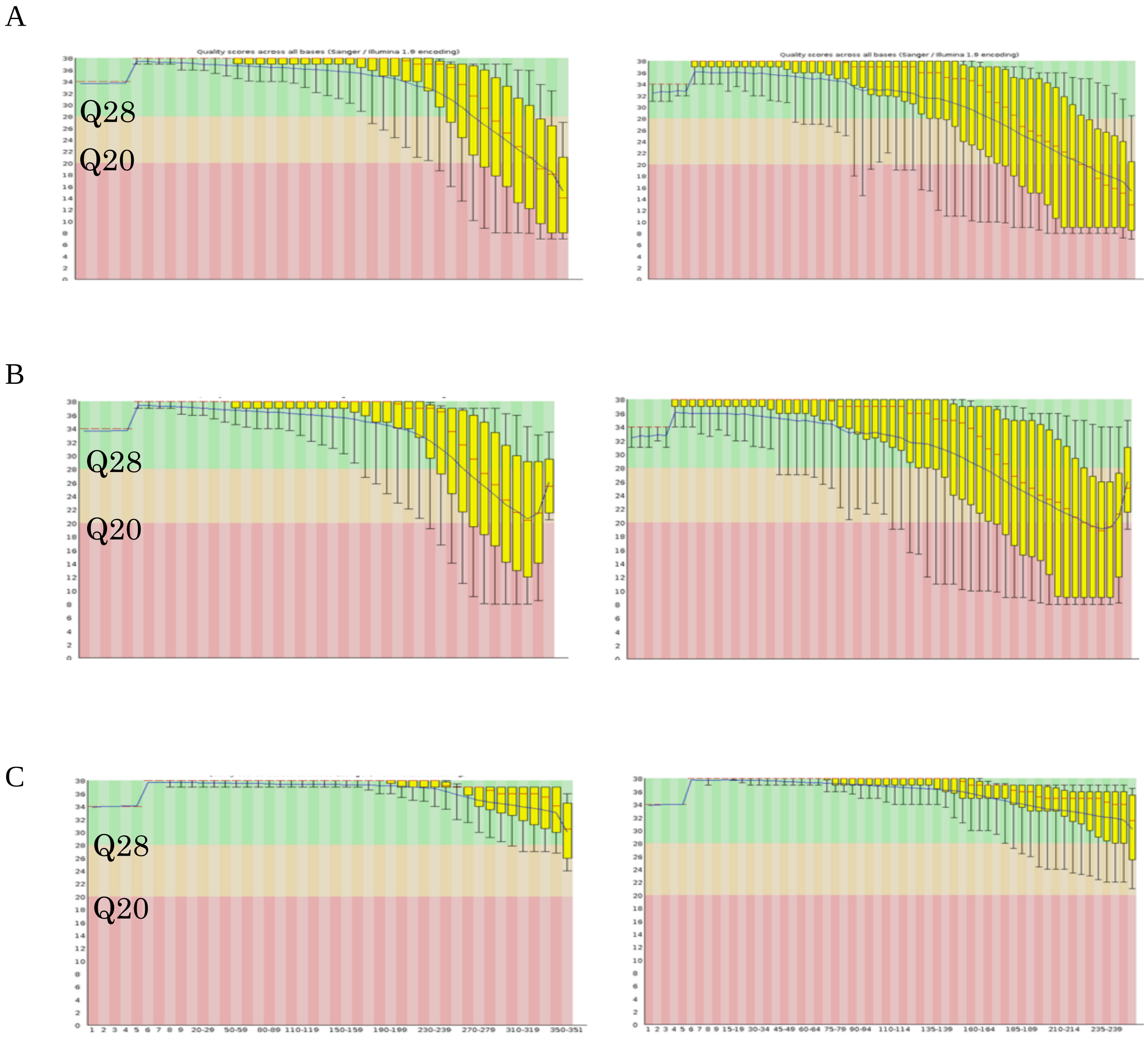


Figure A6.1: Read quality filtering. The left panel represents the quality distribution of the forward reads while the right panel represents the quality distribution for the reverse read. (A) The unfiltered reads (B) Fastx toolkit trimmed (Q20) reads (C) Reads after critical filtering with a python script that I wrote.

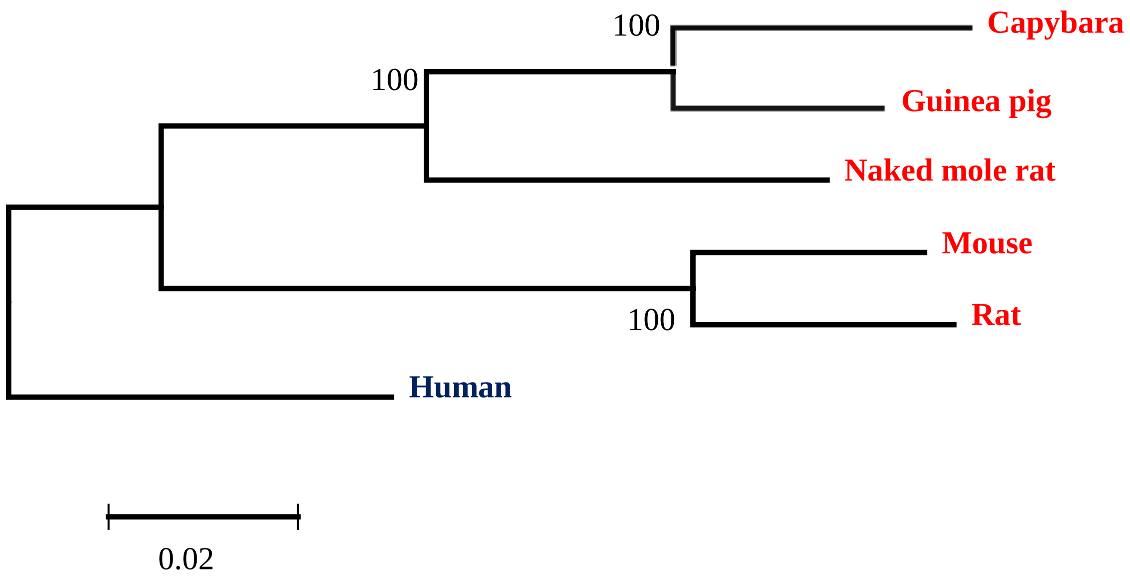
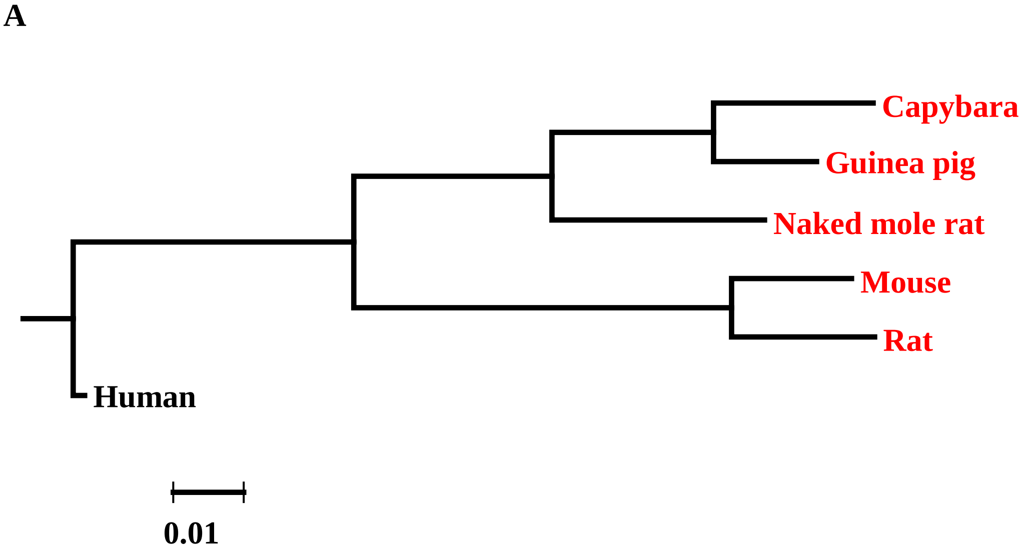


Figure A6.2: Phylogenetic relationship of rodent species plotted from concatenated amino acid sequences. The tree was from ~2.9million sites with no gap in any of the species used. Poisson model was set with Gamma parameter k =5. Notably, the phylogenetic relationship was well reproduced with 100% bootstrap support.



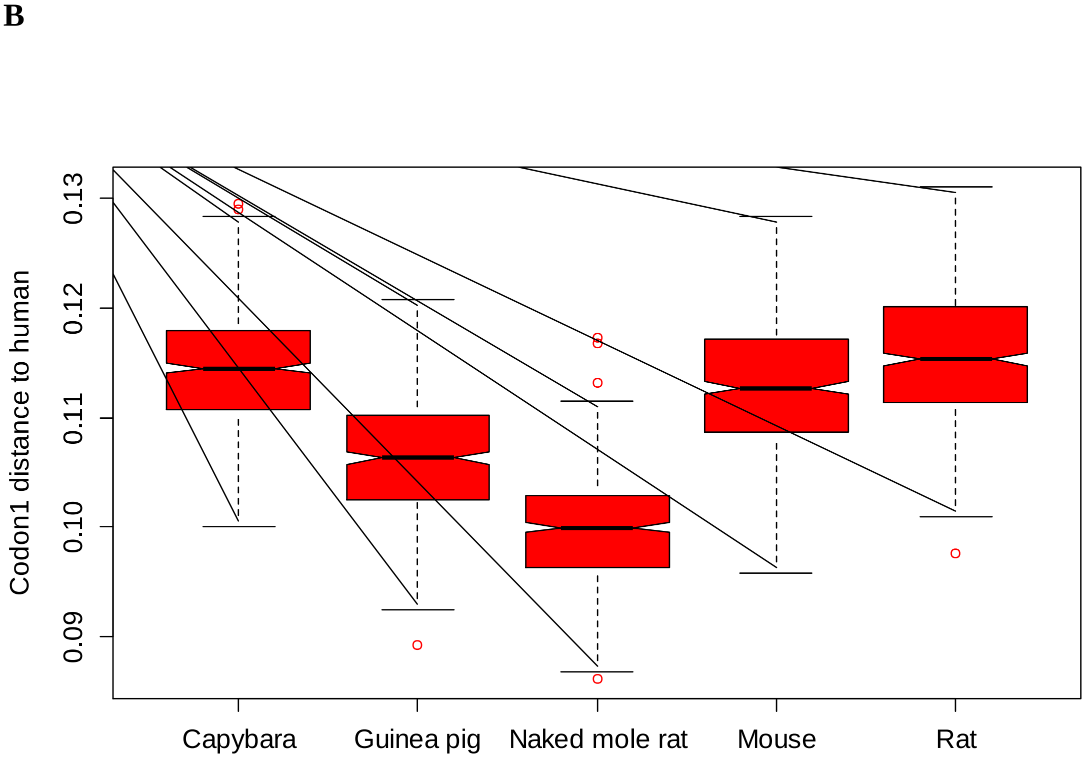
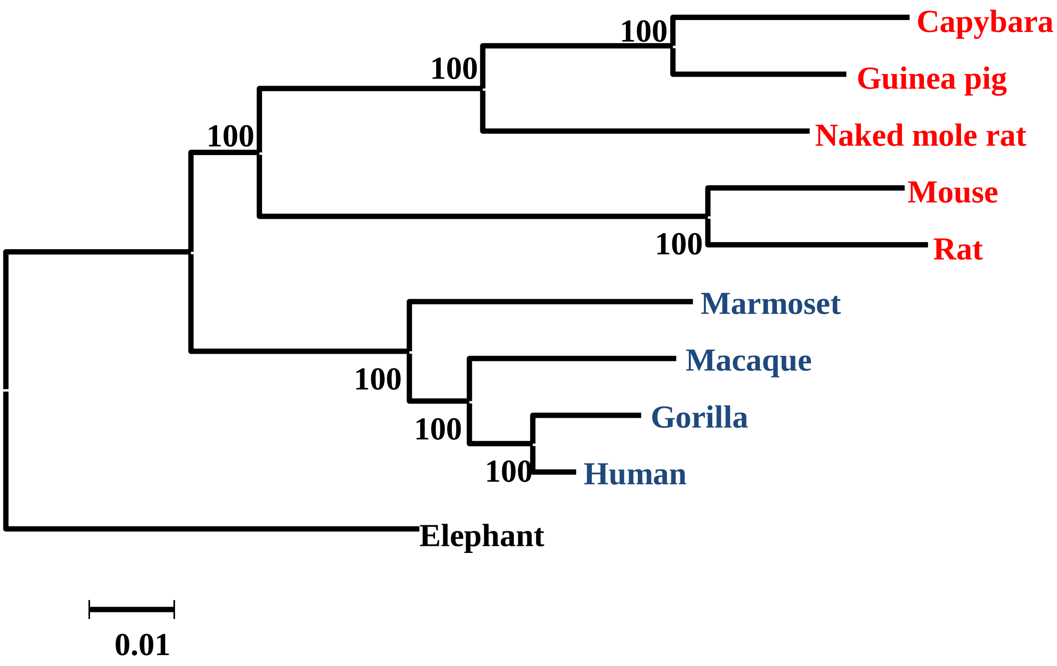


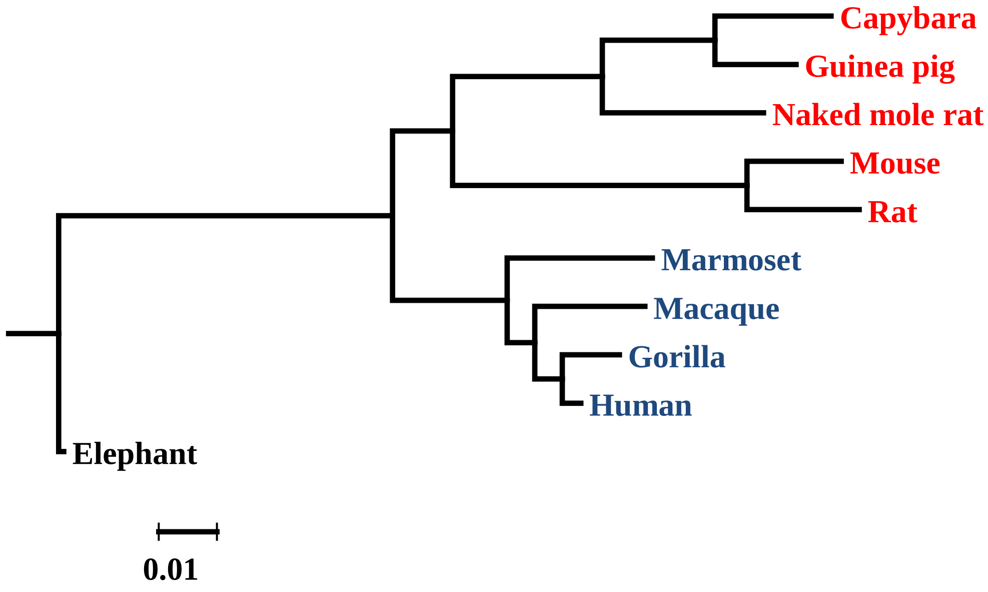
Figure A6.3: Genetic distance of first codon position. (A) Phylogenetic tree plotted with the nucleotides of the first codon positions. (B) Codon 1 distance to human.



T

Figure A6.4: Phylogenetic relationships of the Euarchontoglire from amino acid sequences.

A



B

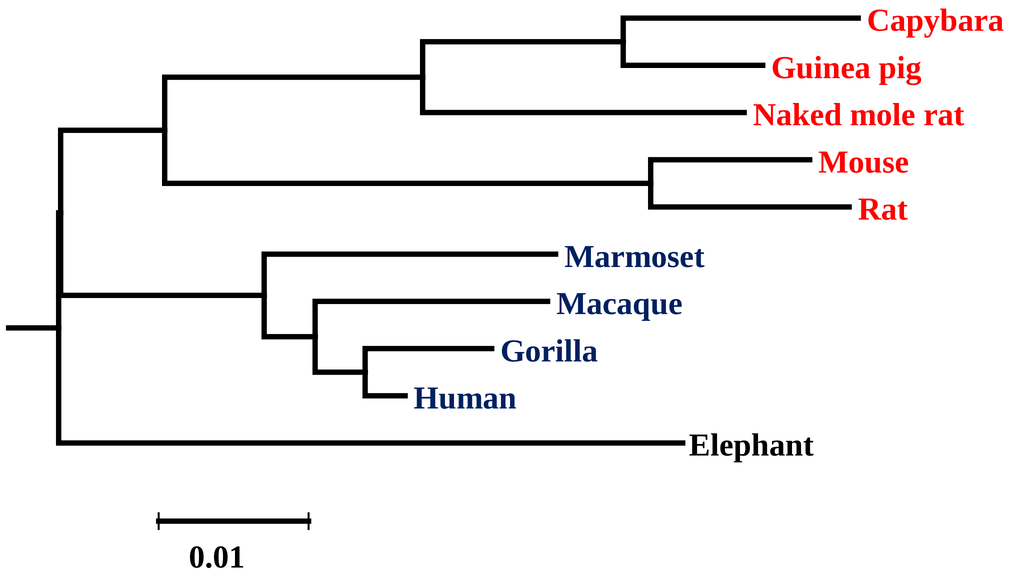


Figure A6.5: The phylogenetic tree of the Euarchontoglire from first (A) and second (B) codon positions.

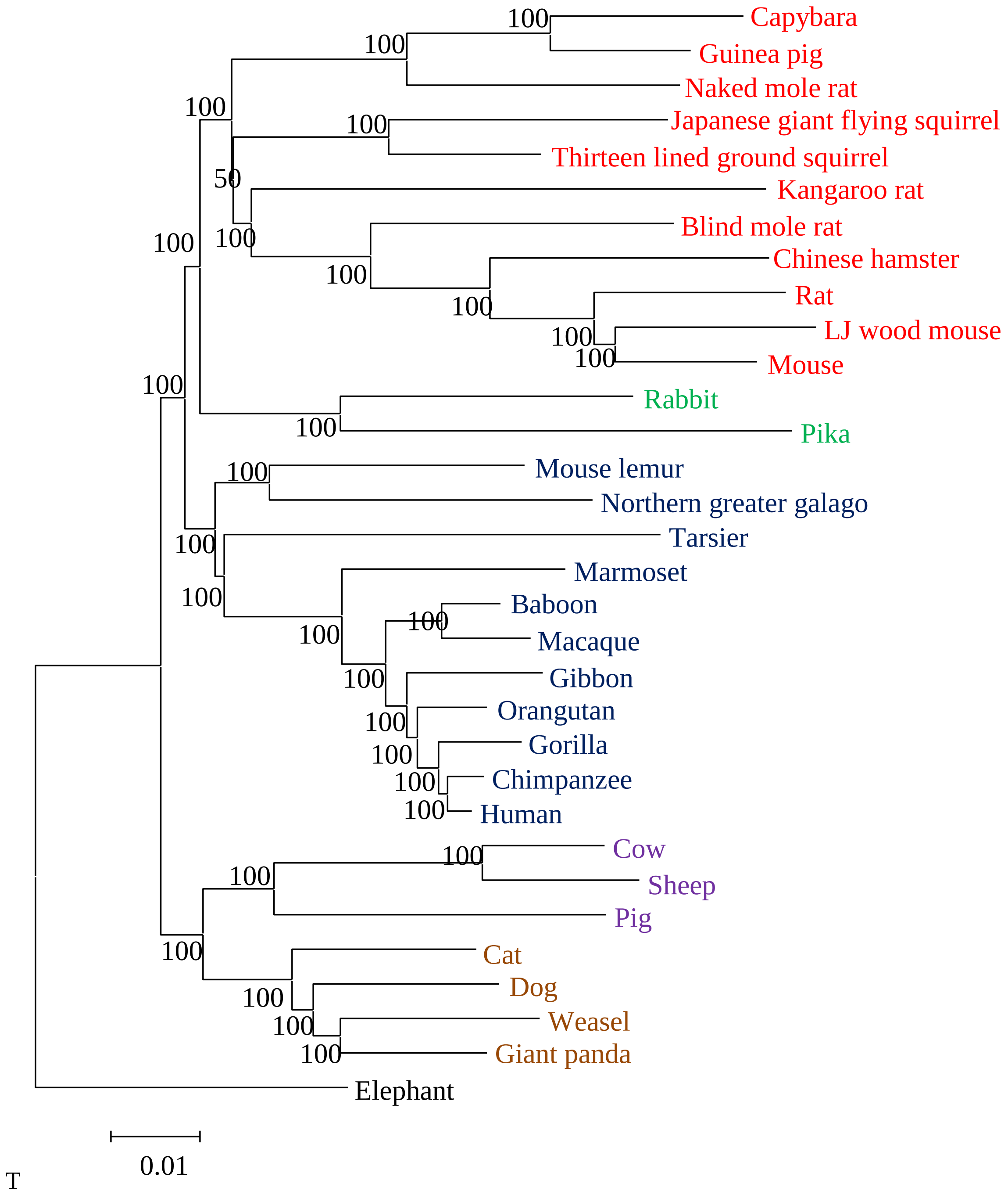


Figure A.6.6: Amino acid phylogenetic relationships of higher number of species. The phylogenetic tree was computed from ~112 thousands gapless amino acid positions.

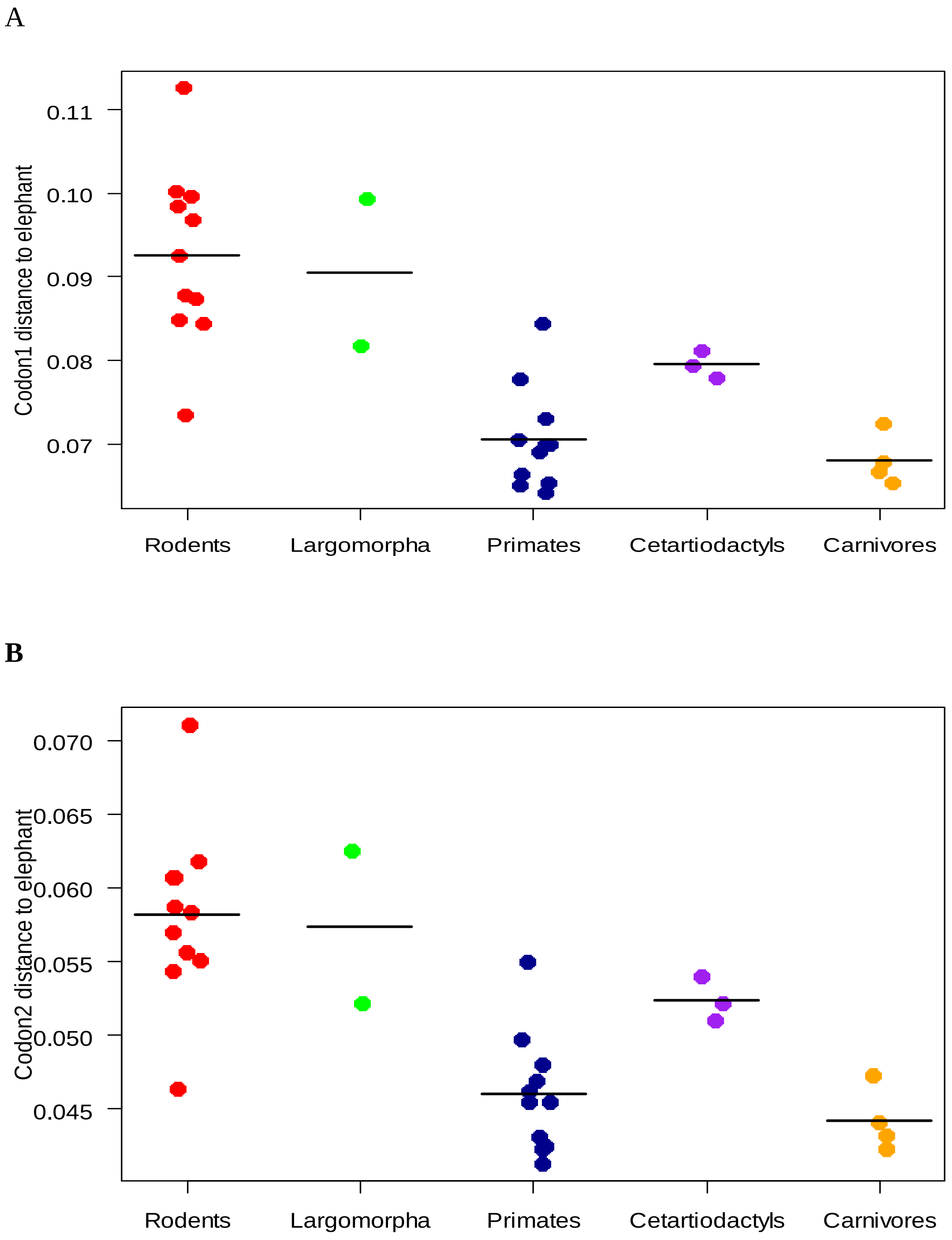


Figure A6.7: Rodents have higher and more heterogeneous distances. (A) First codon position distance to elephant. (B) Second codon position distance to elephant. The horizontal black line represents the mean for each order.

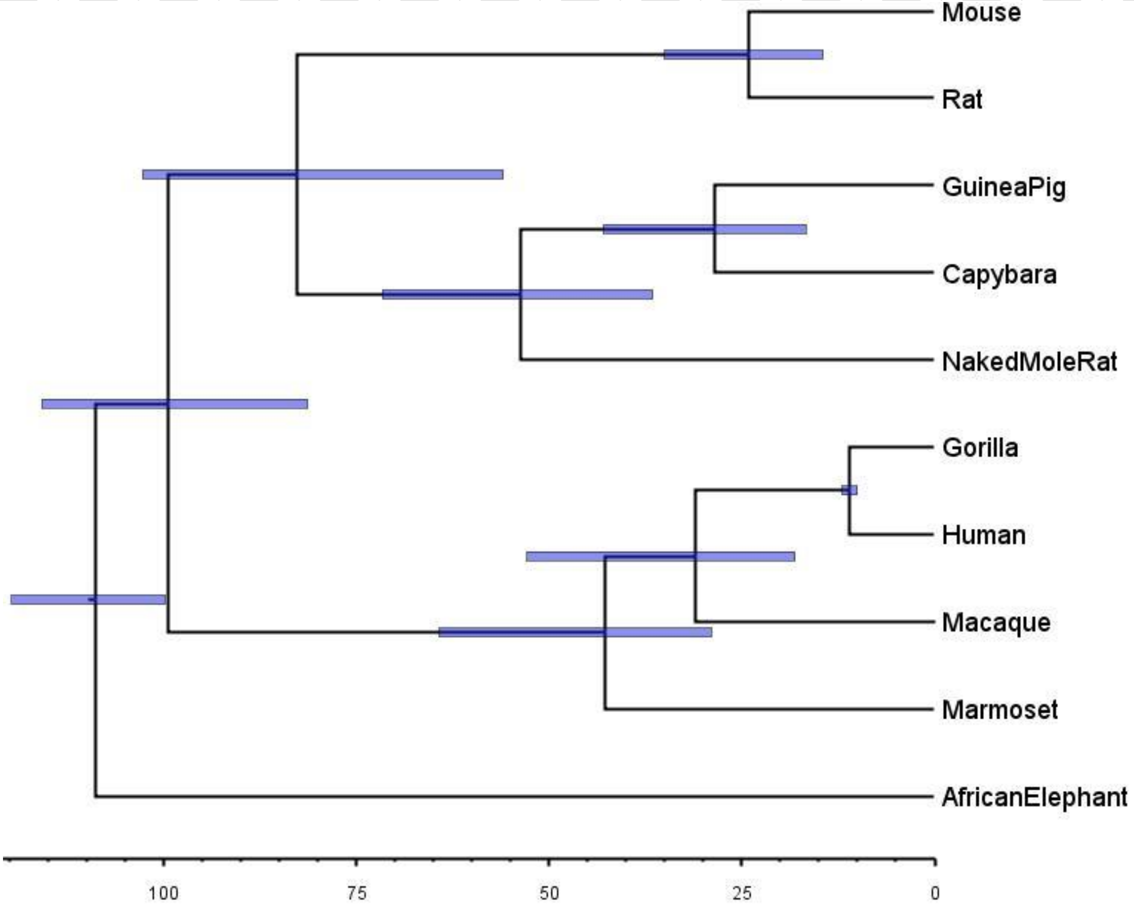


Figure A 6.8: Divergence time estimation with MCMCtree. The blue bars represent 95% confidence intervals.