Heterogeneous characteristics of Conserved Noncoding Sequences (CNSs) in Eukaryotes

Nadeeka Nilmini Hettiarachchi

Doctor of Philosophy

Department of Genetics
School of Life Science
SOKENDAI (The Graduate University for Advanced Studies)

Heterogeneous characteristics of Conserved Noncoding Sequences (CNSs) in Eukaryotes

By Nadeeka Nilmini Hettiarachchi

Acknowledgements

I am indebted to all those who assisted me in numerous ways, throughout the past several years, filled with hard work and tasking times. I owe everyone who were there by me and guided me with their thoughtful remarks, probing, challenging questions and suggestions including those who assisted me in numerous ways with life and living in Japan. Although one page acknowledgement would not suffice to express my sincere heartfelt gratitude to all those who sincerely meant my well-being, their names and deeds will always find room in my heart.

First I wish to extend my sincere gratitude to Prof. Naruya Saitou for his valuable advices and support during the time of my research. Thank you very much for your encouragement and patience.

Next I wish to thank my committee members, Prof Kakutani Tetsuji, Prof. Nori Kurata, Prof. Asao Fujiyama and Associate Prof. Kazuho Ikeo for their indispensable suggestions and support during the gradual progress of this study. Also I am sincerely thankful to Associate Prof. Naoki Osada of Hokkaido University for various advices during his term as one of the committee members when he was in National Institute of Genetics.

At the same time I am thankful to the members in my lab for creating a pleasant working environment. Also I wish to extend my sincere gratitude to Mr. Isaac Adeyemi Babarinde for his constant support and consideration at all times.

I am sincerely obliged to thank the Ministry of Education, Sports, Culture, Science and Technology of Japan for providing me this opportunity to pursue my graduate studies in Japan.

Finally, to my parents, my brothers and my dearest friend Morteza Mahmoudi Saber who helped me maintain a proper perspective through times of adversity, I can no more than reaffirm my sincere gratitude and devotion to them.

Table of Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of figures

List of tables

List of abbreviations

CNSs: Conserved Noncoding Sequences

UCEs: ultraconserved elements

CNEs: conserved noncoding elements

UTR: Untranslated region

Ds: The distance of synonymous substitution

GO: Gene Ontology

TF: Transcription factor

Abstract

Comparative genomics approach has made it feasible to determine potential functional elements in the eukaryote genome via computational analyses. Studies on Conserved Noncoding Sequences (CNSs) have reached its climax where the computationally discovered functional elements in genomes have also been experimentally verified where they were found to function as regulatory elements governing gene expression. CNSs have been verified to function as enhancers, and as well repressors. Earlier all noncoding region of genomes was referred to as “junk DNA” but now it became clear that some noncoding regions which are evolutionarily conserved, namely CNSs, are as important as the coding regions.

I first determined lineage specific CNSs of eudicots, grasses, monocots, angiosperms, land plants and plants. Here I identified 27, 6536, 204, 19, 2 eudicot, grass, monocot, angiosperm, vascular plant specific CNSs respectively. Average percentage identity for CNSs in all groups was more than 80% and the average length varied across groups.

Since the number of CNSs varied considerably across lineages, I tried to determine if this could be due to evolutionary rate, number of species in each group or relatively short divergence time for grass lineage. Even with pairs of species from eudicots and grasses with same divergence time, eudicots had less number of CNSs than monocots. Next I showed that the number of species in each group cannot be a reason for the difference by considering same number of species from eudicots and monocots. In this case also eudicots had less number of CNSs (69) than monocots (204). With respect to the evolutionary rate I found that eudicots have a saturated Ds (2.44) compared to grasses and monocots. If evolutionary rate played a key role in number of CNSs, lineage specific genes should follow the same pattern in abundance as CNSs. Surprisingly Eudicots had more lineage specific genes (2439) than CNSs whereas monocots and grasses (113, 444) had less lineage specific genes. I also found that UTR (untranslated region) CNSs were overrepresented with respect to other regions (introns and intergenic regions). The GO (gene ontology) analysis for the likely target genes of CNSs showed that these genes are related to transcription regulation and development. Another discovery was that CNSs are flanked by sequences showing an increase of GC content and CNSs were also GC rich. Next I tested if these high GC regions are recombination hot spots, as recombination hot spots are known to be related with high GC content in the human

genome. None of the CNSs overlapped with recombination hot spots, while high GC CNSs have a higher propensity to form nucleosomes.

The lineage specific plant CNSs I identified were all GC rich, but mammalian CNSs have been reported to have low GC content. CNSs of diverse lineages follow different patterns in abundance, sequence composition and location. I therefore conducted a thorough analysis of CNSs in diverse groups of Eukaryotes with respect to GC content heterogeneity. I examined a total of 55 Eukaryote genomes (24 fungi, 19 invertebrates, and 12 non-mammalian vertebrates) as to find lineage specific features of CNSs. Fungi and invertebrate CNSs are predominantly GC rich, whereas non-mammalian vertebrate CNSs are GC poor, similar to mammalian CNSs. This result suggests that the CNS GC content transition occurred from the ancestral GC rich state of invertebrates to GC poor in the vertebrate lineage probably due to the enrollment of GC poor binding sites that are lineage specific. To test how the transition of GC content could have occurred, I determined the GC contents of transcription factor (TF) binding sites and their level of conservation in outgroup lineages. GC content of CNSs also showed a correlation with the location in the genome; GC poor CNSs showed a higher probability to be located in open chromatin regions and GC rich CNSs showed a tendency to locate in heterochromatin regions. The histone modification signal analysis showed that CNSs overlapped with more H3K27Ac and H3K4Me1 compared to the random expectation. These histone marks are signatures of active enhancer regions. The predicted target genes of CNSs also agreed with the previous analysis where the most overrepresented GO term was related to transcription and development. Vertebrate ubiquitous TF binding sites are significantly GC rich compared to tissue specific ones. In contrast, plants have GC rich tissue specific binding sites compared to ubiquitous ones. This heterogeneity in GC content therefore must be attributable to TF binding sites and I found that vertebrate tissue specific TFs are more lineage specific than ubiquitous ones, whereas plant tissue specific and ubiquitous binding sites showed no significant difference in conservation implying the underrepresentation of tissue specific TFs among conserved TFs is a specific feature in vertebrate lineage.

Chapter 1

General Introduction

1.1 A general overview on Conserved Noncoding Sequences (CNSs)

Conserved sequences in the noncoding regions of genomes have been studied and examined for more than 20 years for their functional importance (Brenner et al. 1994; Jareborg et al. 1999; Pennacchio and Rubin 2001). Long stretches of persistently conserved regions in intron, intergenic and untranslated regions in the genomes has captured the attention of many researchers for their unique properties. In some instances these conserved noncoding regions were showing higher homology across species even than for protein coding regions (Pennacchio et al. 2006; Taher et al. 2011). Also these elements have been found spanning across diverse evolutionary times. For example Woolfe et al. (2007) documented that the conserved elements they found to span over 400 million years of evolution. I would like to call such sequences as “conserved noncoding sequences” or CNSs in this dissertation.

In numerous studies these CNSs have been designated in varying terms such as UCEs - ultraconserved elements (Bejerano et al. 2004), CNEs - conserved noncoding elements, HCNEs - highly conserved noncoding element (Lindblad et al. 2005), HCNRs - highly conserved noncoding regions (de la Calle-Mustienes et al. 2005), Hyper-conserved elements (Guo et al. 2008) etc. These terminologies have been adapted in each investigation based on the criteria they used for the analysis of the conserved elements. For example ultraconserved elements followed a criterion of more than 200 base conservation with 100% for very closely related species such as human and mouse. Whereas Shin et al. (2005) and Pennacchio et al. (2006) adapted a relaxed threshold of 70% identity for a set of diverse organisms.

Nonetheless their functional importance stays unhindered irrespective of the terminology. Various studies on vertebrate CNSs (e.g., Bejerano et al. 2004; Lee et al. 2010; Matsunami et al. 2010; Takahashi and Saitou 2012; Matsunami and Saitou 2013; Babarinde and Saitou 2013), invertebrate CNSs (woolfe et al. 2004; Siepel et al. 2005) and plant CNSs (Kaplinsky et al. 2002; Guo et al. 2003; Inada et al. 2003; Kritsas et al. 2012; Baxter et al. 2012; Hettiarachchi et al. 2014) reported CNSs to potentially have regulatory functions related to transcription and development. It has also been found that CNSs are under purifying selection (Drake et al. 2006; Casillas et al. 2007; Takahashi and Saitou 2012; Babarinde and Saitou 2013). Some recent studies have experimentally verified the function of CNSs (Lee et al. 2010). And also there is evidence of active histone modification signals associated with the CNSs. Zheng et al. (2010) reported that Foxp3 gene is associated with CNSs enriched with H3K4Me1, H3K4Me3 histone modification signals. These histone modifications are typically known to be related with distal enhancers and active promoter regions consecutively (Heintzman et al.2007).

Primary investigations on CNSs have mainly been focused on vertebrates, but recent studies have shown CNSs are also present in invertebrates and plants. These conserved regions of which the functions are yet to be fully understood still remain an enigmatic area of discovery which is yet to be fully explained.

1.2 Conserved noncoding sequences in animal genomes

Conserved noncoding sequences have been identified in various organisms. One initial study on Ultraconserved elements in the human genome (Bejerano et al. 2004) kindled the curiosity of researchers with regards to conservation in the noncoding portions of the genome. Though it is expected that coding regions are conserved for their functional constraint, the noncoding conservation for long stretches with a very high percentage identity was a novel idea at the time. Bejerano et al. (2004) reported on 481 segments in the human genome longer than 200bp with 100% identity conserved across rat and mouse genomes. These segments were a combination of coding and noncoding regions with a high degree of conservation. Later numerous efforts were invested into studying CNSs in animals. It was found that these conserved regions are under purifying selection (Casillas et al. 2007) and also they are found close to genes that act as developmental regulators (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Elgar 2009; McEwen et al. 2009; Lee et al. 2010; Takahashi and Saitou 2012; Matsunami and Saitou 2013; Babarinde and Saitou 2013). Lee et al. (2010) studied ancient vertebrate CNSs in bony vertebrate lineages, and CNSs they found were in abundance in tissue specific enhancers and they are likely to be cis-regulatory elements that are functionally conserved through evolution. Furthermore, *C. elegans* has a unique set of CNSs that are not found in vertebrates and are associated with nematode regulatory genes (Vavouri et al. 2007). Clarke et al. (2012) experimentally verified the functions of two CNSs conserved between vertebrates and invertebrates which have repression function on central nervous system and hindbrain. The functions of the CNSs have also been experimentally verified in numerous instances. Lee et al. (2011) found that many of the ancient vertebrate CNSs overlapped with experimentally verified human enhancers and their functional assay of ancient vertebrate CNSs in transgenic zebrafish embryos showed that some elements were able to recapture the mouse mid brain and hind brain expression in zebrafish. More recently Sun et al. (2015) reported that some

CNSs related to lhx5 gene showed expression in the forebrain region of the teleost fish. Lineage specific CNSs specifically have been verified to establish lineage specific features in taxa. Since the protein coding genes are more or less persistently conserved across lineages even with long divergence times the morphological and physiological differences of organisms might have risen from regulatory elements that are specific to each lineage (Davies et al. 2014; Hettiarachchi et al. 2014). Cretekos et al. (2008) reported that bat Prx1 enhancer increases the forelimb length of tested mice. Also the loss of an enhancer region *Pitx1*gene causes pelvic reduction in stickleback (Chan et al. 2010). Above mentioned are some of the evidences to support lineage specific features of organisms.

For a while it was argued that these persistently conserved sequences in the noncoding regions might merely be mutational cold spots (Clark 2001). However, Drake et al. (2006) reported that the conserved noncoding sequences are not mutational cold spots but actually under selective constraint by showing that CNSs have suppressed derived allele frequencies. Takahashi and Saitou (2012) and Babarinde and Saitou (2013) also showed this pattern in mammalian CNSs.

These numerous studies have set various selection criteria for the CNSs, such as the length, e-value and percentage identity in BLAST homology searches. Some studies have used very stringent criteria for identification of CNSs such as >100bp regions of 100% identity (Bejerano et al. 2004). Some of the other criteria being used are 70% over 100bp (Duret et al. 1993; Lee et al. 2011), 95% over 50bp (Sandelin et al.2004) and 95% over 500bp (Janes et al. 2011).

Irrespective of the selection criteria, in general, animal CNSs tend to be much longer and more abundant in number. Babarinde and Saitou (2013) reported having found about 52,000 primate specific, about 12,000 eutherian specific conserved noncoding sequences by considering sequences over 100bp and 94% identity set based on protein level conservation of genes. Kim and Pritchard

(2007) reported on about 82,000 mammalian specific CNSs (over 90% identity without length criteria). It is important to consider various filtering thresholds when dealing with closely related species to remove false positives from the set of putative CNSs that might appear conserved due to lack of time necessary for them to accumulate mutations and diverge.

Even though there is a plethora of documented evidence for conserved noncoding sequences in animal genomes, CNSs are not restricted to animals. CNSs have also been identified and studied in invertebrates and in plant genomes to varying degrees.

1.3 Progression of conserved noncoding sequence studies on plant genomes

Initially instigated studies on plant conserved noncoding sequences focused on a limited number of orthologous genes in a set of genomes. Kaplinsky et al. (2002) compared the *lg1*gene in two grass species, namely maize and rice, to identify putative regulatory elements associated with this gene. They also tried to identify CNSs associated with 5 more genes (*Chalcone synthase, wx1, adh1, sh2, H^+ ATPase*) and reported that grass CNSs are smaller and less abundant than those of mammalian genomes studied at that time. Later Guo and Moose (2003) analyzed 11 orthologous maize-rice gene pairs to identify putative regulatory elements in their flanking noncoding regions. And the CNSs they identified had a minimal length of 20 bases. Inada et al. (2003) reported the first comparatively large scale analysis in to grass CNSs. This study included 52 annotated maize-rice genes considering 1000bp or more toward the promoter region of each gene and reported that the majority of the genes in the analysis at least had one CNS and most CNSs were short as less than 20 bases in length. This analysis agreed with what was reported by Kaplinsky et al. (2002). Still late into 2012 the plant CNS studies stayed limited to gene wise analyses which considered the upstream region of the genes of interest in search of regulatory elements. As genome sequencing expanded and more plant genomes became available plant CNS studies started to gain pace.

On the contrary to the previous studies that were restricted to a limited set of orthologous regions, I started my analysis on CNSs in plant genomes in 2011considering 15 species.

Kritsas et al. (2012) analyzed complete genomes of four species (*Arabidopsis thaliana, Vitis vinifera, Brachypodium distachyon,* and *Oryza sativa), and identified ultraconserved like elements. T*hey identified 36 highly conserved elements with at least 85% identity and are longer than 55bp *between Arabidopsis thaliana* and *Vitis vinifera* and also 4572 highly conserved elements between *Oryza sativa* and *Brachypodium distachyon*.D'Hont et al. (2012) compared *Brachypodium* *distachyon*, *Oryza sativa* and *Sorghum bicolor* and identified 16,978 CNSs which were defined as pan-grass CNSs. Further with the addition to *Musa acuminata* they identified 116 CNSs in the commelinid monocotyledon lineage. Both these studies have concentrated on CNSs that are commonly found in the groups of species in their analyses, and also these CNSs might also be present in other lineages. Haudry et al. (2012) reported on conserved noncoding sequences responsible for crucifer regulatory network. Hettiarachchi et al. (2014) found that grasses and monocots generally had more linage specific CNSs compared to eudicots.

Similar to CNS analyses on other organisms Plant CNS studies also found that the predicted target genes of the CNSs are related to transcription regulation and DNA binding (Kritsas et al.2012; Hettiarachchi et al. 2014)The plant CNSs are also found to be enriched in transcription factor biding sites (Hettiarachchi et al. 2014; Burgess and Freeling 2014). Haudry et al. (2012) further reported that the 44% of the CNSs they identified overlapped with DNase I hypersensitive sites implying regulatory activity. Function wise plant and animal CNSs seem to be following the same pattern but the lengths and abundance of CNSs seem to vary widely between animals and plants. Even though many efforts have been invested into functionally verifying the importance of the CNSs in animals, plant CNSs are not extensively studied or analyzed with respect to function. This is one aspect of plant CNS analyses that is yet to be fully examined and expanded if we need to fully understand the regulatory architecture of plant genomes.

1.4 Objectives of this study

So far some of the documented CNS studies emphasize on lineage specificity brought about by regulatory elements that are unique to a particular taxa or clade. Mainly these studies have focused on vertebrate genomes (e.g., Takahashi and Saitou 2012; Matsunami and Saitou 2013; Babarinde and Saitou 2013) and elaborated on lineage specific characteristics related to vertebrates that must have driven their unique features.

There have also been studies on lineage specific expression of experimentally verified regulatory elements such as enhancer element related with *Olig2* (Sun et al. 2006) that drive expression in ventral spinal cord of tested transgenic mice and further they report this element is responsible for lineage specific expression in motor neurons. Also craniofacial enhancer I37-2 is therian specific and it has contributed to the craniofacial development in the mammalian lineage (Sumiyama et al. 2012).

These evidences propose the existence of the lineage specific regulatory architecture that governs group specific features. Computational comparative genomics analyses on DNA level is the first approach to uncover such regions that are conserved throughout evolutionary history. Finding conservation in the noncoding regions of genomes through computational analyses draws the baseline in identifying such potential regulatory elements. This idea of lineage specificity with regards to conserved noncoding sequences in plants has not been tested or adapted before. This prompted me to uncover the lineage specific conserved noncoding sequences that are most likely to be putative regulatory elements of plants and also to elaborate on the CNS features compared to vertebrates.

In chapter 1 I address the lineage specificity of CNSs in several plant taxa namely eudicots, grasses, monocots, angiosperms, land plants and plants. The initial objectives of the first part of the study was,

- To identify the CNSs that presumably originated in their respective common ancestors
- Determine likely target genes and their potential functions
- Determine any abundance differences in the lineage specific CNSs
- Determine specific characteristics of these CNSs with regards to
    - Dinucleotide distribution in and around CNSs
    - Relationship of CNSs and nucleosome occupancy
    - CNSs with regard to methylation level
- Identify patterns of lineage specific loss of ancestral CNSs
- Identify patterns of lineage specific CNSs and lineage specific genes

I focused on the GC content heterogeneity of CNSs across different lineages such as fungi, invertebrates, non-mammalian vertebrates and plants in Chapter 2. As identified in Hettiarachchi et al. (2014) plant CNSs are GC rich. Babarinde and Saitou (2013) showed that mammalian CNSs they identified are GC poor. This lineage difference in GC content is the main focus of this chapter. Objectives of this study involved

- Determining the origin of the GC poor CNSs
- Various characteristic features of CNSs belonging to different lineages
- Plausible reasons for GC content heterogeneity

All the analyses were performed via custom made perl scripts which were produced by myself and genome data were retrieved from public genome repositories.

1.5 Overall map of the Dissertation

This dissertation all together entails four chapters with two main chapters that address different aspects of conserved noncoding sequences in organisms.

Chapter 1(this chapter) is documented to give a general overview of conserved noncoding sequences, the development of CNS studies throughout the years with regards to animal and plant CNSs.

Chapter 2 mainly deals with the analysis of the lineage specific conserved noncoding sequences in plant genomes belonging to various taxa. This chapter is fragmented into subdivisions that handle the materials and methods, results and discussion and is intended to answer the objectives listed in section 1.4. Chapter 2 has been published in Genome Biology and Evolution in 2014 September (doi: 10.1093/gbe/evu188).

Chapter 3 was documented with the expectation of answering the key question that arose with regards to CNS GC content heterogeneity of different lineages. This chapter tries to elaborate the work flow to the analysis, results based on GC contents of CNSs of different lineages, discussion and a conclusion for GC content heterogeneity of CNSs in diverse taxa.

Chapter 4 was written to summarize the discussions from chapter 2 and 3, explain plausible reasons to the observations and analyses and finally give a description on future directions with regards to this study.

Chapter 2

Determination of lineage specific plant CNSs

2.1 Introduction

With the ever-increasing high throughput genomic data it has become possible to understand and decipher the genomic properties and evolutionary aspects of various organisms. The handling of genomic data on various levels has made it possible to elucidate the properties of organisms on a functional level. Comparative genomic analyses have been considered to identify conserved potential regulatory elements through varying divergence times. This has been a sound method in identifying elements that are conserved on sequence level. It has been found that these CNSs that are computationally discovered are also functionally important in shaping the morphological and physiological aspects of an organism (Sun et al. 2015; Ritter et al. 2010).Many studies have been conducted in CNSs on animals and this study intended to set light on lineage specific conserved noncoding sequences in plants for the first time.

In this study I decided to focus on whole genome analysis of available plant genomes to identify lineage specific CNSs. My focus of this study is in finding all the CNSs, thus find potential regulatory elements specific to different plant lineages.

Here I searched eudicot lineage specific CNSs by analyzing genome sequences of the following seven eudicot species: *Arabidopsis thaliana, Brassica rapa, Populus tricocarpa, Ricinus communis, Vitis vinifera, Cucumis sativus,* and *Aquilegia coerulea.* To determine grass specific CNSs, genome sequences of *Oryza sativa, Brachypodium distachyon, Sorghum bicolor,* and *Setaria italica* were compared. The genome sequences of *Musa acuminata* were also analyzed to determine monocot specific CNSs in addition to the four grass species mentioned above. It has to be noted that in order to look for the specific CNSs in the analysis I have included the most basal species sequenced so far, assuming that if a CNS is present in the most diverged species, it is highly likely to be found in closer species inside a group. The most basal eudicot species used in the study is *A. coerulea* which diverged about 120 mya (Anderson et al. 2005) from the rest of the eudicot species used in this study. *Musa acuminata* is considered as the basal monocot species, which diverged from grasses about 115 mya (D’Hont et al. 2012). The other species used in the study are *Selaginella moellendorffii* which diverged from angiosperms about 400 mya (Banks et al. 2011), *Physcomitrella patens* which diverged 450 mya (Rensing et al. 2008) from vascular plants and *Chlamydomonas reinhardtii* that diverged from land plants more than 1000 mya (Heckman et al. 2001). A total of 15 species (see Figure 3 for their phylogenetic relationship) were used with the expectation of finding the group specific CNSs in this study.

2.2 Materials and Methods

2.2.1 Genomes considered in the analysis

Repeat masked genome sequences of *Arabidopsis thaliana, Brassica rapa, Populus tricocarpa, Oryza sativa, Brachypodium distachyon, Sorghum bicolor, Selaginella moellendorffii, Chlamydomonas reinhardtii* and *Physcomitrella patens* were downloaded from Ensembl release 12, whereas *Ricinus communis, Vitis vinifera, Cucumis sativus*, *Aquilegia coerulea, Setaria italica* were downloaded from Phytozome version 8.0. Genome sequences of *Musa acuminata* were downloaded from banana genome project database. Since the analysis was focused on the Conserved Noncoding Sequences (CNSs) in the nuclear DNA, the mitochondria and chloroplast genomes were removed from the analysis where they were known and annotated in the databases. Since there is also a possibility of mitochondrial and chloroplast sequences being transferred into nuclear genome, I further removed any sequences which showed homology to mitochondrial or the chloroplast genome before initiating any analyses on the respective sequences.

2.2.2 Identification of lineage common CNSs

Common to eudicots: BLAST 2.2.24+ (Altschul et al. 1997) was used for performing homology searches in this study. BLASTn search was done with *A. thaliana* as the query and *B. rapa* as the subject database. The cut off e-value for the search was 0.001. Only the alignments without any overlap with a coding region for both query and subject were used for subsequent analysis. From the remaining (nuclear DNA) BLAST hits, the best hits of overlapping alignments were selected using the e-value. If the coordinates of two sets alignments overlapped with each other only the alignment with the lower e-value was retained. Thus a dataset with the best alignments for *A. thaliana* and *B. rapa* conserved noncoding sequences was produced for further analysis. The obtained best hits were searched against *C. sativus*, thus *A. thaliana, B. rapa* and *C. sativus* best hits were obtained the same method explained above. Similarly, the best hits of *A. thaliana, B. rapa*, *C. sativus* were searched against *P. tricocarpa*. This method was carried out in form of a chain (best hits of previous step used to search a new species) for the following species, *R. communis, V. vinifera* and *A. coerulea* in the sequence given, to obtain the common CNSs to eudicots. These CNSs were in turn searched in Rfam v10.1 (June 2011) and the CNSs with overlaps with noncoding RNA were removed from further analysis.

Common to grasses: BLASTn search was done with *O. sativa* as the query and *B. distachyon* as the subject database. The cut off e-value for the search was 0.001. Only the alignments without any overlap with a coding region for both query and subject were used for subsequent analysis. The remaining hits were filtered based on the e-value and only the best hits were retained to search against *S. italica*. Then *O. sativa, B. distachyon* and *S. italica* best hits were searched against the last monocot genome, *S. bicolor*. This procedure finally achieves the CNSs that are found in all the monocots used in the study and thus were considered as grass common CNSs. The common CNSs were searched in Rfam v10.1 (June 2011) and the CNSs with overlaps with noncoding RNA were removed from further analysis.

Common to all monocots: The grass-common CNSs discovered from the previous step were searched in *M. acuminate* to obtain CNSs that are common in all monocot species used in the study. The cut off e-value for the search was 0.001.

Common to all angiosperms, to all vascular plants, and to all plants: The eudicot common and monocot common CNSs were searched against each other with a cut off e-value of

0.001, using the eudicot common CNSs as the query and the monocot common CNSs as the subject, and the best hits selected based on the e-value were searched in *S. moellendorffii* which is a lower vascular plant in order to identify the CNSs common to vascular plants. The best hits from this step were searched in *P. patens* to identify any CNSs that could still be remaining as common to all land plants. Finally the best hits from this step were searched in *C. reinhardtii* with the expectation to find any noncoding sequences conserved in the group viridiplantae irrespective of their long divergence time.

2.2.3 Identification of lineage-specific CNSs

Eudicot, monocot, angiosperm, vascular plant lineage-specific CNSs: All the eudicot common CNSs found, were searched in all the outgroups used in the study (all the monocot species, *S. moellendorffii, P. patens and C. reinhardtii*) The CNSs that are common to eudicots and not found in any of the outgroups were designated as eudicot-specific CNSs. Similarly, in order to identify monocot specific CNSs, the monocot common CNSs were searched in the following outgroups, all the eudicots, *S. moellendorffii*, *P. patens* and *C.reinhardtii* used in the study. The angiosperm specific, vascular plant specific and plant specific CNSs were identified the same way by searching against their outgroups. The flowchart for the analysis is depicted in Figure 2.1.

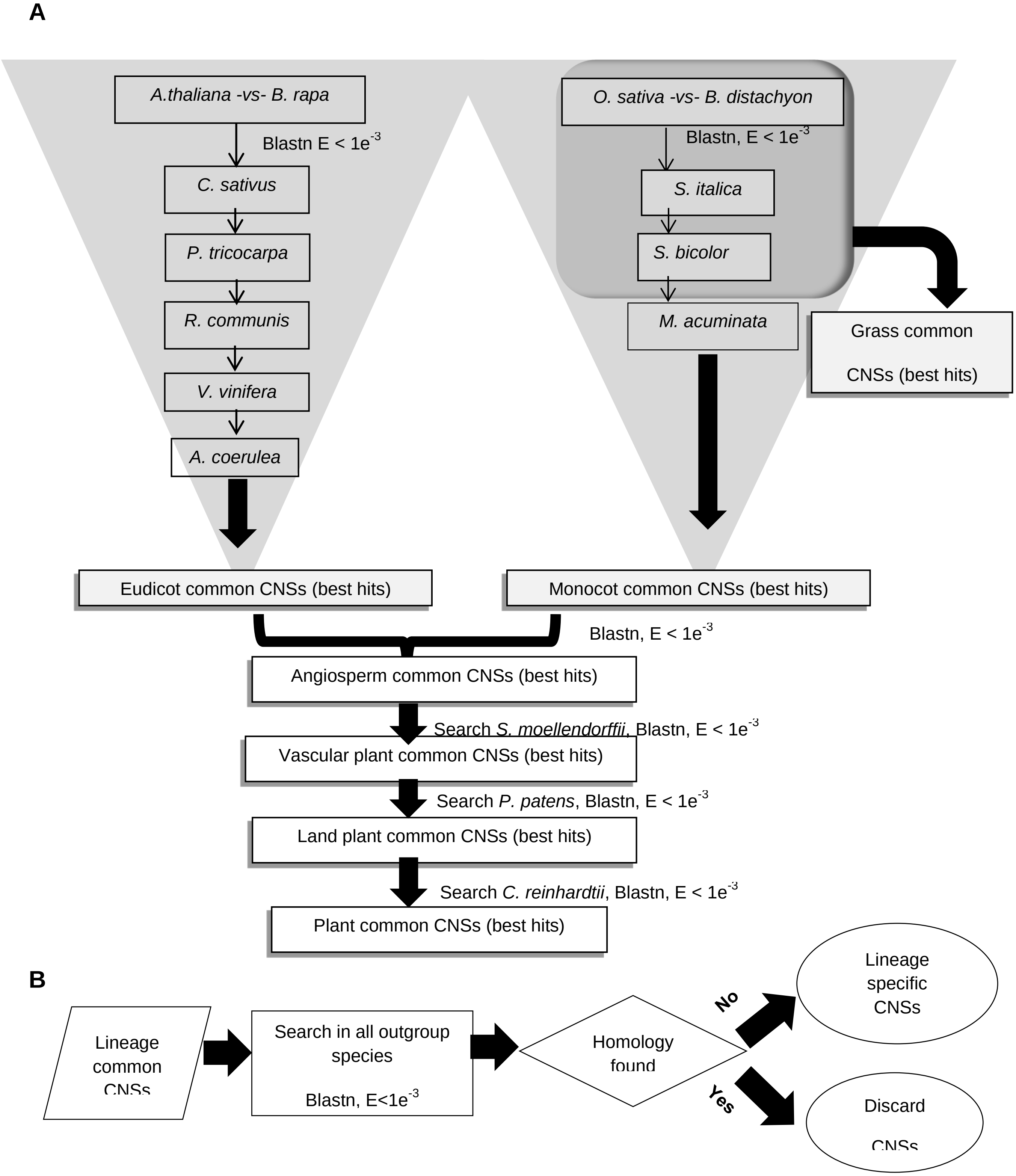

A
A.thaliana -vs- B. rapa
Blastn E < 1e-3
C. sativus
P. tricocarpa
R. communis
V. vinifera
A. coerulea
O. sativa -vs- B. distachyon
Blastn, E < 1e-3
S. italica
S. bicolor
M. acuminata
Grass common CNSs (best hits)
Eudicot common CNSs (best hits)
Monocot common CNSs (best hits)
Blastn, E < 1e-3
Angiosperm common CNSs (best hits)
Search S. moellendorffii, Blastn, E < 1e-3
Vascular plant common CNSs (best hits)
Search P. patens, Blastn, E < 1e-3
Land plant common CNSs (best hits)
Search C. reinhardtii, Blastn, E < 1e-3
Plant common CNSs (best hits)
B
Lineage common CNSs
Search in all outgroup species
Blastn, E<1e-3
Homology found
No
Yes
Lineage specific CNSs
Discard CNSs

Figure 2.1 - The flow charts of the lineage specific CNS determination. (A) Flow chart for lineage common CNS determination. (B) The Flow chart for lineage specific CNS determination.

Lineage specific loss of CNSs: One main reason for the differences in abundance of lineage-specific CNSs is partially due to retention or loss of ancestral CNSs. Consideration of ancestral CNSs give a comprehensive outline of the dynamics of the retention or loss of CNSs. To study the loss of ancestral CNSs I considered *C. reinhardtii* as the basal species for all land plants for this analysis. I conducted independent homology searches for all the in-group species, namely *A. thaliana, B. rapa, R. communis, P.tricocarpa, C. sativus, V. vinifera, A. coerulea, S. italica, S. bicolor, B. distachyon, O. sativa japonica, M. acuminata, S. moellendorffii* and *P. patens* with *C. reinhardtii* as the query genome. Hits overlapping with any genes were filtered out and then a superset of all the ancestral CNSs found in all in-group species was made by merging overlapping hits. This superset represents the aggregate of ancestral plant CNSs that are still found in *C. reinhardtii.* Based on this set of CNSs, the ancestral CNSs lost in each branch were found.

2.2.4 Identification of CNSs for all pairs of species

I conducted an analysis to determine CNSs that are present in all pairs of species and their common ancestors to provide a comprehensive view on presence of CNSs in each pair of species. In total 105 searches were performed for this analysis between different pairs of species. The cut off e-value for the search was 0.001. Only the alignments without any overlap with a coding region for both query and subject were considered. The Schematic representation of this analysis is depicted in Figure 2.

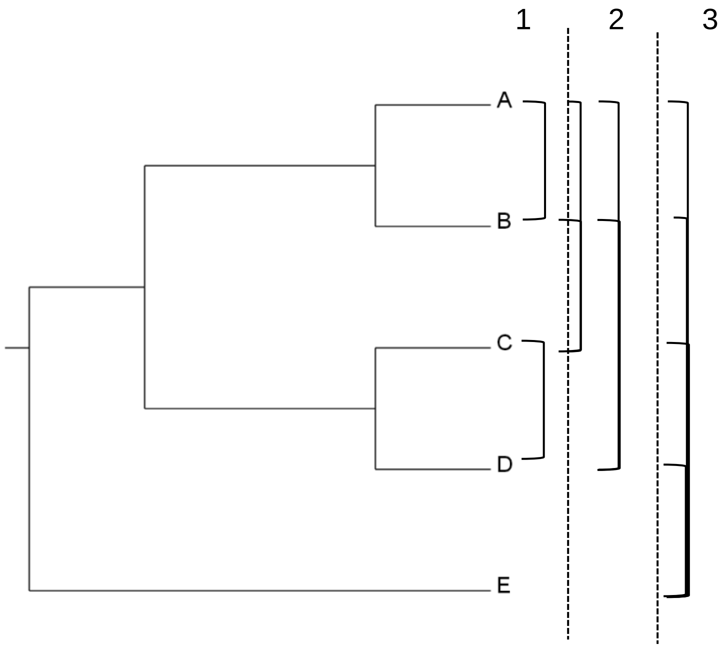


Figure 2.2- Example schematic diagram for identification of CNSs for all pairs of species.

This schematic example with 5 species (A, B, C, D and E) shows how the pairwise searches are performed to determine the union of CNSs for each pair of species and separate lineages. In the example the searches are performed in three levels (1, 2, and 3). The bars on the right side connecting species represent separate searches. For all the species used in the analysis a total of 105 searches were performed in a similar manner to determine the CNSs present in all pairs.

2.2.5Analysis of protein coding genes

Predicted target genes of CNSs: The gene that lies closest to a particular CNS was considered as the likely target gene. For CNSs that were found inside a gene in intron or UTR, the gene it resides was considered as the likely target gene. The likely target gene is with respect to the reference genomes used in the study. For monocot and grass specific CNSs *O. sativa japonica* was considered as the reference genome and for eudicot specific CNSs *A. thaliana* was the reference genome. These genomes have better genome annotation and quality, therefore were considered as the reference.

Identification of lineage specific genes or orphan genes: To establish a preliminary understanding of any correlation between lineage specific CNSs and lineage specific genes, I determined the numbers of lineage specific genes for eudicots, monocots and grasses. I considered protein coding gene sequences of all eudicot and monocot species used in the analysis to run blastp searches. The cut off e-value used was 0.00001 following Yang et al. (2013). The blastp searches were performed as in CNS search in a step wise manner. The lineage specific genes are defined as genes found in all the in group species but absent in all the outgroup species. It has to be noted that this analysis solely depends on the annotated protein coding sequences and that I might have missed on some unannotated genes.

Gene enrichment analysis for the likely target genes: In order to identify the functional groups for the likely target genes the gene enrichment analysis was carried out for grass and monocot specific CNSs using The Database for Annotation, Visualization and Integrated discovery (DAVID) by Huang et al. (2009).

2.2.6 Characterization of the CNSs

A+T content in the flanking regions and the inside of CNSs: Another analysis to characterize CNSs was done by exploring the A+T content in 1000bp flanking regions and the center (20bp) of grass and monocot specific CNSs by a moving window analysis (10bp window with 1 base step size). CNSs with flanking regions that ran into coding regions were removed from the analysis, altogether 4993 grass specific CNSs and 188 monocot specific CNSs were considered for this analysis. The statistical significance was assessed by t-test.

Nucleosome occupancy probability: Kaplan et al. (2009) built a probabilistic model of sequence preferences of nucleosome regions. This model considers the dinucleotide signals along with specific pentamer sequences that are favored or disfavored in known nucleosome sequences to produce a score for each sequence under study. I downloaded their program from **http://genie.weizmann.ac.il/software/nucleo_prediction.html**.

Nucleosome occupancy probabilities for grass and monocot specific CNSs were computed by considering a 4000 base region to each side starting from the center of the CNSs by using this probabilistic model. The average nucleosome occupancy probability was then computed for both sides of each nucleotide site (in total of 8000 sites) along the length of sequences. The same analysis was carried out for a random sample with same AT content as the CNSs (random sequences to have the same length as the CNSs) and also for a random sample with no specific AT preference. These random samples contained the same number of sequences as the CNSs and also same lengths with additional extending flanking regions. All the sequences were extracted from the noncoding regions of the rice genome. The average occupancy probability was calculated for all the 8000 sites for all random sequences. Statistical significance was determined by using t-test.

CNSs and recombination hot spots: The eudicot specific CNSs were searched against recombination hot spot data for *A. thaliana* published by Horton et al. (2012).

Methylation marks on eudicot specific CNSs: Further methylation marks for eudicot specific CNSs were determined by using bisulphite sequencing data for *A. thaliana* published by Cokus et al. (2008) and is available via **http://epigenomics.mcdb.ucla.edu/BS-Seq/**.Twenty seven random samples of 27 sequences in each with the same lengths as eudicot specific CNSs were extracted from the noncoding regions of *A. thaliana* and were searched for methylation marks. The eudicot specific CNSs and the random samples were compared with two proportion z-test at 95% confidence level.

Phylogenetic tree reconstruction with CNSs: The multiple sequence alignments were constructed for the lineage specific CNSs. The aligned multiple sequences were concatenated and the neighbor-trees (Saitou and Nei 1987) for eudicots, grasses, monocots and angiosperms were constructed with MEGA version 5 (Tamura et al. 2011).

2.3 Results

2.3.1 Lineage specific CNSs

I identified 27 eudicot, 6536 grass, 204 monocot, 19 angiosperm, 2 vascular plant specific CNSs (Figure 2.3) and these lineage specific CNSs are likely to have originated in their respective common ancestors. A large number of grass-specific CNSs were observed and as a whole monocots showed more lineage specific CNSs than eudicots. The average lengths of lineage specific CNSs are in the range of 35-60bp except for grass-specific CNSs whose average CNS length was 140bp (Table 2.1). Length distributions of four lineage-specific CNSs are shown in Figure 5. Although most of lineage specific CNSs are shorter than 100bp, 3306 grass-specific CNSs and 14 monocot-specific were longer than 100bp, and the longest grass-specific CNS was 1517bp (Table 2.1). The minimum length for CNSs spans from 16 to 46bp for all lineage specific CNSs. The average percentage identity for all the lineage specific CNSs was found to be more than 80% sequence similarity.

The average percentage identity for each length groupings for CNSs provided in Appendix A1shows that the shorter CNSs have higher percentage identity leading up to >90% and have a higher conservation level while the longer CNSs tend to have lower conservation level. It should be noted that distinction between one long CNS with varying degrees of conservation and a cluster of short CNSs separated with short non conserved regions is not easy. Therefore, a change in the definition of CNSs by varying thresholds, the CNS length distribution may change.

The difference in the number of CNSs may be due to evolutionary rate differences of the lineages. In order to get an understanding if evolutionary rate could contribute to the differences, I calculated the synonymous substitutions (Ds) between the reference genome and the most basal species inside the lineage. Eudicots have a very high saturated Ds value of 2.4363, while monocots and grasses have lower Ds values of 1.5118 and 0.6304, respectively. Therefore evolutionary rate could possibly be one contributing factor for the heterogeneity in number of CNSs.

Can the short divergence time be the reason for the high number of grass specific CNSs? To address this issue I selected pairs of species from eudicots and grasses with approximately the same divergence time: *O. sativa* and *S. bicolor* - divergence time of 60-70 mya (Woolfe et al. 1989), *C. papaya* and *A. thaliana* - divergence time about 70mya (Woodhouse et al. 2011). I then determined the number of CNSs for each pair. The eudicot pair had 1324 CNSs whereas the grass species pair had 16,029 CNSs. Even with the approximately similar divergence times, the pattern of CNSs remained the same as for the lineage specific CNSs. I also determined if the number of species used in the analysis for eudicots be responsible for the difference in the number of CNSs between eudicots and monocots. I randomly selected 4 eudicot species (*B. rapa, R. communis, C. sativus, V. vinifera*) and determined the number of lineage specific CNSs for them. The number of their lineage specific CNSs was 118, which is still much less than the number of CNSs I obtained for grasses. I further selected 5 eudicot species (*B. rapa, P. tricocarpa, R. communis, C. sativus, V. vinifera*) which has a total of 120 million year divergence times, and determined the number of CNSs for them, in comparison with the number of CNSs obtained for the 5 monocot species in the study. The number of lineage specific CNSs for those 5 eudicot species was 69, whereas the number of lineage specific CNSs for 5 monocot species was 204. It should be mentioned that the total divergence time of these 5 monocot species is 115 million years. Even with the same number and similar divergence times of eudicot and monocots, the number of eudicot specific CNSs remain much less than monocot specific CNSs. Therefore, the difference in number of CNSs is not due to the number of compared species.

It is important to note that this analysis started with an initial pair of species and try to determine lineage specific CNSs for each group that are present only in all members of that lineage. But if I consider lineage common CNSs (CNSs that are commonly found in all members of a group but also might be found in out-group species; see Appendix A3), it is clear that the lineage specific CNSs are much less and represents a small fraction of elements that is likely to be functional in common ancestor. And further with this analysis to determine CNSs that are present in all pairs of species and their common ancestors (Figure 2.5) I found that the CNSs in each branch is higher (in this analyses I considered the union of CNSs) than lineage common or lineage specific CNSs. But many of these may have gone through independent losses inside a lineage, therefore would not fall under my criteria for determination of lineage specific CNSs.

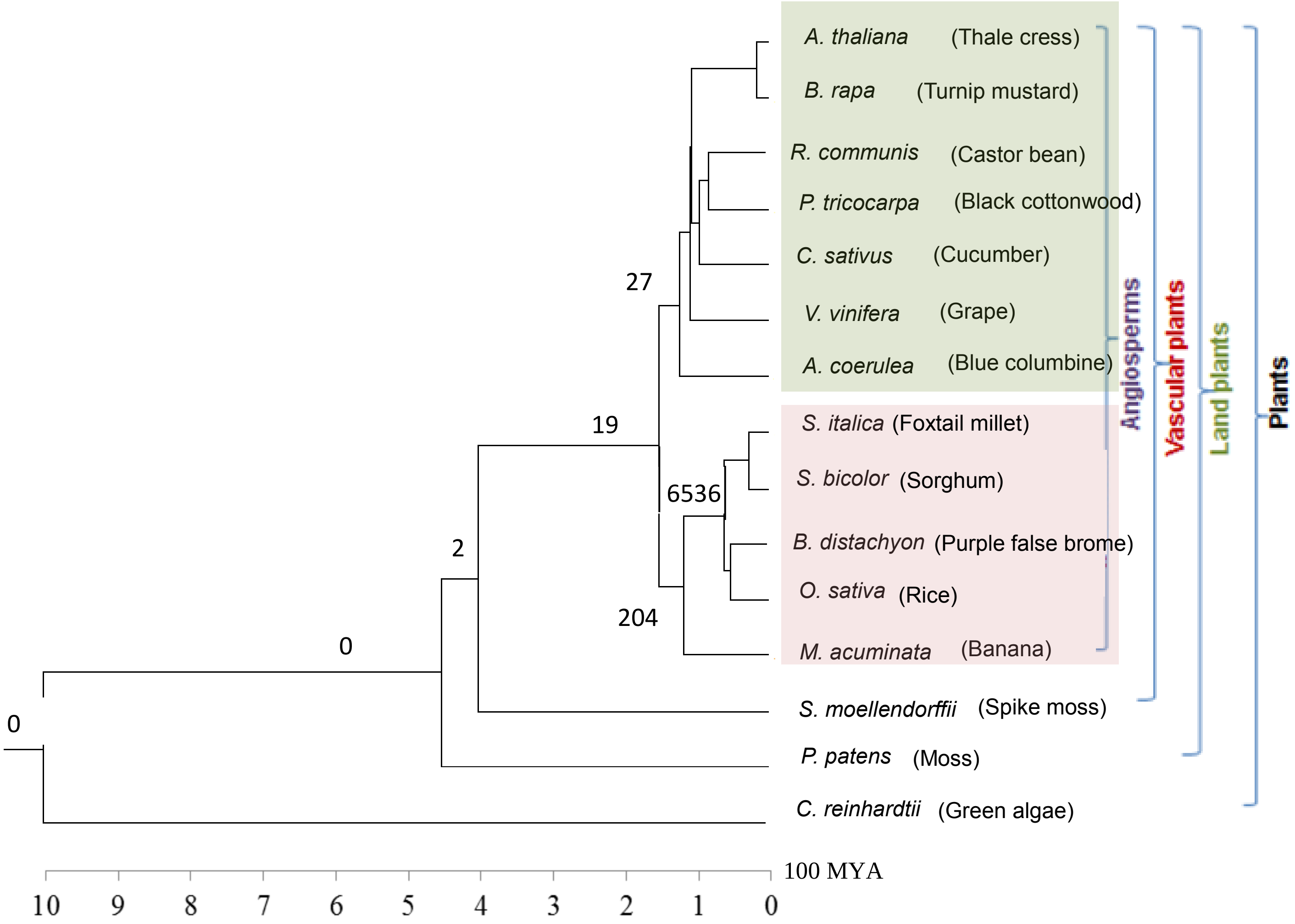

A. thaliana (Thale cress)
B. rapa (Turnip mustard)
R. communis (Castor bean)
P. tricocarpa (Black cottonwood)
C. sativus (Cucumber)
V. vinifera (Grape)
A. coerulea (Blue columbine)
S. italica (Foxtail millet)
S. bicolor (Sorghum)
B. distachyon (Purple false brome)
O. sativa (Rice)
M. acuminata (Banana)
S. moellendorffii (Spike moss)
P. patens (Moss)
C. reinhardtii (Green algae)
Angiosperms
Vascular plants
Land plants
Plants
27
19
6536
2
204
0
0
100 MYA
10
9
8
7
6
5
4
3
2
1
0

Figure 2.3 - Phylogenetic tree with the number of lineage specific CNSs.

The numbers on each branch represents the number of lineage specific CNSs found in the study. The main plant groups considered in the study are depicted on the right. The phylogenetic tree was constructed with verified divergence times taken from Anderson et al. (2005), D'Hont et al. (2012), Banks et al. (2011), Heckman et al. (2001), Rensing et al. (2008). Eudicots and monocots are shaded in green and light pink respectively.

	Eudicot Specific	**Monocot specific**	**Grass specific**	**Angiosperm specific**	**Vascular plant specific**
Number of CNSs	27	204	6536	19	2
Minimum length (bp)	22	23	23	16	46
Maximum length (bp)	63	186	1517	95	50
Average length (bp)	38.5	58.5	140.7	42.8	48.0
Average pid (%)	89.8	84.3	80.25	87.5	82.0
CNSs ≥ 100bp	0	14	3306	0	0

Table 2.1 - Summary of lineage specific CNSs

2.3.2 Lineage specific CNSs and lineage specific genes

If the evolutionary rate had been a major contributing factor for the differences in the numbers of CNSs, the lineage specific genes should also follow the same pattern as CNSs (unless the lineage specific genes are under higher selective constraint). To investigate this scenario I determined the numbers of lineage specific genes and identified 2439, 444, 113 eudicot, grass and monocot lineage specific genes consecutively. The number of eudicot lineage specific genes is much higher than grass and monocot lineage specific genes, this is quite the opposite scenario to what was observed for lineage specific CNSs. This observation gives light that apart from evolutionary rate there should be additional factors that contribute to the differences in CNSs. If I consider individual lineages, dicots evolved more lineage specific genes whereas monocots and grass common ancestors gave rise to more lineage specific CNSs.

I also found that these lineage specific genes are predominantly plant defense related.

2.3.3 Lineage specific loss of ancestral CNSs

One factor for the differences in the number of CNSs could be the loss or retention of ancestral CNSs by either rapid divergence or complete deletion (Hiller et al. 2012). Assuming an unbiased parallel loss between the reference genome (*C. reinhardtii*) and the other plant species, this analysis would provide an overall pattern on the rate of loss of CNSs of all plant species used.

I found that *C. reinhardtii* has 4355 CNSs conserved in one or more of the species used. Based on this result the number of CNSs lost in each branch was calculated. The result shows that eudicot common ancestor has lost twice as much CNSs than the monocot or grass common ancestors which indirectly answers the heterogeneity of the lineage specific CNSs (Figure 2.4, see

Appendix A3 for ancestral CNSs found between *C. reinhardtii* and other species). Of all the eudicots, *V. vinifera* and *A. coerulea* have lost the most number of CNSs independently in their respective lineages. It was observed that within grass lineage CNS loss has been lowered in *O. sativa japonica* and *S. bicolor.*

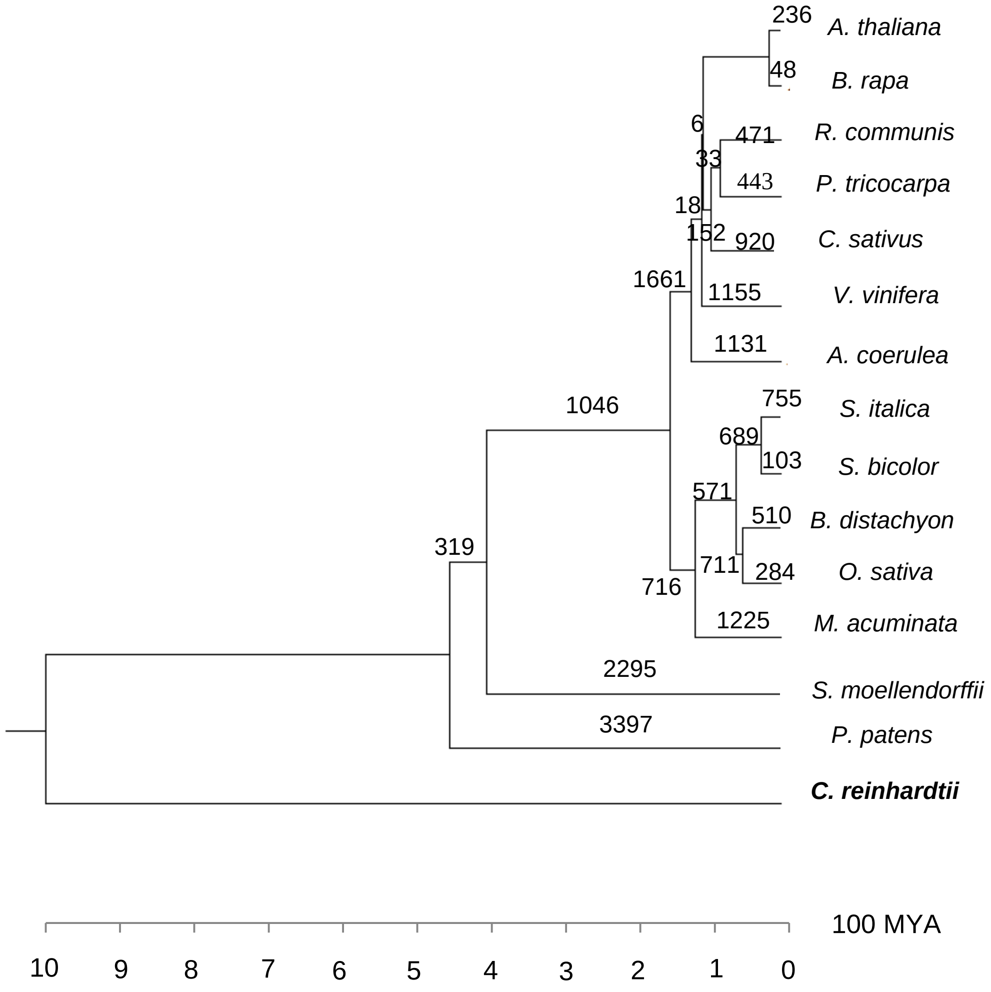


Figure 2.4 – The lineage specific loss of ancestral CNSs. The values on branches represent the number of CNSs lost on that specific branch. The reference genome used for this analysis is *C. reinhardtii.*

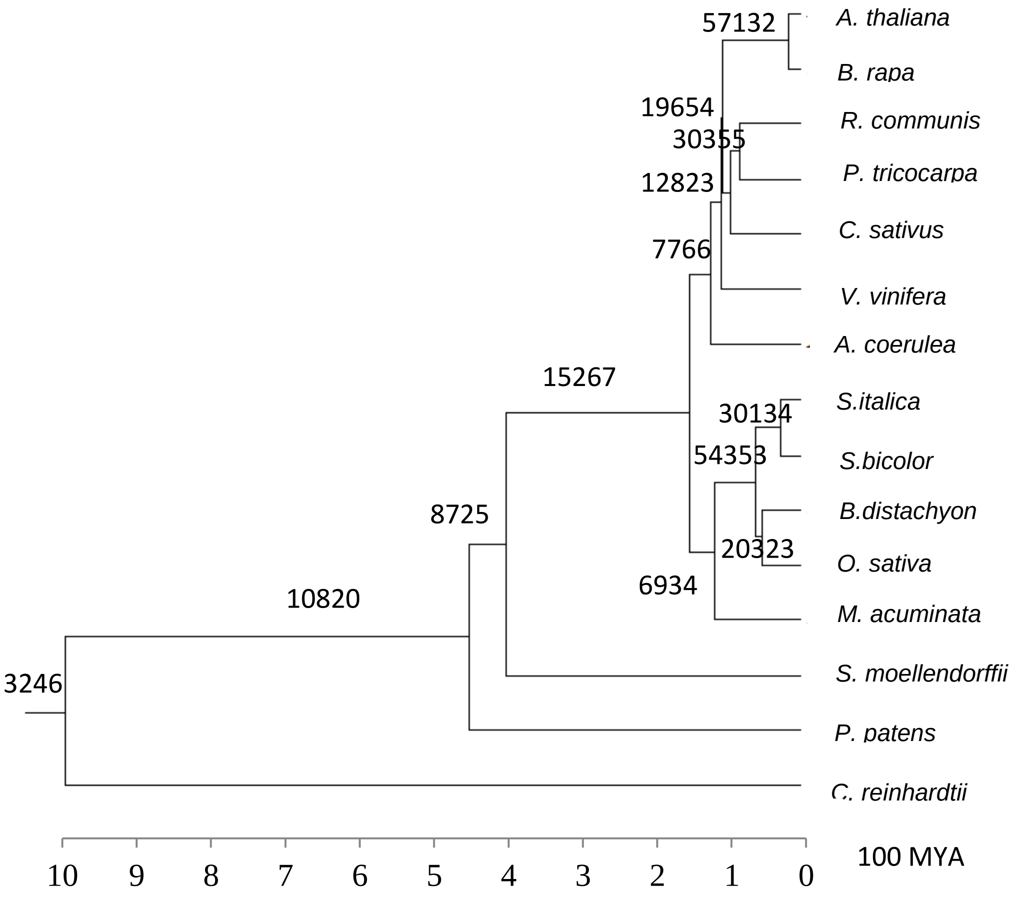


Figure 2.5 - Number of CNSs found from all pairwise searches. CNSs between all pairs of species were determined to have an overall comprehensive view on gain of noncoding conservation. These pairwise analyses consider the union of all CNSs. The number on each node reflects the gain of CNSs obtained via pairwise searches. These CNSs are common to each group of species and therefore are likely to be found in out-group species.

2.3.4 The genomic locations of identified lineage-specific CNSs

I examined the genomic locations of CNSs to see if they are in UTR, intron, or in intergenic regions. Table 2.2 shows the frequencies of these three locations of the grass and monocot specific CNSs identified with respect to the genome of *O. sativa japonica* as the reference. The grass and monocot specific CNSs located in the intergenic regions (53.7% and 55%, respectively) are significantly less (P-value for grass and monocots are 2.6E-161 and 0.000236, respectively) than the expectation from the genomic coverage, 70% for the reference rice genome. In contrast, the number of grass specific and monocot specific CNSs located in the UTR (25% and 22%, respectively) are significantly higher than the expectation from the genomic coverage (6%).

Although the eudicot and angiosperm specific CNSs are less in number, the CNSs located in the UTR regions followed a similar pattern having a significantly higher representation than the genomic coverage for the UTR regions in the reference genome (Table 2.3; *A. thaliana* was used as the reference genome).The result implies a stronger constraint on the CNSs located in the UTR regions.

	Rice noncoding genome composition (%)	**Grass specific**	**Monocot specific**
Intergenic	70.0	53.7 (3503)	54.9 (112)
Intron	24.2	21.0 (1374)	22.1 (45)
UTR	5.8	25.3 (1658)	23.0 (47)
3’UTR	3.4	19.2 (1259)	11.3 (23)
5’ UTR	2.4	6.1 (399)	11.7 (24)

Table 2.2 - Genomic locations of the grass and monocot lineage specific CNSs.

Genomic locations of grass and monocot specific CNSs with respect to the reference genome *O. sativa japonica* are provided as a percentage in third and fourth columns. Rough percentage estimations of the intergenic, intron and UTR regions for the reference genome are provided under rice noncoding genome composition in the second column.

Note - The exact number of CNSs in each region is given in parentheses.

	Arabidopsis noncoding genome Composition (%)	**Eudicot specific**	**Angiosperm specific**
Intergenic	56.8	63.0 (17)	47.4 (9)
Intron	35.0	7.4 (2)	31.6 (6)
UTR	8.2	29.6 (8)	21.0 (4)
3’ UTR	4.0	14.8 (4)	10.5 (2)
5’ UTR	4.2	14.8 (4)	10.5 (2)

Table 2.3 - Genomic locations of the eudicot and angiosperm lineage specific CNSs.

Genomic locations of eudicot and angiosperm specific CNSs with respect to the reference genome *A. thaliana* are provided as a percentage in third and fourth columns. Rough percentage estimations of the intergenic, intron and UTR regions for the reference genome are provided under Arabidopsis genome composition in the second column.

Note - The exact number of CNSs in each region is given in parentheses.

2.3.5 Distribution of the CNSs in chromosomes

I next examined chromosomal distributions of lineage-specific CNSs. Grass and monocot specific CNSs were found distributed among all the 12 chromosomes with respect to *O. sativa japonica* – the reference genome (see Appendix A2 A and B for grass-specific and monocot-specific CNSs, respectively). However, the numbers of CNSs on each chromosome varied and further, intra-chromosomal distributions of the CNSs were observed to be uneven as well. One example is chromosome 10 with CNSs concentrated in several areas in reference genome for grass specific CNSs (encircled in black – 3 clear clusters). This indicates that rather than being distributed randomly, CNSs tend to exist in clusters. A similar pattern was observed for monocot specific CNSs. Some example likely target genes related to CNSs in cluster 3 are, FAD dependent oxidoreductase domain containing protein, aldo/keto reductase family protein, transcription factor BTF3, double-stranded RNA binding motif containing protein, cytokinin dehydrogenase precursor etc. Functions of some of these genes are documented to be important in regulation of genes. Even though it has been reported that much of plant BTF3 functions still remain obscure, previous researches suggest that BTF3 is associated with HR (hypersensitive-mediated) mediated cell death and involved in biotic stress regulation in the nucleus (Huh et al. 2012), and also double stranded RNA binding protein plays a vital role in viral defense and development by regulation of cellular signaling events and gene expression (Waterhouse et al. 2001).

2.3.6 Predicted target genes of CNSs and their enrichment analysis

I considered the genes closest to the CNSs as the likely target gene based on the premise that the regulatory elements reside close to the gene it regulates.

The gene enrichment analysis for the predicted likely target genes of the grass and monocot-specific CNSs indicate that these genes are predominantly involved in the regulation of transcription and DNA binding. Table 2.4 and 2.5 shows the top twenty groups in which genes are enriched in grass specific and monocot specific CNSs. The P-values of the result suggest that the groupings are highly statistically significant. However, the gene ontology groupings obtained for a random sample of 6536 genes (*O. sativa japonica*) also showed groupings with statistically significant P-values up to P = 1.7 X 10^{-8} (Appendix 4-A). If we normalize the P-values acquired for the actual data set shown in Table 3A by dividing this randomly obtained P value, still all genes are highly statistically significant. As for gene enrichment for monocot-specific CNSs, now only the top seven groups become statistically significant after the same normalization. One important feature shared between grass-specific CNSs and monocot-specific CNSs is that they are predominantly enriched with genes related to transcription regulation. GO groups involved in enzymatic activity and various catabolic processes were significantly under represented, meaning that CNSs are less associated with such genes (Appendix 4-B). It has to be noted that functional classification is not an absolute decision making procedure regarding the likely target genes of CNSs but rather indeed an exploratory method to identify the possible likely functions of flanking genes.

My result agrees with several animal CNS studies reported so far, stating that CNSs are found near genes involved in regulation of transcription and development (Hardison2000; Sandelin et al. 2004; Shin et al. 2005; Venkatesh et al. 2006; Janes et al 2010). Kritsas et al. (2012) stated their gene ontology analysis showed that genes associated with *A. thaliana* ultra-conserved like elements (ULE) were involved in development and are likely to be developmentally regulated. Some of the eudicot specific CNSs in this analysis were found close to transcription factors such as Kanadi2 which regulates embryo development (Heyndrickx et al. 2012), transcription factor jumonji which regulates circadian clock (Lu et al. 2010) and *DELLA* protein RGL1 which is a negative regulator of plant hormone gibberellin (Tyler et al. 2004). Seven likely target genes of eudicot specific CNSs were found to be enzyme encoding. The two vascular plant CNSs were found in the intron region of the genes, suppressor of auxin resistance 3and pre-mRNA splicing factor 38A**.** Plants deficient in suppressor of auxin resistance show pleiotropic growth defects such as shorter primary root, fewer lateral root. And also it has been found that flowering occurs earlier than the wild type, flowers are smaller and less fit during life cycle (Parry et al. 2006). Further Parry et al. (2006) reported that this protein plays an important role in hormonal regulation and development in plants. Whereas pre-mRNA splicing factors and regulators are found to be quite important in splice site selection (Reddy et al. 2011) to ensure accurate transcript formation, including instances such as alternative splicing (Mabon and Misteli 2005).

I examined overrepresented motifs of lineage specific CNSs using MEME (Bailey et al. 2009). Results (Appendix A5) suggest that some motifs are related to genes with enriched GO terms (Buske et al. 2009) related to DNA binding and transcription factor activity.

Functional group	Percentage of genes in the group	P-value
Functions related to nucleus	61.9	0.0E0
Regulation of transcription	70.5	0.0E0
DNA-binding	51.5	9.6E-309
Transcription	46.1	4.5E-278
Transcription regulator activity	69.1	5.7E-272
Transcription factor activity	65.6	8.6E-269
Regulation of RNA metabolic processes	41.7	4.6E-171
Zinc-finger related	23.3	3.4E-106
Activator	14.6	1.3E-86
Sequence-specific DNA binding	21.9	2.0E-83
Zinc ion binding	36.6	2.8E-72
Metal ion binding	25.1	1.1E-59
Homeodomain related	12.7	2.2E-54
Response to organic substance	24.5	1.5E-51
Myb-type HTH DNA-binding domain	10.0	5.6E-46
Cellular response to hormone stimulus	13.9	1.4E-43
Hormone mediated signaling	13.9	1.4E-43
Myb, DNA binding	10.3	1.8E-43
Pathogenesis related transcription factor and ERF, DNA binding	7.7	3.9E-39
Transition metal ion binding	39.3	1.2E-38

Table 2.4 - Gene enrichment analysis for the likely target genes of grass specific CNSs.

Functional group	Percentage of genes in the group	P-value
Transcription factor activity	92.3	2.4E-51
Transcription regulator activity	92.3	5.8E-48
Regulation of transcription	90.8	1.3E-46
Functions related to nucleus	72.3	5.8E-45
DNA binding	93.8	1.2E-43
Sequence specific DNA binding	35.4	1.8E-17
Zinc-finger related	26.2	3.0E-12
Homeodomain related	16.9	3.6E-8
Basic-leucine zipper transcription factor	10.8	1.3E-7
Metal binding	27.7	1.3E-7
Activator	13.8	2.1E-7
Transcription factor, GATA, plant	13.8	6.2E-5
Myb-like DNA binding region	9.2	5.5E-6
No apical meristem protein	9.2	2.0E-5
Heat shock factor (HSF) type, DNA binding	6.2	5.1E-5
Homeobox conserved site	7.7	1.4E-4
Anther development	6.2	1.6E-4
Androecium development	6.2	8.2E-4
Stamen development	6.2	8.2E-4
Post-embryonic development	18.5	8.7E-4

Table 2.5 - Gene enrichment analysis for the likely target genes of monocot specific CNSs.

2.3.7 Synteny of target genes

I found all target genes of dicot specific CNSs to have orthologs in *A. coerulea* genome (the most basal dicot species used in the analysis) in a 5kb range from the *A. coerulea* CNSs (subject genome in this context) implying that CNSs are conserved along with the syntenic positions even in species with long divergence times. Similarly for grass and monocot specific CNSs, 5770/6536 and 180/204 target genes, respectively, had orthologs located in same syntenic positions along with the CNSs. This implies that CNSs and their target genes have been conserved as one block during evolution. The conservation level of the orthologous target genes was found to be statistically significant (t-test) when compared with random samples of orthologous genes (p-values 1.05E-11, 5.87E-05, and 0.00 respectively for eudicot, monocot and grass orthologous genes).

2.3.8A+T content in the flanking regions of CNSs

To determine if there are specific characteristics in CNSs and their flanking regions, the A+T content in the flanking regions and the inside of the CNSs were determined. One thousand bp of flanking regions in the 5’ and 3’ directions and 20bp from the middle of the CNS were considered. A moving window analysis (10bp window and 1bp step size) was used to determine the A+T frequency around the margin and the inside of the CNSs for grass and monocot specific CNSs. A decline in the A+T content was observed near the start of the grass specific CNSs as shown in Figure 2.6. A very similar pattern was observed for monocot specific CNSs (Figure 2.8).The average A+T content of the flanking regions is 56%, which is same as the genomic A+T content of rice and the average A+T content of the CNSs is about 54%. T-test showed that there is a statistically significant difference between the flanking regions and the CNSs with respect to the A+T content (p-value < 0.0005). Interestingly, the drop of A+T content at the flanking regions has also been observed in animal CNSs (Walter et al. 2005; Vavouri et al. 2007). Kritsas et al. (2012) also reported the similar scenario being present in plant CNSs, in which they considered *A. thaliana, V. vinifera* and *O. sativa, B. distachyon* to look for ultra-conserved like elements (ULE) in plants. This finding agrees with previous literature. It has to be noted that the feature of the drop of A+T frequency near the border of CNSs seems to be conserved between animals and plants similarly as was reported by Kritsas et al. (2012). A similar tendency of A+T drop was observed for eudicot specific CNSs with a significant difference between the flanking regions and the inside of CNSs (95% confidence p-value < 0.05).The lineage specific CNSs I identified are GC rich compared to the genomic average. Babarinde and Saitou (2013) reported that animal CNSs in their analysis are GC poor, this finding stands opposing to my observation for plant CNSs.

2.3.9 Prediction of nucleosome positioning

The A+T content can affect the DNA topology and nucleosome positioning. Jansen and Verstrepen (2011) showed that A+T rich sequences in *S. cerevisiae* have a very low tendency to form nucleosomes. Nucleosome positioning pattern facilitates the access of transcription factors to their target sites and it plays a pivotal role in determining the transcription level (Bai and Morozov 2010). Therefore the drop of A+T around the flanking regions of CNSs may be contributing towards nucleosome formation. I determined the nucleosome occupancy probability for the grass and monocot specific CNSs and their flanking regions with nucleosome prediction software by Kaplan et al. (2009). From the center of each CNS, a 4000bp region in 5’ and 3’ directions were considered. Two additional control samples were selected from the noncoding regions of the rice genome. A clear peak can be observed (Figure 2.7) in the averaged nucleosome occupancy directly overlapping with the center and surrounding regions of the CNSs indicating a possibility of nucleosome positioning in the CNS regions. Even though a slight increase in the nucleosome occupancy probability can be seen in the random sample with the same AT content as the CNSs, the nucleosome occupancy probability of the CNSs is highly statistically significant compared to the random sample with same AT content (p-value 2.80E-10). A very similar pattern was observed for the monocot specific CNSs (Figure 2.9) again indicating a clear nucleosome positioning in and around CNSs. Similarly a high statistical significance (p-value 0) was observed between the CNSs and the random sample with same AT content. The random samples for both grass and monocot specific CNSs with no AT preference stayed roughly constant with regards to the nucleosome occupancy probability throughout the length. One interesting feature observed for the UTR CNSs was the higher nucleosome occupancy obtained for 5’ UTR regions in comparison with 3’ UTR CNSs. Nucleosome occupancy for 3’ UTR were almost similar to random expectation (Appendix A6).As reported by Bai and Morozov (2010), Jiang and Pugh (2009) nucleosome positioning is related to gene regulation and this result gives thorough evidence that CNSs can be involved in regulation of their target genes. Also Tillo et al. (2010) has reported that there is high nucleosome occupancy at regulatory sequences in the human genome. Also Baxter et al. (2012) observed a similar pattern of nucleosome positioning for CNSs in four eudicot plants.

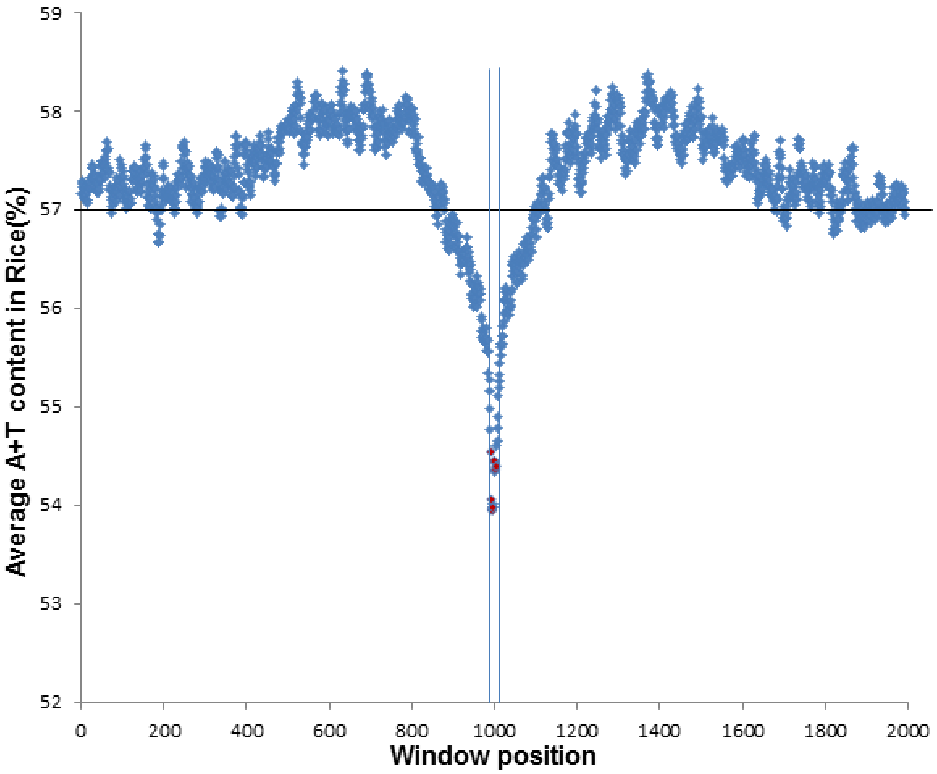


Figure 2.6 -Distribution of A+T content in the flanking regions and within CNSs (for grass specific CNSs).

Black line – average A+T content in rice genome. Red dots – A+T content inside CNSs (20bp from the center of each CNS was considered as mentioned in the methodology) acquired via moving window analysis. Blue vertical lines give the borders of 5' and 3' flanking regions around the CNS.

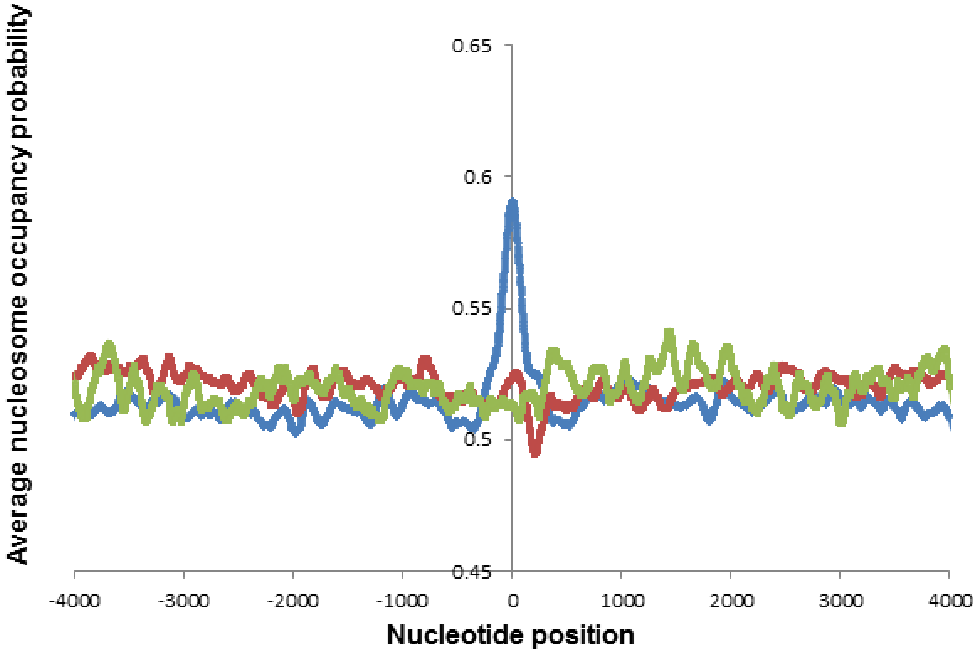


Figure 2.7 - Nucleosome occupancy probability for grass specific CNSs including flanking regions.

0th nucleotide position represents the center of each CNS and also the center of the random samples. Blue, Red and green graphs respectively show nucleosome occupancy probabilities of the CNSs, random sample with same AT content as CNSs and the random sample without specific AT preference.

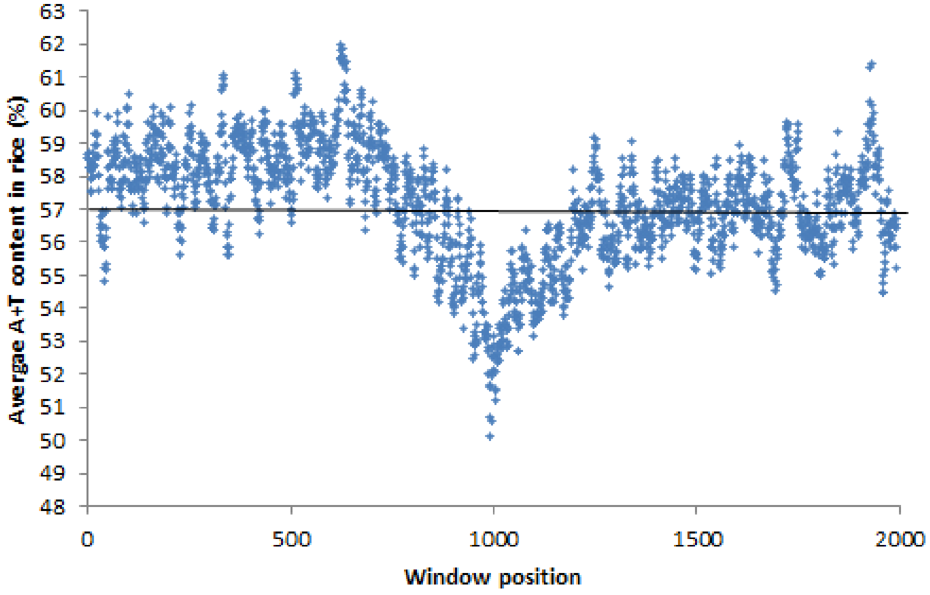


Figure 2.8 - Distribution of A+T content in the flanking regions and within monocot specific CNSs, Horizontal line shows the average A+T content in the rice genome.

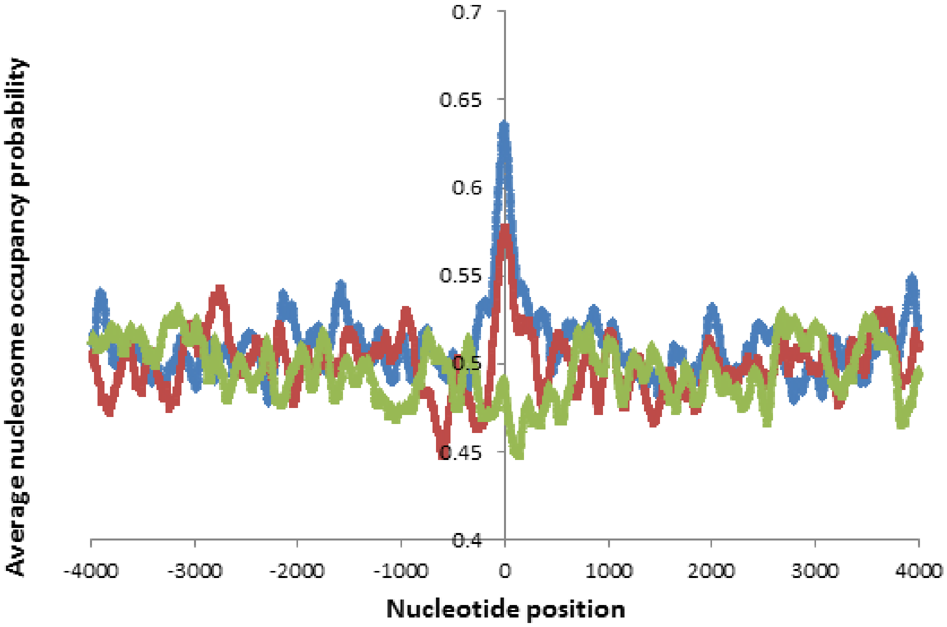


Figure 2.9 - Nucleosome occupancy probability for monocot specific CNSs including flanking regions.

0th nucleotide position represents the center of each CNS and also that of the random samples. Blue, Red and green graphs respectively show nucleosome occupancy probabilities of the CNSs, random sample with same AT content as CNSs and the random sample without specific AT preference.

2.3.10 CNSs are not related with recombination hotspots

It has been found that the A+T content might be related to recombination rates. Spencer et al. (2006) found that recombination hotspots for the human genome are associated with increased GC content or in other words lower A+T content. Guo et al. (2009) found that for rice and human genomes microsatellites with motifs consisting of only A & T such as AT, TA have lower recombination rates. As there was a decline in the A+T content in the flanking regions of the CNSs, I checked if the eudicot specific CNSs overlapped with any recombination hot spots that were found by Horton et al. (2012) for *A. thaliana* genome. I found that none of the CNSs overlapped with recombination hot spots documented in the above mentioned study. An observation similar to this was found for Kritsas et al. (2012) for eudicot ULEs. The decline in the A+T content may not be related to any recombination rate variation, but may be related to an unknown feature that is related to the function of the CNSs.

2.3.11 Methylation in eudicot specific CNSs

Methylation of the cytocine residues is a well observed and understood phenomenon in many organisms including *A. thaliana*. DNA methylation is known to be related to regulation of gene expression and many other numerous cellular processes such as embryonic development, genomic imprinting, and preservation of chromosome stability (Phillips. 2008; Gutierrez-Arcelus et al. 2013; Geiman and Muegge 2010). To see if the eudicot specific CNSs contain any signature of methylation (Kritsas et al. 2012), those CNSs were compared with *A. thaliana* whole genome methylation data obtained by Cokus et al. (2008). Only 2 of the 27 eudicot specific CNSs showed methylation mark in CG, CHH or CHG sequence contexts, in the minus or the plus strand. About 20% of the Arabidopsis genome is methylated including transposable elements and repeats (Bilichak et al. 2012). The methylation signature for the selected random samples provided sufficient evidence that the proportion of methylated CNSs is less than the proportion of methylated sequences in the random sample (z- value for the test statistic is – 2.66, p-value < 0.05). This result implies that the probability of observing methylation in eudicot specific CNSs is less than the probability of observing methylation in the random samples at 95% confidence level according to two proportion z-test. Therefore I conclude that eudicot specific CNSs found in this study are not predominantly modified by DNA methylation.

2.3.12 Phylogenetic tree construction with CNSs

I constructed the phylogenetic trees for each plant group considered in the study by using lineage specific CNSs. The trees were constructed with 1000 bootstrap replication and the model used was maximum composite likelihood method. The neighbor-joining trees (Saitou and Nei 1987) constructed for grass and monocot specific CNSs (Figure 2.10 A and B) exactly comply with monocot and grass phylogenies (Bennetzen et al. 2012) with >80% bootstrap probability. Branching of *S. italica* and *S. bicolor* showed 100% bootstrap value whereas *O. sativa* and *B. distachyon* showed 80% bootstrap value for the tree constructed with monocot specific CNSs. The expectation of phylogenetic tree construction with CNSs is that, if the CNSs are orthologous it should be possible to reconstruct the species tree with high statistical support.

Even though I constructed similar trees for eudicot and angiosperm CNSs, only some branching showed high bootstrap values and agreed with the known phylogeny. *A. thaliana* and *B. rapa* were always clustered with 100% bootstrap value in both eudicot and angiosperm trees (Appendix A7). Monocots were also clustered together as expected with 100% bootstrap confidence.

The fact that the exact topology could not be achieved with eudicot and angiosperm CNSs does not mean the CNSs are not orthologous, one reason for not being able to achieve the exact known topology could be that the concatenated sequence lengths for eudicot and angiosperm CNSs are too short (which were less than 1500bp) or in other words number of informative sites are less for phylogenetic tree construction. When sequences are short and the divergence among sequences is low it affects the phylogenetic tree construction.

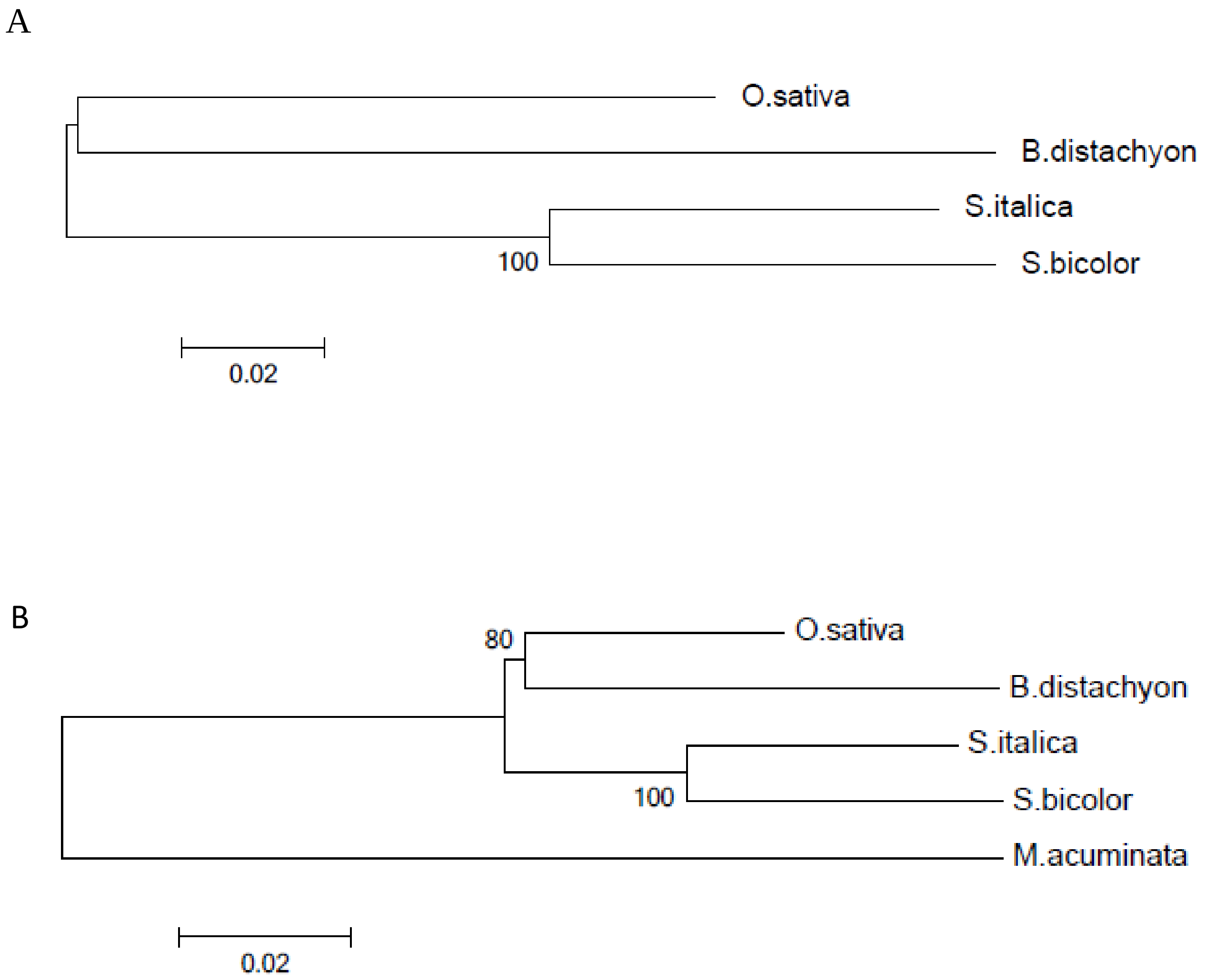


Figure 2.10 - The phylogenetic trees of the CNSs.

The concatenated lineage specific CNSs were used to construct the phylogenetic trees with neighbor-joining method. (A) Phylogenetic tree constructed for grasses with grass specific CNSs (B) Phylogenetic tree constructed for all monocots in the study with monocot specific CNSs.

2.4 Discussion

I identified lineage specific CNSs that originated in their respective common ancestors in this study. These CNSs likely define their lineage specific characters and functions. I observed a large number of CNSs that originated in the grass common ancestor and shared by all the grass species used in the study. One plausible reason for this high number could be the short divergence time of the grass species, but when compared with eudicot species with the same divergence time as the grasses the pattern of CNSs remained the same, in other words eudicots still had much less CNSs than grasses irrespective of the divergence time. This implies that grasses may have developed their own specific regulatory mechanisms to withstand diverse conditions such as biotic, abiotic stresses and to facilitate various other molecular processes. Furthermore, this suggests that grasses might be sharing similar kind of regulation as a group of species that might be contributing to their lineage specific features.

In contrast, eudicots have much smaller number of lineage specific CNSs shared by all the species used in the study. One explanation could be that many of the CNSs that originated in the eudicot-common ancestor have diverged beyond recognition, that they no longer can be detected by homology search. It is likely that the CNSs have degraded over evolutionary time but the binding functions may be conserved even though they cannot be detected by sequence alignment due to binding site turnover. And also whole genome duplication events that occurred during evolution of eudicots can be another plausible reason for the less number of eudicot CNSs (Tang et al. 2008).

Since the number of monocot-specific CNSs is seven folds higher than eudicot-specific CNSs, it is possible that, after the divergence from common ancestor of angiosperm, monocots gained more CNSs to establish monocot specific features, which are still found conserved in monocot species. This could be due to various physiological and morphological complexity differences between monocots and eudicots. The vascular bundle formation and arrangement is more complex in monocots than in eudicots, monocot embryogenesis differs broadly from that of eudicot and the architecture of monocotyledonous embryo is far more complex, complexity in differences in root formation (multiple layers of cortical cells in monocot root while eudicots have one layer of cells; Grunewald et al. 2007) are some of the differences known with respect to complexity. Zimmerman and Werr (2007) reported that even at very early stages of monocot embryogenesis the cell division patterns are variable and unpredictable, further they report that primary root of cereals is formed endogenously deep inside the embryo which is a major difference with the dicots. And also the embryonic axis of the monocots is displaced laterally respect to scutellum in contrast to apical-basal axis of dicots (Zimmerman and Werr 2005). Also it has been reported that the shoot apical meristem differs in structure and function in eudicots and monocots (Sussex 1989; Jurgens 1992; Kerstetter and Hake 1997).Therefore, primarily, complexity could be one reason for monocots to have more CNSs as they require more regulation.

The fact that only 2 specific CNSs were found to be conserved in all the vascular plants suggests that in general, plant CNSs have a high turnover rate and many CNSs originated in vascular plant, land plant and plant common ancestors have diverged beyond recognition.

One interesting feature observed in this analysis is the difference in the numbers of lineage specific CNSs and lineage specific genes. The lineage specific genes showed quite the opposite pattern to CNSs. The eudicots had the least number of CNSs but the highest number of lineage specific genes, whereas grasses had the highest number of CNSs but lineage specific genes were five folds less than that of eudicots. It appears that eudicots gave rise to more lineage specific genes whereas grasses and monocots evolved more CNSs in their respective common ancestors. It would be very interesting to find out what factors actually govern organisms in originating genes and CNSs in their respective common ancestors. In other words which factors determine to have more CNSs or more genes?

The frequency of CNSs in the UTR regions was observed to be higher than the genomic coverage for the UTR regions in the reference genomes, thus shows a stronger selective pressure on the CNSs located in the UTR regions. Most of the UTR CNSs were found in 3'UTR regions for the grass specific CNSs .This finding is consistent with earlier reports of conservation in the 3'UTR regions (Duret et al. 1993; Lipman 1997; Grzybowska et al. 2001, Siepel et al. 2005). In addition, the enriched conservation in UTR was observed for genes in DNA binding proteins (Duret et al. 1993). However, it is likely that 3'UTR conservation found in this study could be involved in post-transcriptional regulatory mechanisms as well directing subcellular localization, transcript stability or translatability. In accordance with this assumption I observed lower nucleosome occupancy probability for the 3' UTR CNSs compared to CNSs in other regions of the genome.

The drop in A+T content near the borders of the CNSs is a feature that is also seen in animals (Walter et al. 2005; Vavouri et al. 2007). Therefore it is possible that this orientation of nucleotides such as the drop of the A+T content near the boundaries is an important feature for the CNS function. This shows a required functional property of CNSs, even though the reason for this CNS layout conservation between animals and plants is not yet known. But one candidate explanation lies with A+T content and the nucleosome formation.

With the nucleosome positioning analysis it was observed that the CNSs tested, showed high nucleosome occupancy probability in and around the CNSs implying CNSs may have a higher probability to form nucleosomes. The finding by Bai and Morozov (2010), Jiang and Pugh (2009) stating nucleosome positioning is related to gene regulation give evidence to support the fact that CNSs may be involved in transcriptional regulation of their target genes. Also Tirosh and Barkai

(2008) reported that high nucleosome occupancy near transcription start site is associated with transcription and that regulatory elements with high occupancy are more responsive to external and internal signals in the yeast genome. These findings further support the view of CNSs playing a regulatory role. One important feature in the result is the A+T increased flanking regions just before the drop of A+T content. These A+T increased regions level off to the genomic average of the reference genome in the study. It can be argued that the regions with high A+T content does not fold into nucleosomes, rather they can be acting as linker regions with low G+C content (Nishida 2012) that is adjacent to nucleosomes. I also found that the A+T drop may not be related to recombination rate variation in the genome.

Gene enrichment analysis carried out for grass and monocot-specific CNSs suggests that CNSs tend to locate close to genes involved in DNA binding, transcription regulation and transcription factor activity. Animal genome analyses demonstrated that CNSs are found near genes involved in regulation of transcription and development (Sandelin et al. 2004; Shin et al. 2005; Venkatesh et al. 2006; Matsunami and Saitou 2013; Babarinde and Saitou 2013). This finding for the lineage specific CNSs also agree with animal CNS studies reported so far. One interesting feature to note down is that lineage specific genes and lineage specific CNSs have different functional classifications. Lineage specific genes were found to be plant defense related whereas lineage specific CNSs are related to regulation of transcription and development. Therefore it appears that lineage CNSs and genes are functioning in two diverse arenas to ensure thorough overall accurate functioning of the plant. Interestingly Babarinde and Saitou (2013) reported that two underrepresented terms for their GO analysis for CNSs include categories related to stimulus and defense.

Even though I considered the closest gene to the CNSs as the likely target gene, it is noteworthy that without experimental support and evidence it is hard to establish the actual target genes of CNSs. Also it is important to note that there are exceptions to the above mentioned scenario. As reported by Lettice et al. (2003) a regulator designated as ZRS responsible for early spatio-temporal expression pattern in the limb of tetrapods lies in intron 5 of *Lmbr1* gene where the target gene *Shh* lies1 Mb away from the enhancer.

The result achieved for grass and monocot CNSs and also for certain grouping for eudicots and angiosperms are consistent with the established phylogeny of plants (Bennetzen et al. 2012; Angiosperm phylogeny group 2003) thus agrees with the expectation of CNSs being orthologous in different species. This analysis also shows that CNSs can be used to construct species trees provided that concatenated sequence lengths are of considerable lengths with enough informative sites.

In this study I identified 27 eudicot, 204 monocot, 6536 grass, 19 angiosperm and 2 vascular plant lineage specific CNSs that originated in their respective common ancestors. I also observed a stronger constraint on CNSs located on UTR regions. The CNSs are flanked by genes involved in transcription regulation and also a drop of A+T was observed near the borders of the CNSs. Further the CNSs showed a high nucleosome occupancy probability. This study provides candidates of regulatory elements that can be experimentally tested for their potential functionality. These findings along with other investigations on plant CNSs will help to establish an understanding to shape the regulatory landscape of plants, governed by conserved noncoding sequences.

Chapter 3

Determination of GC content heterogeneity of CNSs in Eukaryotes

3.1 Introduction

CNSs have been extensively studied for its functions related to transcription regulation and development (e.g., Lee et al. 2011; Clarke et al. 2012). Vavouri et al. (2007) reported a drop of AT content in the flanking regions of CNSs in *Takifugu rubripes, Homo sapiens, Caenorhabditis elegans* and *Drosophila melanogaste*r genomes. Babarinde and Saitou (2013) reported a sharp decrease in GC content of CNSs.

The same pattern of AT drop near the boundaries of CNSs has been observed in plant CNSs. Kritsas et al. (2012) report AT drop of flanking regions of *Arabidopsis thaliana* and *Brachypodium* *distachyon* CNSs. Hettiarachchi et al. (2014) reported AT drop of the boundaries of grass, monocot and eudicot lineage specific CNSs, as well the CNSs. Along with the AT drop certain groups reported an increase in the nucleosome occupancy probability for these CNSs (Baxter et al. 2012; Hettiarachchi et al. 2014). A recent study (Seridi et al. 2014) on highly conserved noncoding elements (HCNEs) of *Drosophila melanogaster* showed a drop of nucleosome occupancy toward the center of the HCNEs. So far there have not been many studies on CNSs and its relations with nucleosome occupancy. Further it has to be noted that since it is documented and known that nucleosome which is the repetitive unit of chromatin is inhibitory to transcription factor binding, experimental evidence is required to verify the functionality and the molecular mechanism by which CNSs located in folded chromatin regions can act as regulatory elements. This aspect of structural architecture of CNSs and their functionality has yet to be fully explained. Apart from the above, a recent study showed that distance between CNSs follow a power-law like distribution pattern by investigating the chromosomal distribution of CNSs (Polychronopoulos et al. 2014). They tested this feature for amniotic, mammalian, fly and worm CNSs and found that this pattern for CNSs remained even after they removed the closest genes to the CNSs from the analysis. In addition to the function of CNSs so far various structural features and distribution patterns have been investigated with regards to CNSs.

In this analysis I tried to focus on several aspects of CNSs which includes GC content, nucleotide frequency patterns, nucleosome occupancy probability and substitution pattern. In my previous study on lineage specific plant CNSs (Hettiarachchi et al. 2014), the determined CNSs were GC rich. However Babarinde and Saitou (2013) reported that mammalian CNSs are GC poor. The first impression out of the two results is that there seem to be a GC content heterogeneity in CNSs of different groups of organisms. This heterogeneity might be related to lineage specific nucleotide preferences in regulatory elements. In order to determine where in the line of evolution this GC content heterogeneity for CNSs first appeared, I initiated the analysis with fungal genomes which is a sister group of animals. Then I expanded the analyses to invertebrates and non-mammalian vertebrates to obtain eukaryote wide perspective of the features of CNSs.

In summary I found that fungi and invertebrate lineage common CNSs were predominantly GC rich whereas non-mammalian CNSs were GC poor following the pattern similar to mammalian CNSs. The invertebrate and non-mammalian vertebrate CNSs show a consistent GC=>AT substitutions, whereas fungi CNSs showed a heterogeneous pattern of substitutions. I found that non-mammalian vertebrate common CNSs are positioned in open chromatin regions, whereas invertebrate CNSs except Diptera CNSs were predicted to be located inside well positioned nucleosomes. Here I also report that the GC poor feature is characteristic to vertebrates and that specifically appeared in the vertebrate lineage when compared to other lineages which is governed by evolutionary dynamics of transcription factor binding of the vertebrates. Another evident feature was that the vertebrate CNSs were located in long stretches of GC rich isochore-like regions.

3.2 Materials and Methods

3.2.1 Genomes considered in the analysis

Repeat masked genomes of 24 fungi, 19 invertebrates and 12 non-mammalian vertebrates were downloaded from Ensembl release 78. The analyses were focused on the nuclear genome.

3.2.2 Identification of lineage common CNSs

BLAST 2.2.25+ (Altschul et al. 1997) was used for performing homology searches in this study.

CNSs common to invertebrates: The BLASTn search was done for individual orders in group invertebrates. The genomes considered for this analysis include orders Diptera, Lepidoptera, Hymenoptera and Nematoda. BLASTn search was done with one species as the query and the second species as the subject database. The cut off e-value for the search was 0.001. The alignments without any overlap with a coding region for both query and the subject were considered for subsequent analyses. The best hits selected based on the e-value were searched in the third species. Similarly the four mosquito genomes were searched against each other (*Anopheles gambiae*-vs-*Anopheles darlingi* and *Aedes aegypti*-vs-*Culex quinquefasciatus*) and best hits obtained from the mosquito genomes were searched in best hits obtained for fly genomes to obtain the Diptera common CNSs. Similarly Lepidoptera, Hymenoptera and Nematode common CNSs were obtained by pairwise chain search.

CNSs common to non-mammalian vertebrates: The BLASTn searches for non-mammalian vertebrates were performed in a similar manner with a cutoff e-value of 0.001. The initial search for birds was done with *Gallus gallus* (Chicken) as the query and *Meleagris gallapavo* (Turkey) as the subject database. The best hit results were searched in *Anas platyrhynchos* (Wild duck). The best hits from this step were searched in *Taeniopygia guttata* (Zebra finch) finally to obtain bird common CNSs. The best hits from previous step were searched in the following new species, *Pelodiscus sinensis* (Chinese softshell turtle), *Anolis carolinensis* (Anolis lizard) and *Xenopus tropicalis* (Xenopus frog) with the expectation of finding reptilian, reptilian and amphibian shared CNSs. The CNSs that are found in all teleost fishes were found with the same strategy within the group non-mammalian vertebrates.

CNSs common to fungal genomes: Fungal common CNSs were determined for nine different orders. Determining lineage common CNSs for fungi follows the same method used for invertebrate and non-mammalian vertebrates.

A depiction of the pipeline used in identifying the lineage common CNSs is provided in Figure 3.1.

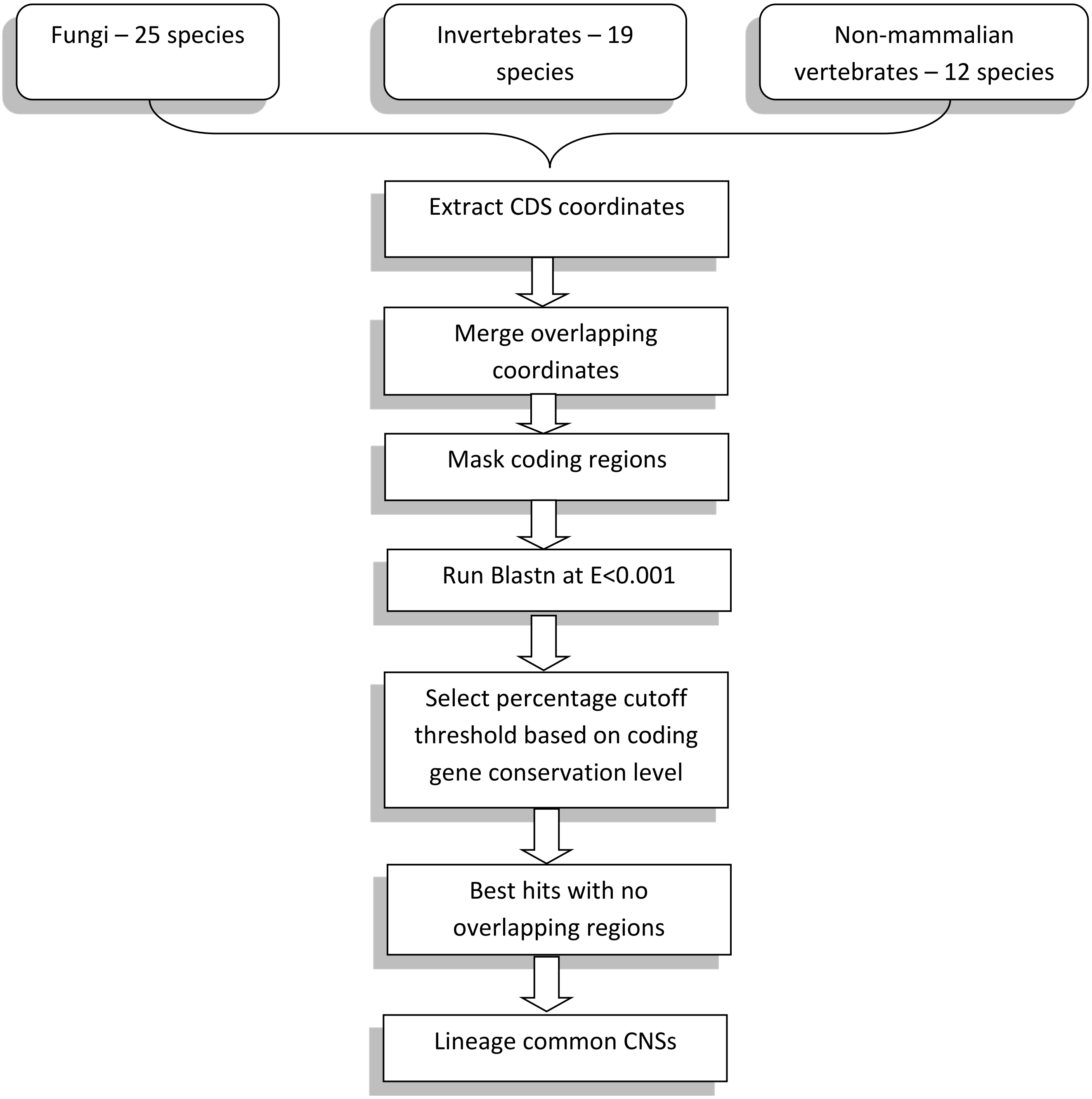

Fungi – 25 species
Invertebrates – 19 species
Non-mammalian vertebrates – 12 species
Extract CDS coordinates
Merge overlapping coordinates
Mask coding regions
Run Blastn at E<0.001
Select percentage cutoff threshold based on coding gene conservation level
Best hits with no overlapping regions
Lineage common CNSs

Figure 3.1 - The pipeline followed to identify the lineage putative lineage common CNSs for fungi, invertebrates and non-mammalian vertebrate species used in the analysis.

3.2.3 Setting percentage identity cutoff for the CNSs

I used the gene based approach as Babarinde and Saitou (2013) to set the percentage identity cutoff for CNSs. This step is needed to identify the conserved regions that might actually be under selective constraint against regions that are not under functional constraint but appear conserved since they didn't have enough time to accumulate mutations. For this analysis I considered only one-to-one orthologous cDNA sequences for the reference genome of a particular group and the most basal species within the same group. For the invertebrates cDNA searches I considered *D. melanogaster* and *A. darlingi* with respect to Diptera, *Danius plexippus* and *Bombyx mori* for lepidotera, *Atta cephalotes* and *Nasonia vitripennis* for hymenoptera and *C. briggsae* and *C. japonica* for nematode. Similarly for non-mammalian vertebrate cDNA searches I used *G. gallus* and *T. guttata* for birds, *G. gallus* and *P. sinensis* for protein conservation between birds and group chelonia, *G. gallus* and *A. carolinensis* for all reptiles, *G. gallus* and *X. tropicalis* to find level of protein conservation between reptiles and amphibians. *T. nigroviridis* and *Danio rerio* were used to determine the protein conservation level for teleost fish. Same strategy was followed for fungal genomes (Appendix A8) after performing BLASTn searches on query and the subject, reciprocal best hits were selected for each of the above mentioned pairs of species. The average percentage identities for the reciprocal best hits were considered as the cutoff threshold for the CNSs in the respective groups.

3.2.4 Determining GC content of CNSs and the reference genome

The GC content of the CNSs were determined and compared with the noncoding GC content of the reference genomes.

3.2.5 Determining ancestral GC content of CNSs and background noncoding regions

Multiple sequence alignments for the CNSs and the background noncoding regions were constructed with clustalw. Based on these multiple sequence alignments ancestral sequences were constructed using FASTML (Ashkenazy et al. 2012).

3.2.6 Substitution pattern determination for CNSs

The Multiple sequence alignments of CNSs constructed with clustalw were used to determine substitution patterns in MEGA6 (Tamura et al. 2013). Tamura-Nei model (Tamura and Nei 1993) was considered for substitution pattern determination.

3.2.7 GC content distribution in the flanking regions and inside of CNSs

I analyzed the GC content distribution in 1000bp flanking regions and the center (20bp) of the CNSs by a moving window analysis (10bp window with 1 base step size). The statistical significance was assessed by t-test.

3.2.8 Isochore-like regions identification

The flanking regions of bird, bird-Chelonian shared, reptilian, reptilian and amphibian shared CNSs were considered for this analysis. The flanking regions were extended up to 12kb regions and the classification of isochore like regions were based on Costantini et al. (2006). Mammalian common CNSs were retrieved from Babarinde and Saitou (2013) to compare with the non-mammalian vertebrate CNSs.

3.2.9 Determination of nucleosome occupancy probability

Nucleosome occupancy probability was determined by using the computational model produced by Kaplan et al. (2009) by considering nucleotide preferences in nucleosome regions. The link to the program is http://genie.weizmann.ac.il/software/nucleo_prediction.html. The nucleosome occupancy probabilities for all groups in the study were computed. Initially I extracted a total of 8000 bases from the center of the CNSs. Then the average nucleosome occupancy probability was calculated for each site along the complete length of 8000 bases. The same analysis was done for random samples with the same length and the same number of sequences as the CNSs. The statistical significance was determined by t-test.

3.2.10 Association of histone modifications with CNSs

Certain histone modification signals are known to be signatures for some genomic regulatory regions such as promoters and distal enhancers.H3K4Me3 has been found to be highly associated with gene promoter regions (Tserel et al. 2010) whereas H3K4Me1 and H3K27ac are known to be related with nucleosome regions that flank enhancer elements (Creyghton et al. 2010; Heintzman et al. 2009).The regions with these histone modification signals are considered to be active enhancer positions in numerous studies as stated above.

I determined histone modifications associated with zebrafish and nematode CNSs. The coordinates for histone modifications data for *C. elegans* were downloaded from modencode (http://www.modencode.org/) project and zebrafish chromatin signature marks were retrieved from Bogdanović et al. (2012). This analysis was done only for the two above mentioned species as the histone modification data is not available for the rest of the species used in this study.

3.2.11 Predicted target genes of CNSs

I considered the closest gene to the CNS as the most plausible likely target gene. For the CNSs that were found inside introns or UTR regions, the gene that they reside in was considered as their target gene. The genes were considered based on the reference genomes used for each group in the analysis. The GO analysis for the target genes were performed using DAVID (The Database for Annotation, Visualization and Integrated Discovery, version 6.7).

3.2.12 Transcription factor binding site (TF binding site) analysis for the CNSs in vertebrates

The TF binding site data for human was downloaded from UCSC table browser (GRch37/hg19). A total of 4286829 binding sites were considered for 150 transcription factors. In order to determine the ancestral TF binding sites that are shared across lineages I searched (Blastp) the human TF gene sequences in *Arabidopsis thaliana*, *Oryza sativa and C. briggsae* protein coding genes and determined the union of TF genes that are shared across lineages. Also I tried to compare this data with random expectation by searching all the longest transcripts of protein coding genes of human against *A. thaliana, O. sativa and C. briggsae* protein coding genes. All Blastp searches were performed with e-value < 0.00001.

3.2.13 Transcription factor binding site analysis for plants

Fifty six thousand five hundred and twenty eight transcription factor binding site data for 27 transcription factors were downloaded from supporting data provided by Heyndrickx et al. (2014). The binding site information is based on the *A. thaliana* genome. In order to test for TF binding site genes that are shared across lineages I tested the homology of these sequences in human, chicken and fugu genomes. And to compute the random expectation of *A. thaliana* genes that find homology in other lineages I searched all the protein coding genes of *A. thaliana* in the above mentioned genomes.

3.3 Results

3.3.1 Fungi lineage common CNSs

I identified 467 Eurotiales, 1536 Pleosporales, 339 Hypocreales, 201 Schizosaccharomycetales, 2412 Slerotiniaceae, 288 Magnaporthales, 26 Saccharomycetales, 22053 Pucciniales and 669 Ustilaginales CNSs (Table 3.1). Despite the long divergence times (minimum divergence between two species was 100 million years ago (mya)) the fungal species had considerable number of CNSs. The average lengths of the CNSs were above 50bp for all orders. The length distributions for the CNSs are provided in Appendix A9.

I compared the GC contents for the fungal CNSs along with the reference genome noncoding regions. Eurotiales, Pleosporales, Hypocreales and Schizosaccharomycetales showed statistically significantly higher GC content than the genomic average. Sclerotiniacea and Pucciniales had higher GC CNSs compared to the genomic average, but the values were not statistically significant (Appendix A12). Ustilaginales CNSs were not significantly different from genomic noncoding GC and Saccharomycetales had too low number of CNSs for any statistical inference. The fungal CNSs showed a pattern of being predominantly GC rich.

3.3.2 Invertebrate lineage common CNSs

Invertebrate lineage common CNSs were higher in number compared to fungal CNSs. I identified 50338, 2716, 20513, 15573 and 5121 CNSs for Drosophid, Mosquito, Lepidoptera, Hymenoptera and Nematoda respectively (length distributions are provided in Appendix A9). Diptera which constitute drosophid and mosquitos had 1194 CNSs and this very low number could be due to its long divergence time. In the invertebrate lineage the Lepidoptera, Hymenoptera and Nematode CNSs were GC rich compared to the genomic average. The Diptera CNSs showed a very low GC content of 20.72% compared to all other orders. Inside the order Diptera, drosophids alone showed a slightly higher GC (not statistically significant) than the genomic average of *Drosophila melanogaster*. Whereas mosquito CNSs were considerably GC poor (29.92%) compared with the reference genome (*Aedes aegypti*). The CNSs that are common to the order Diptera (the CNSs that are shared between drosophids and mosquitoes) showed a pattern of being GC poor. Even when I considered species pairs with roughly the same divergence time between drosophids and mosquitoes, namely *D. melanogaster* – *Drosophila ananassae* (44.2 mya) and *Culex quinquefasciatus* – *Aedes aegypti* (43.3 mya) (Saisawang and Ketterman 2014; Marinotti et al. 2013) the pattern of GC remained the same where drosophid CNSs were GC rich and mosquito CNSs were GC poor. Therefore the very low GC content observed for Diptera group could not have solely been brought about by the age of the CNSs. Even though the extant CNS GC content is low, the ancestral GC content analysis showed that the ancestral CNSs of order Diptera had been considerably GC richer (about 26%) compared to the current CNSs GC content. The Diptera CNSs have a GC=>AT substitution pattern thus are becoming more GC poor with time (Appendix A10). This is explained by the lower GC content of current CNSs. In general, GC content of ancestral CNSs of Diptera was lower than the ancestral state determined for other orders.

3.3.3 Non-mammalian vertebrate lineage common CNSs

I found 25011 CNSs that are commonly shared among the four bird species used in the analysis. Between order birds and *Pelodiscus sinensis* (Chinese softshell turtle), there are 8007 shared CNSs. Reptilian, Reptile and Amphibian shared CNSs were 4477 and 2305 respectively. Fifteen thousand one hundred sixty eight CNSs were identified for the five teleost fish species used in the analysis (the length distributions are provided in Appendix A9). All non-mammalian vertebrate CNSs had

GC content lower than the genomic average of the reference genomes (*Gallus gallus* and *Tetraodon nigroviridis*) that were considered in the analysis.

Lineage	**No: species**	**No: CNSs**	**Mean Length (bp)**	**Mode length (bp)**	**Genome size (Mb)**	**Proportion of CNSs in genomes (%)**
Fungi						
Eurotiales	4	467	89.64	61	34.85	0.12
Pleosporales	3	1536	142.89	42	37.99	0.58
Hypocreales	4	339	74.34	35	35.00	0.07
Schizosaccharomycetales	3	201	73.47	64	11.26	0.13
Sclerotiniaceae	2	2412	150.44	50	40.00	0.91
Maganaporthales	2	288	103.03	84	34.53	0.08
Saccharomycetales	2	26	111.69	55	9.11	0.03
Pucciniales	2	22053	85.44	53	126.64	1.48
Ustilaginales	2	669	122.46	46	19.74	0.41
Invertebrates						
Diptera	7	1194	25.46	23	142.57	0.02
Hymenoptera	4	15573	58.41	34	281.12	0.32
Lepidoptera	3	20513	105.26	51	272.85	0.79
Nematoda	5	5121	61.90	37	108.35	0.29
Vertebrates						
Birds	4	25011	261.33	37	1072.54	0.61
Birds and chelonia	5	8007	323.72	48	1072.54	0.24
Reptiles	6	4477	300.22	46	1072.54	0.12
Reptiles and amphibian	7	2305	253.54	98	1072.54	0.05
Teleosts	5	15168	116.38	65	342.41	0.51
Mammals	19	10939	201.30	103	3160.00	0.07

Table 3.1 - Number of lineage common CNSs, mean and mode lengths of CNSs identified for groups' fungi, invertebrate and non-mammalian vertebrates.

3.3.4 The pattern observed in GC content transition

In the previous study I identified the lineage specific plant CNSs to be GC rich (Hettiarachchi et al. 2014). Here I report the fungi and invertebrate CNSs also to be predominantly GC rich. Babarinde and Saitou (2013) reported that mammalian CNSs are GC poor. In this study I observed a transition of the GC content in non-mammalian vertebrate CNSs. Similar to mammalian CNSs I found the non-mammalian CNSs are also GC poor (Figure 3.2). This shows a change in nucleotide preference that occurred in the vertebrate lineage with regards to CNSs or putative regulatory elements compared to other eukaryotes. This transition may be related to vertebrate specific CNSs that originated in the vertebrate common ancestor or a transition in the sequence preferences in transcription regulatory binding sites in vertebrates. The distribution of the CNS GC content along with reference genome is given in Appendix A11.

Close observation of Diptera group revealed that relative GC content change for mosquitoes is much higher than for drosophilids, and the relatively high GC content change in Diptera group was brought about mainly due to mosquito and drosophilid shared CNSs (Figure 3.3).

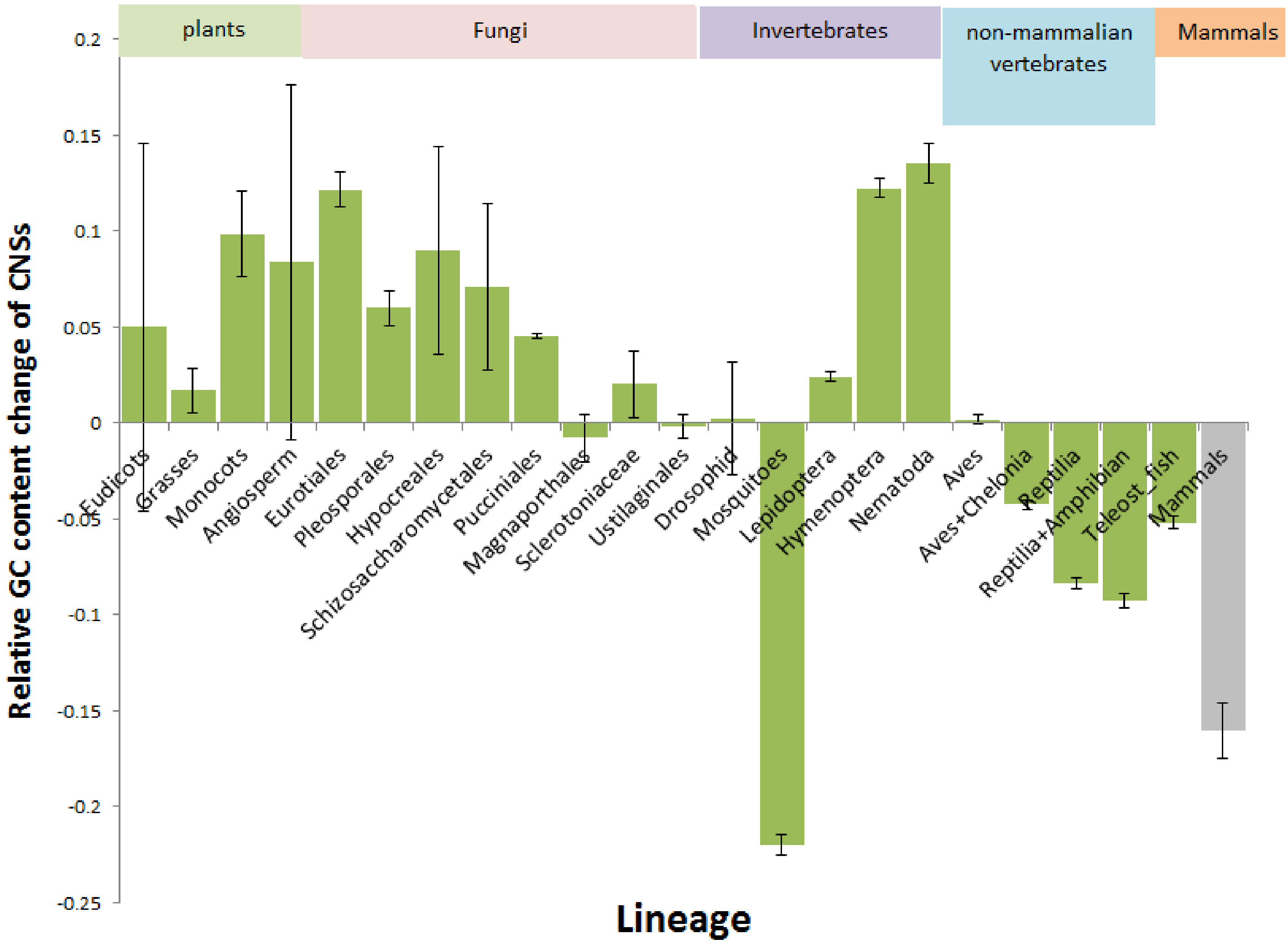


Figure 3 - Relative GC content change for the CNSs. The relative change of GC content was calculated based on the noncoding GC content for the reference genome of each lineage.

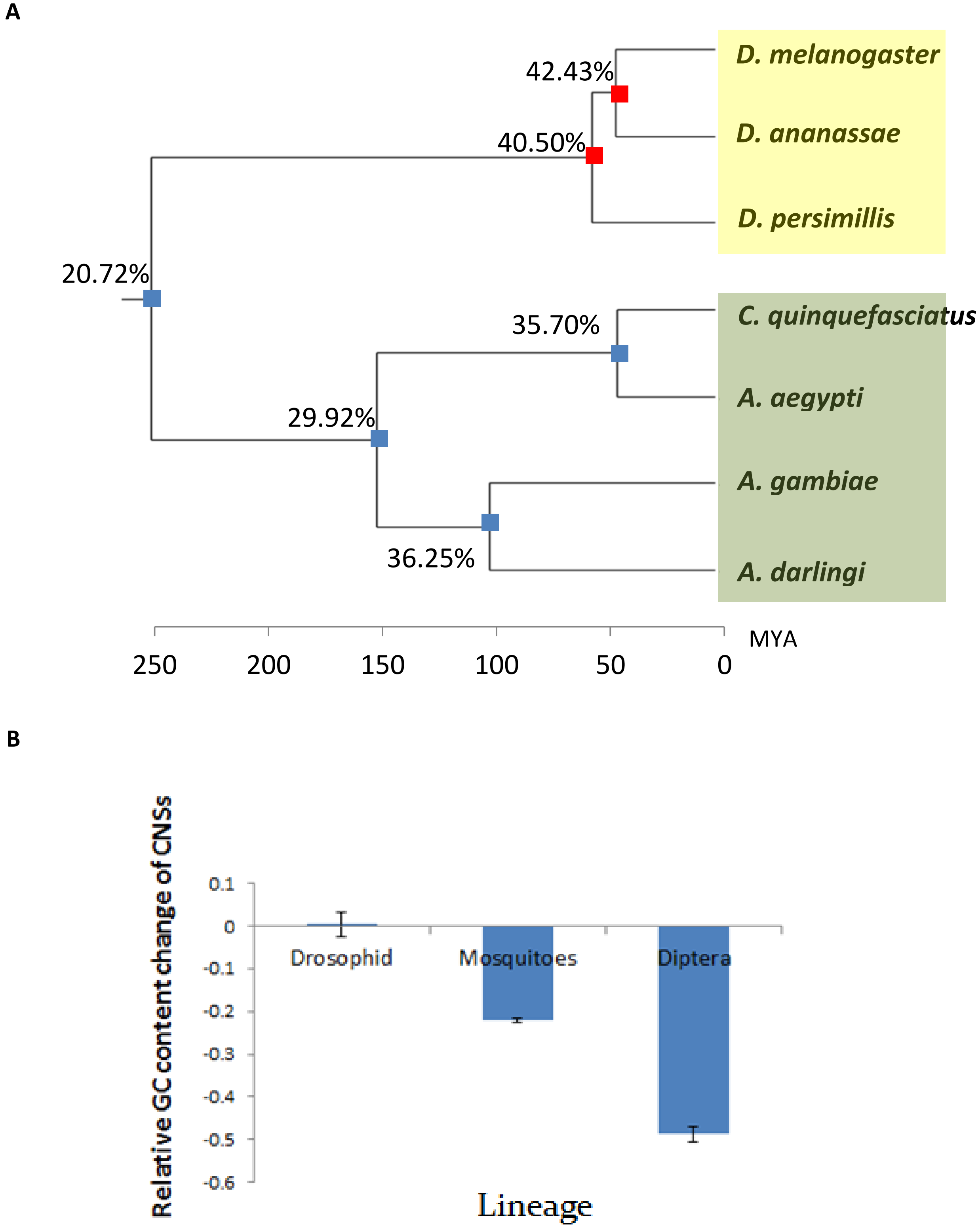

A
D. melanogaster
D. ananassae
D. persimillis
C. quinquefasciatus
A. aegypti
A. gambiae
A. darlingi
42.43%
40.50%
20.72%
35.70%
29.92%
36.25%
250
200
150
100
50
0
MYA
B
Relative GC content change of CNSs
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
Drosophid
Mosquitoes
Diptera
Lineage

Figure 3.3 - (A) GC content change in order Diptera. In these analyses the order Diptera contain drosophids (yellow background) and mosquitoes (green background).

The common CNSs identified at each node was considered for determining the GC content and red squares correspond to high GC CNSs whereas blue squares correspond to low GC CNSs. The actual GC contents are provided in the phylogenetic tree for groups' drosophids, mosquitoes and order Diptera.

(B) Relative GC content change of drosophids, mosquitoes and order Diptera with respect to the reference genome. *A. aegypti* was used as the reference genome for mosquito group was and *D. melanogaster* was considered as the reference for both drosophid and Diptera in general.

3.3.5 Nucleosome occupancy and GC content distribution for CNSs

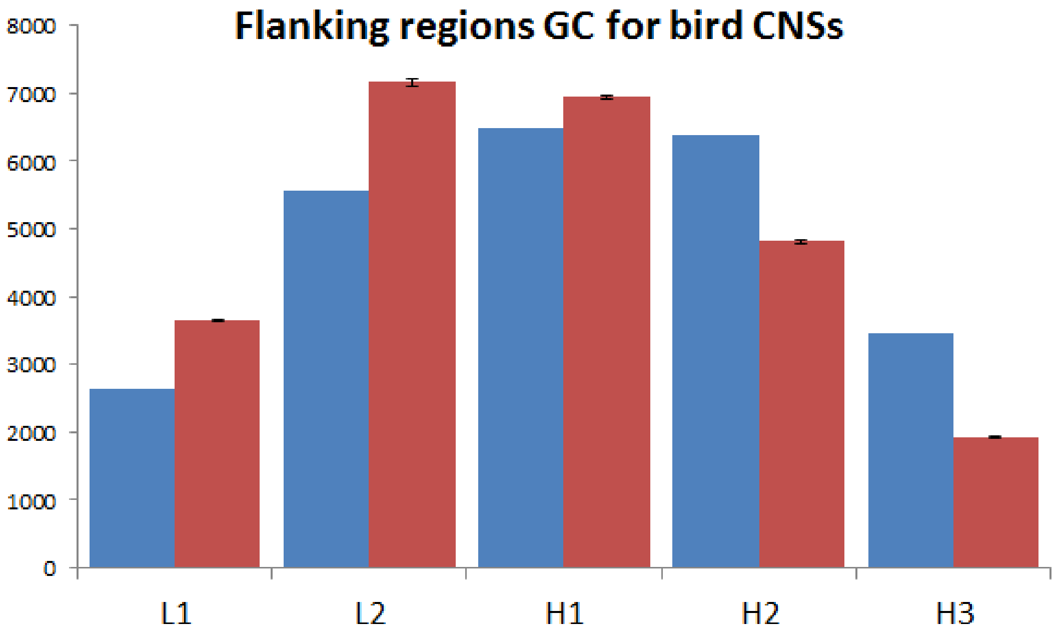
The CNSs are located in numerous structurally diverse regions. One reason for this observation is the diverse nucleotide composition in these regulatory regions. This diversity in GC content of the CNSs in turn results in positioning them in open chromatin or heterochromatin regions. Nucleosome occupancy is known to be directly related to regulation of genes according to Jiang and Pugh (2009). Furthermore, nucleosome occupancy has been reported to be directly associated with nucleotide composition (Kaplan et al. 2008; Tillo and Hughes 2009; Gaffiney et al. 2012). It has been reported that GC rich sequences have a high propensity to form nucleosomes whereas lower GC regions will prefer an open chromatin conformation (Warnecke et al. 2008; Washietl et al. 2008). I found that Lepidoptera, Hymenoptera and Nematode CNSs showed a high nucleosome occupancy probabilities. These CNSs seemed to be located in a well-positioned nucleosome region. Diptera CNSs showed lower nucleosome occupancy which goes in line with their very low GC content.

The Diptera CNSs specially seem to be located in very low GC open chromatin regions margined by two well positioned nucleosomes. This structurally constrained conformation may also be related to functional aspect of the CNSs which are yet to be fully elucidated. These GC- rich flanks have been documented as container sites by Valouev et al. (2011). Kundaje et al. (2012) showed that the container sites are a distinct feature of transcription factor binding sites. To perform regulatory functions, the transcription factors should be able to identify the correct binding site or motif from a large arena of similar regions. Therefore it can be assumed that the accurate finding of the correct binding site may lie in the sequence features along with the unique structural architecture of the binding sites. These high GC flanks might be essential for keeping the proper structural architecture of the actual transcription factor binding sites which are embedded inside these regions.

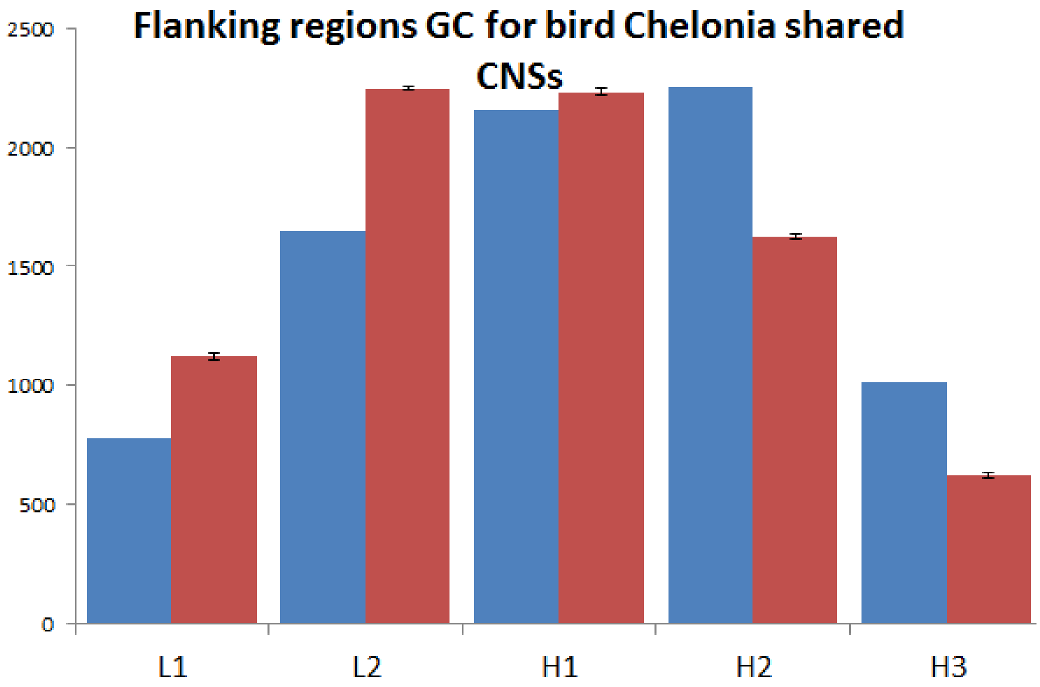
I found that non-mammalian vertebrate CNSs were GC poor and they showed a lower nucleosome occupancy probability. The teleost CNSs showed a low nucleosome occupancy toward the center of the CNSs where the CNS center was flanked by two nucleosome regions. One interesting feature observed was that the bird, bird and Chelonian shared, reptilian, reptilian and amphibian shared CNSs are flanked by long stretches of high GC regions or in other words stretches of GC rich isochore-like regions making CNSs the only low GC genomic area in that region (Appendix A13). One clear observation was that the CNSs were flanked by highly GC rich isochore-like regions (H2 and H3) compared to the randomly sampled genome sequences (Figure 3.4). The low GC regions were more abundant in random samples than for the flanking regions of CNSs. The classification of flanking regions is based on the Costantini et al. (2006).

I conducted the same analysis for mammalian CNSs. The mammalian common CNSs were kindly provided by Mr. Isaac Adeyemi Babarinde from Babarinde and Saitou (2013). The mammalian CNSs are also located in high GC background regions (Appendix A13) but this GC content was lower than the noncoding GC content of the human reference genome which is 43%. Isochore analysis for mammalian CNSs showed that mammalian common CNSs are mostly found in L2 Isochore-like regions. The flanking regions of mammalian CNSs are not as high as for the non-mammalian vertebrate CNSs but mammalian CNSs are also located in comparatively higher GC background to CNSs.

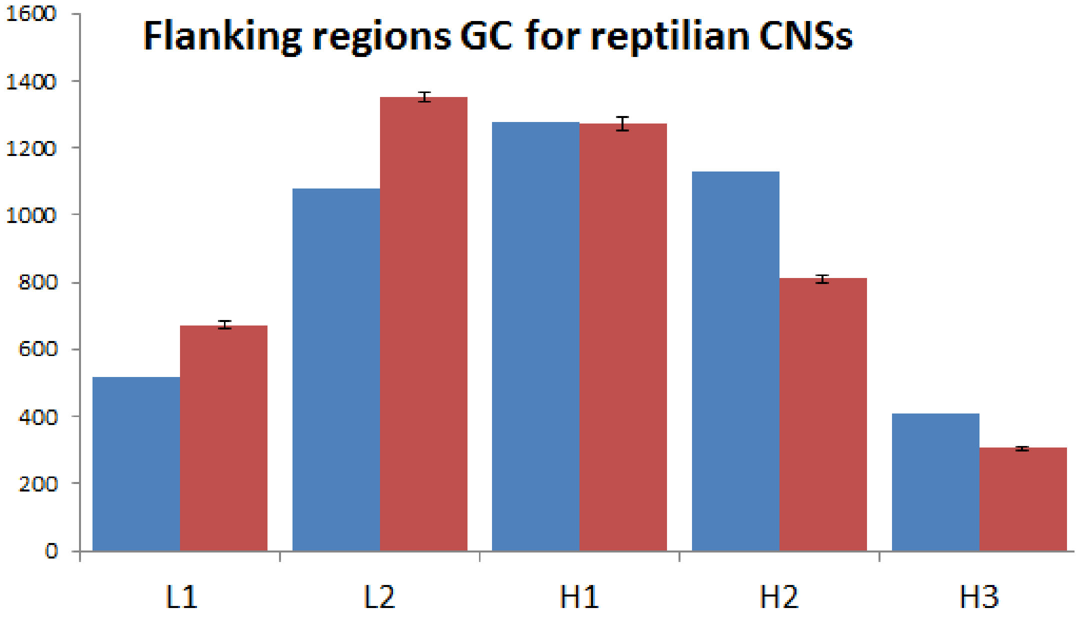
A


Flanking regions GC for bird CNSs
8000
7000
6000
5000
4000
3000
2000
1000
0
L1
L2
H1
H2
H3

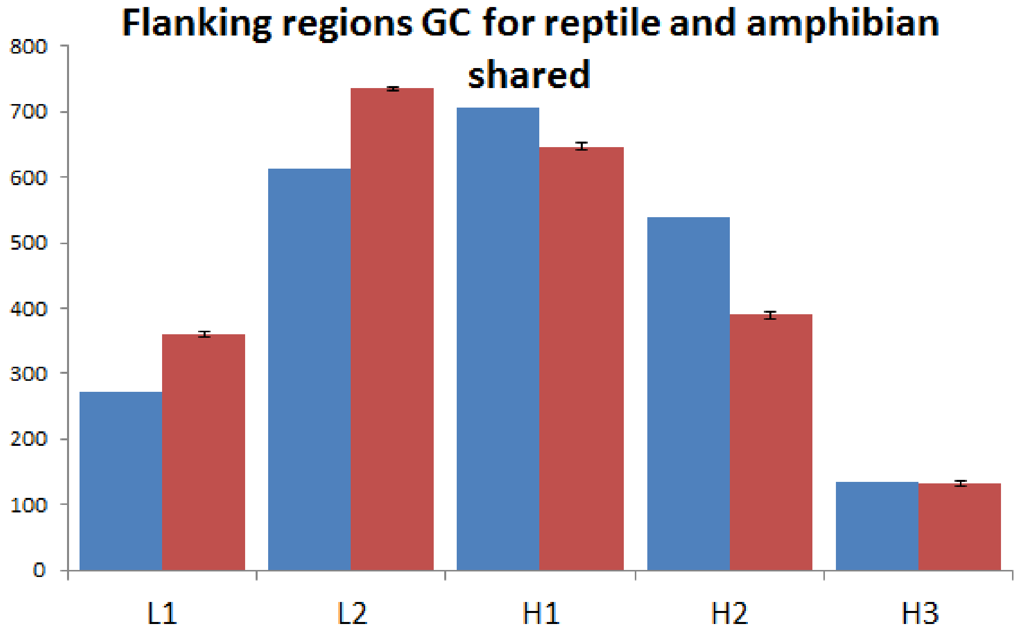
B


Flanking regions GC for bird Chelonia shared CNSs
2500
2000
1500
1000
500
0
L1
L2
H1
H2
H3

C


Flanking regions GC for reptilian CNSs
1600
1400
1200
1000
800
600
400
200
0
L1
L2
H1
H2
H3

D


Flanking regions GC for reptile and amphibian shared
800
700
600
500
400
300
200
100
0
L1
L2
H1
H2
H3

E

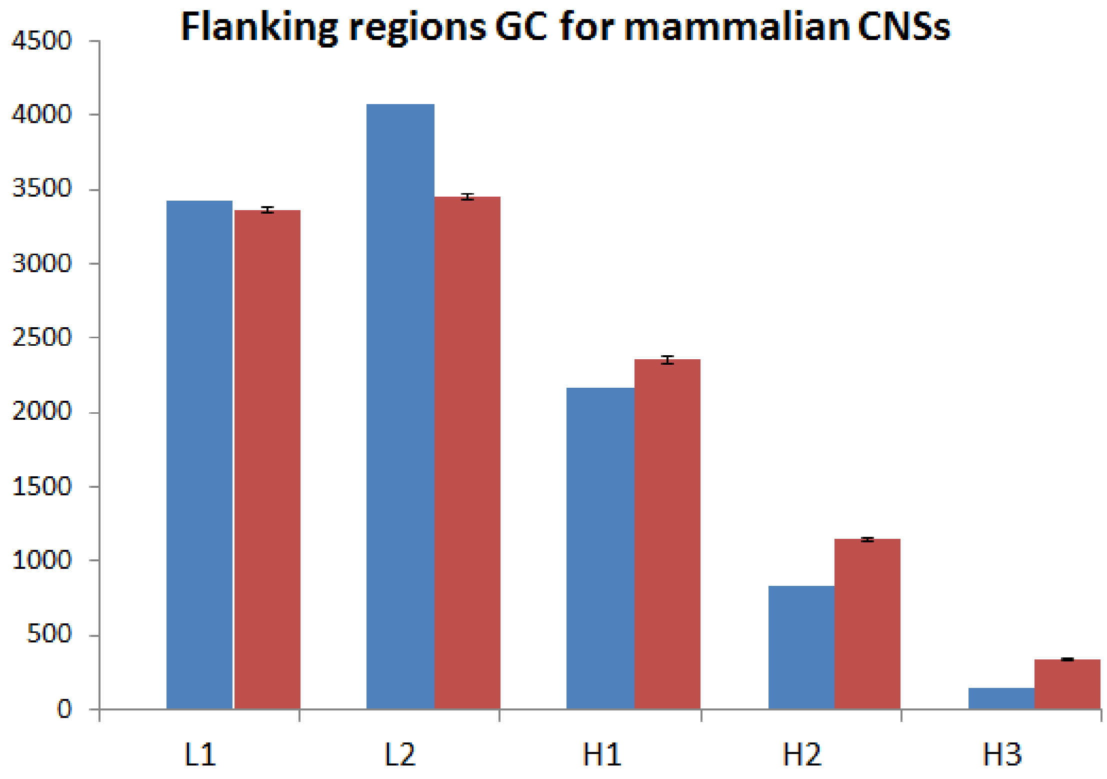
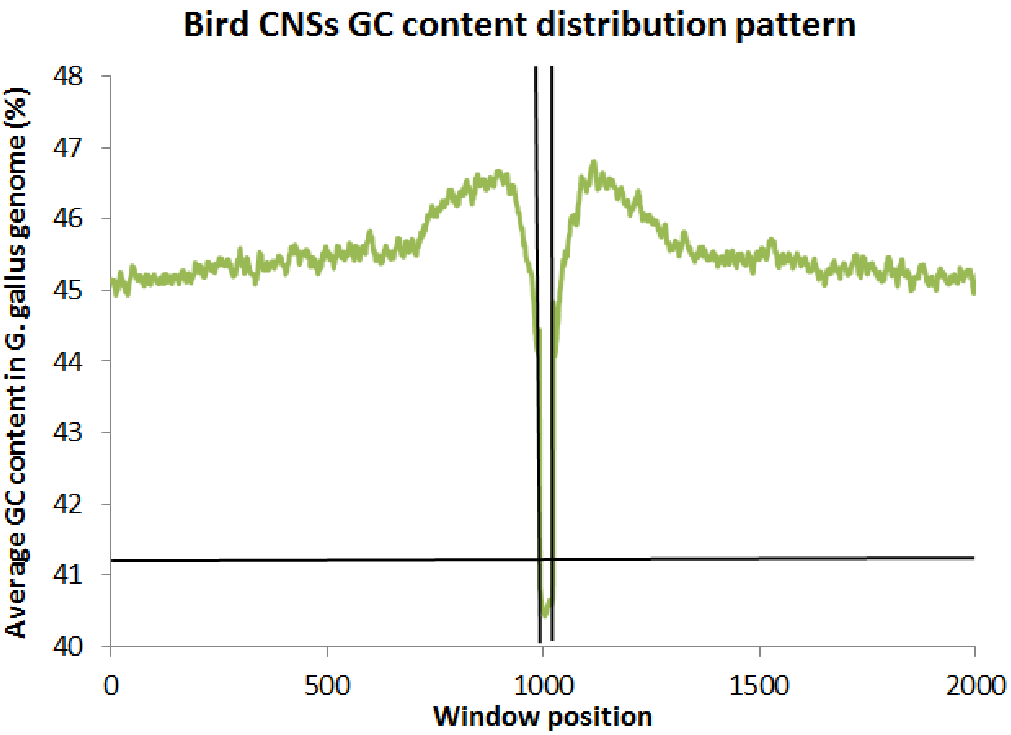


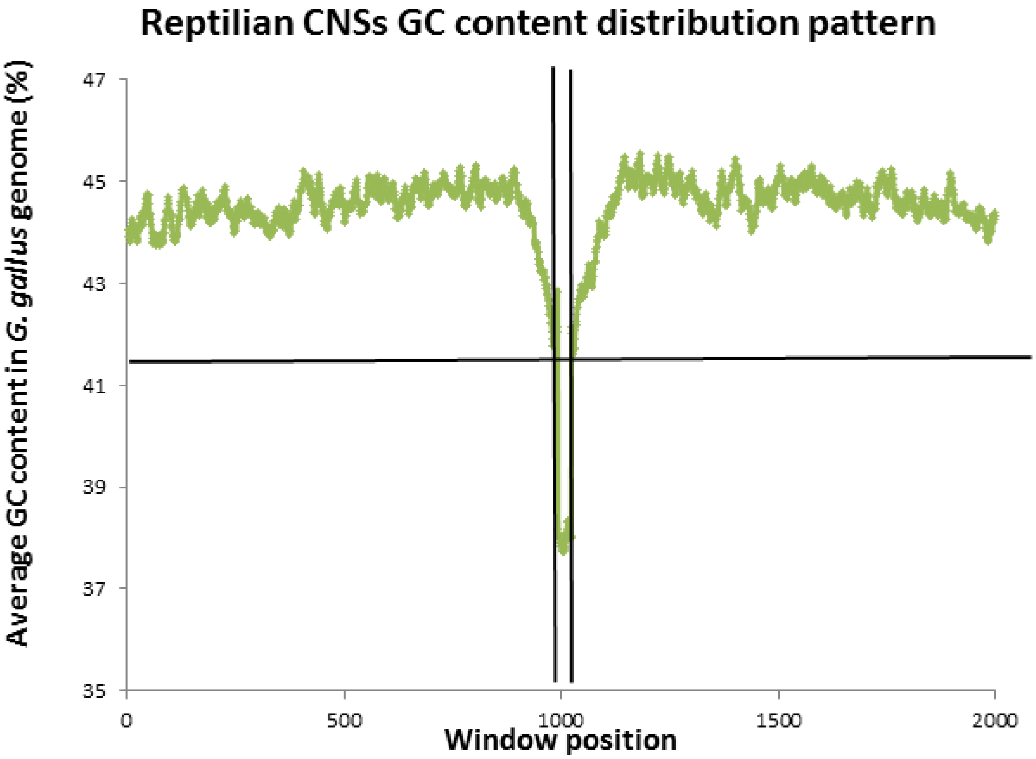
Figure 3.4 – Isochore distribution for the flanking regions of CNSs and random samples. Red bars represent the random samples and the Blue bars represent the flanking regions of CNSs. The X axis gives the groups in to which isochores are classified into (<37-L1, 37>=<40-L2, 41>=<45-H1, 46>=<52-H2, >=53-H3). The Y axis gives the frequency of each isochore segment in flanking regions of CNSs and random samples.

Figure 3.5A, 3.5B, 3.5C and 3.5D gives examples of the GC content distribution for bird, reptoile, nematode and fungi (Eurotilaes) CNSs respectively. The GC content distribution for reptile CNSs show a decline in GC inside the CNSs compared to the surrounding flanking regions. Similar to non-mammalian vertebrate CNSs mammalian CNSs also showed a decline in GC distribution inside CNSs and the flanking regions (Babrinde and Saitou 2013 Figure 5b). Nematode (invertebrates) CNSs and Eurotiales (fungi) show an elevation in the GC content of the CNSs in comparison to the surrounding genomic regions. Figure 3.6A, 3.6B, 3.6C and 3.6D provide nucleosome occupancy probability distributions for bird, reptile, nematode and Eurotiale CNSs, respectively. The nucleosome occupancy follows a similar pattern where the low GC reptile CNSs have low nucleosome occupancy probability whereas nematode, Eurotiales with high GC CNSs show a higher nucleosome occupancy probability compared to the flanking regions. The randomly sampled genome sequences in all instances showed no apparent elevation or decline in the nucleosome occupancy when compared with the CNSs.

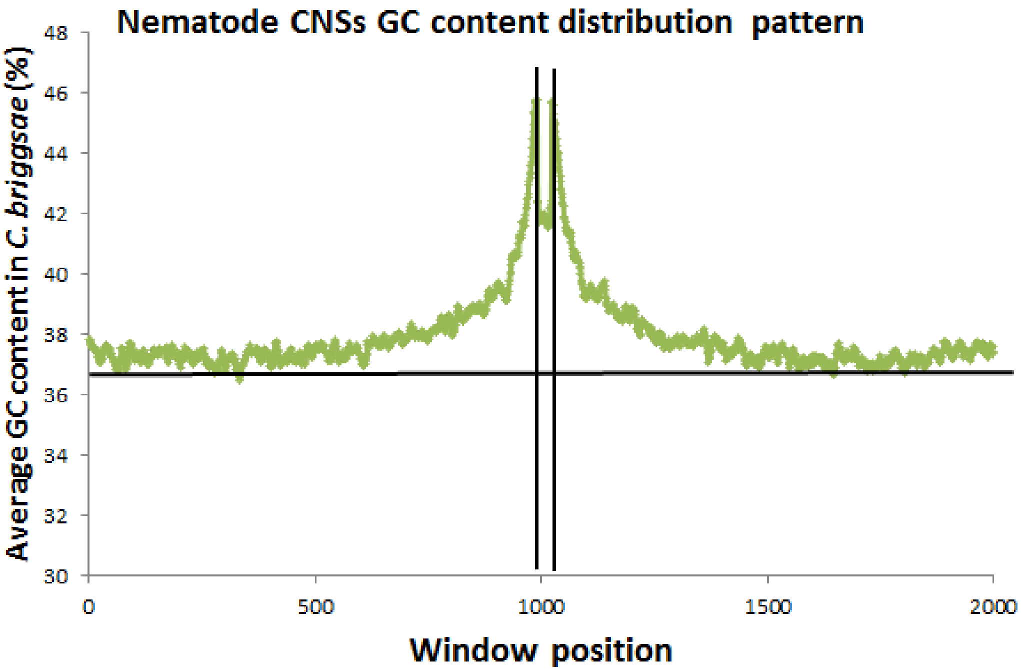
A


Bird CNSs GC content distribution pattern
Average GC content in G. gallus genome (%)
48
47
46
45
44
43
42
41
40
0
500
1000
1500
2000
Window position

B


Reptilian CNSs GC content distribution pattern
Average GC content in *G. gallus* genome (%)
47
45
43
41
39
37
35
0
500
1000
1500
2000
Window position

C


Nematode CNSs GC content distribution pattern
Average GC content in C. briggsae (%)
48
46
44
42
40
38
36
34
32
30
0
500
1000
1500
2000
Window position

D

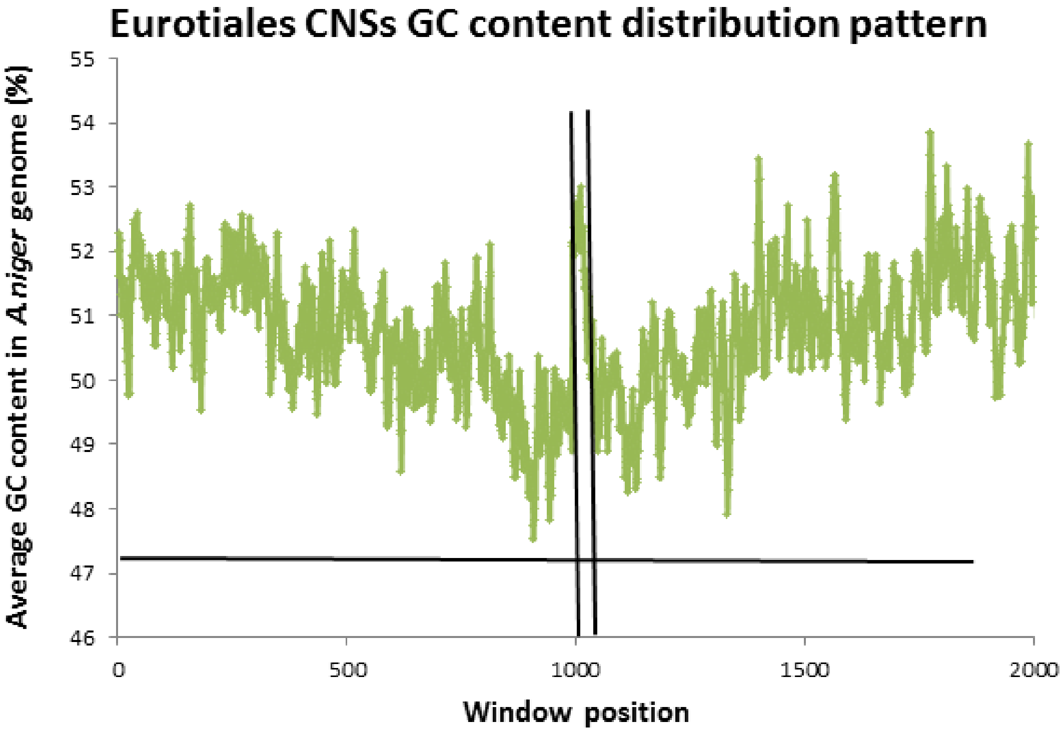
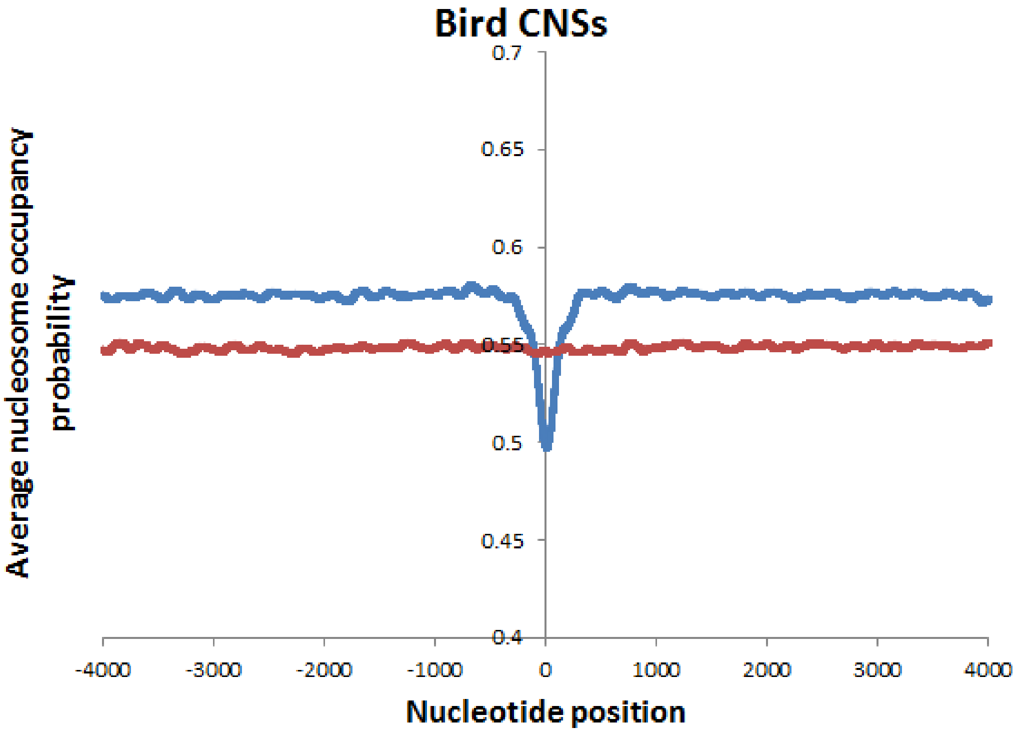

Eurotiales CNSs GC content distribution pattern
Average GC content in A.niger genome (%)
55
54
53
52
51
50
49
48
47
46
0
500
1000
1500
2000
Window position

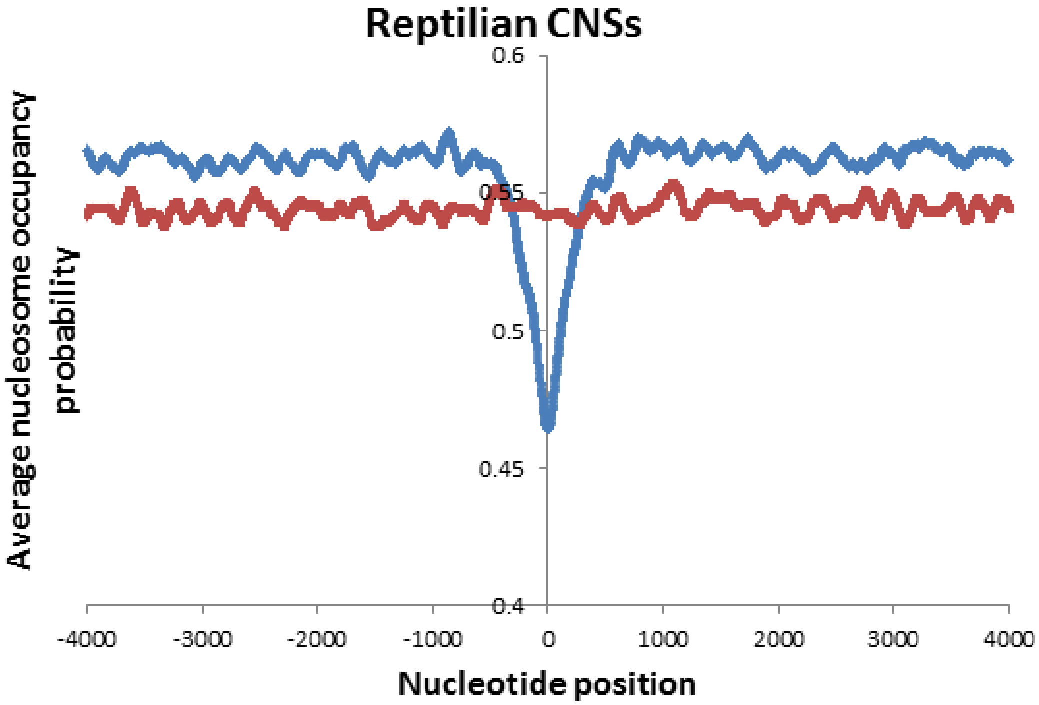
Figure 3.5 - GC content distribution of the CNSs across the center of CNSs and the flanking regions. 1000th nucleotide position corresponds to the center of the CNSs.

The horizontal back line represents the level of noncoding GC content of the reference genome. The vertical lines represent the margins of the flanking regions. (A) Bird CNSs GC content distribution. (B) Reptilian CNSs GC content distribution. (C) Nematode CNSs GC content distribution. (D) Eurotiales CNSs GC content distribution.

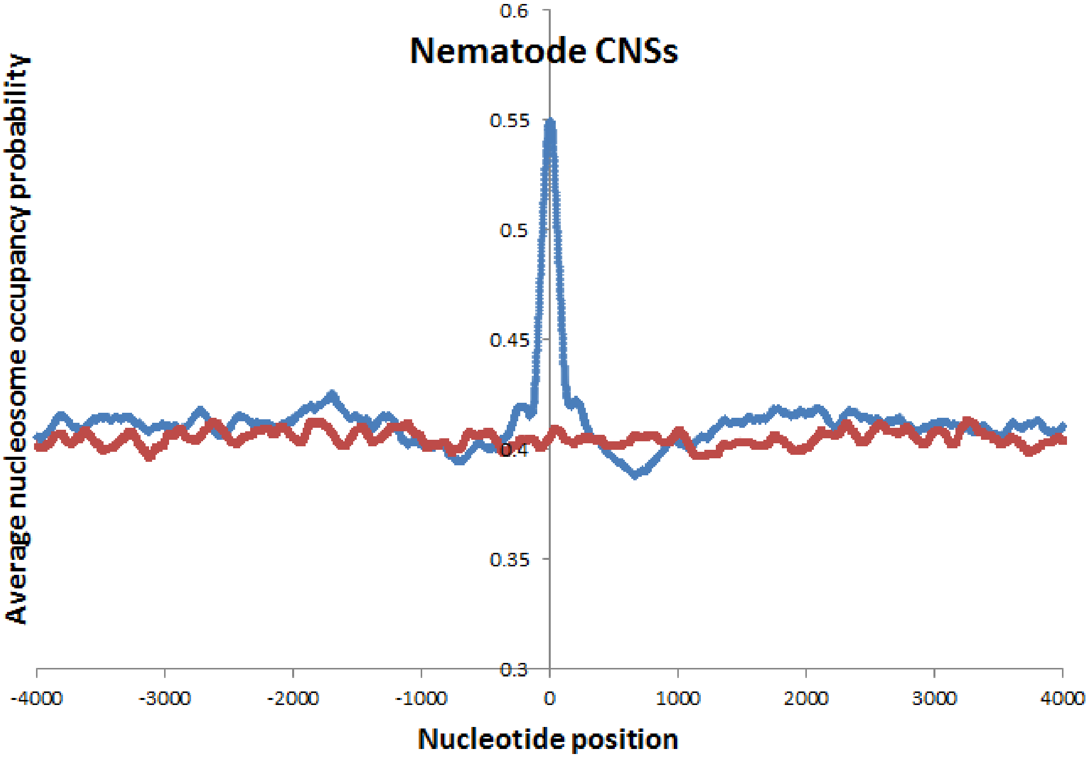
A


Bird CNSs
Average nucleosome occupancy probability
0.7
0.65
0.6
0.55
0.5
0.45
0.4
-4000
-3000
-2000
-1000
0
1000
2000
3000
4000
Nucleotide position

B


Reptilian CNSs
Average nucleosome occupancy probability
0.6
0.55
0.5
0.45
0.4
-4000
-3000
-2000
-1000
0
1000
2000
3000
4000
Nucleotide position

C


Nematode CNSs
Average nucleosome occupancy probability
0.6
0.55
0.5
0.45
0.4
0.35
0.3
-4000
-3000
-2000
-1000
0
1000
2000
3000
4000
Nucleotide position

D

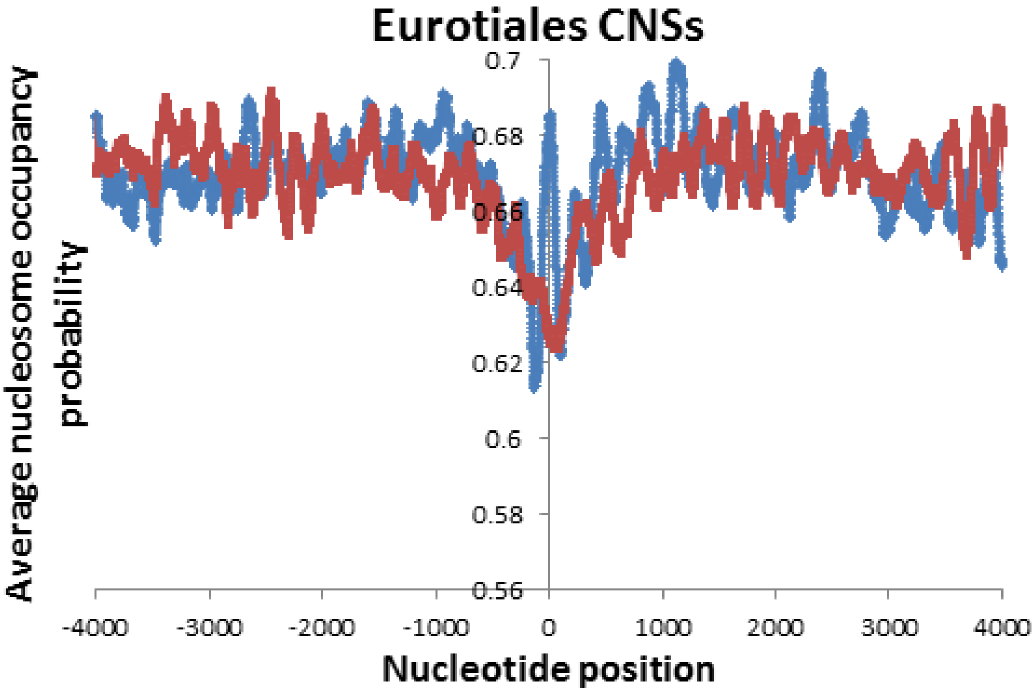

Eurotiales CNSs
Average nucleosome occupancy probability
0.7
0.68
0.66
0.64
0.62
0.6
0.58
0.56
-4000
-3000
-2000
-1000
0
1000
2000
3000
4000
Nucleotide position

Figure 3.6 - Nucleosome occupancy probability for the CNSs of different lineages.

The 0thposition represent the center of the CNSs. 8000bp flanks were considered for this analysis. The blue and red colors represent the nucleosome occupancy for CNSs and random samples respectively. (A) Bird CNSs average nucleosome occupancy probability (B) Reptilian CNSs average nucleosome occupancy probability. (C) Nematode CNSs average nucleosome occupancy probability. (D) Eurotiales CNSs average nucleosome occupancy probability.

3.3.6 Histone modifications related with CNSs

The histone modification signals related to the CNSs were determined for nematode and teleost fish conserved regions found in the analysis. Histone modifications have been studied in many organisms and they are regarded as gene regulatory signals which enable genes to be activated or repressed. Certain histone modification signals are thought to have a direct impact on regulation of genes (Hebbes et al. 1994; Kalmykova et al 2005; Buenrostro et al. 2013).

The teleost fish CNSs (tested with zebra fish chromatin modification data) showed an overrepresentation for H3K27ac and H3K4Me1 with respect to random expectation (Appendix A14-A). These modification signals are known to be related to active enhancer regions (Creyghton et al. 2010). H3K4Me3 which is related with promoter regions also showed an overrepresentation in CNSs. The number of CNSs that overlapped with H3K27ac regions advances with the development stage whereas for H4K4Me1, many CNSs overlap with this chromatin mark at very early stage of development such as the dome stage. Similarly nematode CNSs also showed an overrepresentation with regards to H3K4Me1, H3K27ac at early embryo stage. Also many CNSs overlapped with H3K4Me3 regions during early development stage when compared to later stages such as L3 stage or the young adult (Appendix A14-B).

3.3.7 The predicted target genes for CNSs and functional classification

The closest genes to the CNSs were considered as the likely target genes. The target genes were determined based on the reference genomes used in the analysis. The gene ontology analysis was performed based on the likely target genes. The GO analysis for the likely target genes showed that Diptera and Nematode CNS-associated genes were enriched in transcription regulation and

DNA binding (Table 3.2). This analysis was only done for these two invertebrate groups as the gene ontology data was not available for other groups of interest. The GO analysis for bird CNS-associated genes showed a pattern similar to invertebrates. The highly enriched GO terms were related to regulation of transcription, regulation of RNA metabolic processes and DNA binding whereas the most underrepresented was related to certain receptor classes and enzyme activity related proteins. The GO analysis could only be determined for the Diptera, Nematode and teleost fish due to limited availability of data for other genomes (Table 3.2)

(A) Diptera CNSs

Functional group	**P-value**
DNA binding	5.7E-40
Transcription factor activity	5.1E-35
Functions related with nucleus	3.5E-31
Transcription regulator activity	9.2E-31
Homeobox domain related	1.2E-29
Regulation of transcription	2.0E-26
Developmental protein	1.2E-15
Winged helix repressor DNA-binding	5.5E-13

(B) Nematode CNSs

Functional group	**P-value**
DNA binding	7.2E-61
Functions related to nucleus	3.6E-58
Homeobox related	4.5E-53
Regulation of transcription	5.1E-52
Regulation of RNA metabolic processes	6.4E-52
Transcription factor activity	3.6E-42
Transcription regulator activity	1.0E-39
Developmental protein	2.4E-25
Helix-turn-helix motif, lambda like repressor	8.1E-13

(C) Teleost fish

Functional group	**P-value**
Regulation of transcription	7.5E-89
Regulation of RNA metabolic processes	9.1E-78
Transcription factor activity	6.8E-75
DNA binding	1.6E-74
Functions related to nucleus	1.4E-38
Positive regulation of gene expression	4.1E-29
Regulation of macromolecule metabolic processes	6.5E-24
Nucleoplasm related	7.7E-22
Homeodomain-related	5.3E-21

Table 3.2 - Gene ontology (GO) analysis for (A) Diptera (B) Nematode and (C) teleost fish CNSs. The closest gene to the CNSs was considered as the likely target gene for this analysis.

3.3.8 Transcription factor binding site analysis for vertebrates

The human transcription factor binding site data from UCSC were used as the vertebrate reference in determining the binding site characteristics for non-mammalian vertebrates. I found that many of the binding sites for ubiquitous transcription factors are GC rich (Appendix A15). For example SMC3, SP1, USF2 and ATF1 are known to be ubiquitous transcription factors and they have higher than average genomic GC content: 49.87%, 51.57%, 51.95%, and 49.74%, respectively.

Transcription factors such as SP1 and ATF specifically bind to sites that are overrepresented in housekeeping gene promoter regions (Farréal et al. 2007). The GC content for CNSs I found for non-mammalian vertebrates were lower than the genomic average for the noncoding region of the reference genome. Many of the underrepresented binding sites in CNSs are GC rich and are also related to ubiquitous transcription factor binding sites. The tissue specific binding sites were overrepresented compared to the ubiquitous binding sites. For example the overrepresented binding sites such as SETDB1 are found to be related to repression of genes encoding developmental regulators and help to maintain the embryonic stem cell state (Bilodeau et al. 2009). Another overrepresented binding site was for transcription factor ZNF263 which is found to be related to regulation of cell growth, cell differentiation and development according to Okubo et al. (1995).

Several other transcription factors such as MAFs, MAFF and MAFK are considered to be important in gene expression in mammals (Kannan et al. 2012). MAFF and MAFK correspond to two low GC binding sites which are overrepresented in the set of CNSs for non-mammalian vertebrates. MAFK has a function in neuronal differentiation and in general Maf family transcription factors are considered to be regulators of tissue specific gene expression (Kataoka2006).

In general one evident pattern I observed is that the GC poor binding sites are related to tissue specific gene expression, regulation of transcription and development, whereas high GC binding sites correspond to ubiquitous activity.

3.3.9 Transcription factor binding sites in plants

In this analysis I mainly focused on the *A. thaliana* transcription factor binding sites. Out of the 26 binding sites I analyzed the low GC binding sites (13/26) mostly seem to be related to ubiquitous activity and the high GC sites appear to facilitate binding of transcription factors related to transcription and development and tissue specific expression (Appendix A16). The binding sites such as AP1, AP2, SEP3, SOC1, LFY, GL1, and GLT1 are related to ubiquitous expression and they have GC lower than the genomic average GC for the noncoding region. Binding sites such as PRR5, PRR7, PIF4, PIF5, TOC1, and FUS3 have larger than average GC content (42.07%, 42.51%, 38.77%, 39.72%, 46.56%, and 36.84%, respectively) and are related to transcription regulation and development.

This finding is opposite to what was observed for the vertebrate CNSs. That is, the vertebrates binding sites related to regulation of transcription and development and tissue specific expression were GC poor. It appears that vertebrates and plants have formulated different sequence preferences when it comes to binding sites related to tissue specific and ubiquitous binding. This in a way explains the heterogeneity I observed for high GC plant CNSs and GC poor vertebrate CNSs. Even though the CNSs seem to be related to same GO function in both lineages, their sequence preferences seem to differ. These analyses were restricted to A. thaliana and human, as a comprehensive transcription factor binding site data is not available for other organisms under study. I did not perform the transcription factor binding site overrepresentation analysis for plant CNSs, as the number is too low for any statistical inference.

3.3.10 The GC content transition in the vertebrate lineage

In order to explain the origin of the GC content heterogeneity, I decided to look into the evolutionary dynamics of the transcription factors. Since I found that the CNSs are overrepresented in TF binding sites (Appendix A17), I obtained a cue that the transcription factor binding site evolution may have played an important role in the evolutionary dynamics of CNS GC content. Upon closer examination of the TF binding site data for vertebrates (human as the reference), I found that ubiquitous TF binding sites have higher GC than the tissue specific binding sites when compared with the genomic GC content (Appendix A15). In comparison, I observed that plant TF binding sites follow the opposite pattern, whereby plant tissue specific binding sites were GC rich and the ubiquitous binding sites were found to be GC poor when compared with the genomic average (Appendix A16). This heterogeneity in GC content with regards to the CNSs in different lineages might be attributable to the tissue specific TFs. After testing 150 vertebrate TFs shared with *A. thaliana, Oryza sativa* and *C. briggsae* protein coding genes, I found that vertebrate tissue specific TFs are more lineage specific to vertebrates than the ubiquitous TFs. The tissue specific TFs were significantly less shared than the ubiquitous TFs (Figure 3.7). The lineage specific features should come from tissue specific TFs that are not shared across lineages. Since vertebrates evolved more tissue specific TFs that are lineage specific which are GC poor, in turn more binding sites that are GC poor, they show the characteristic feature of low GC CNSs among other lineages.

As for a high GC lineage, I expected that they should have evolved higher GC TFs with high GC binding sites. To this end I tested the *A. thaliana* TF genes against all human, chicken and fugu annotated genes. However, I found no significant difference in conservation of ubiquitous and tissue specific TFs with regards to plants. This implies that underrepresentation of tissue specific TFs among conserved TFs is a specific feature to the vertebrate lineage (Figure 3.8).

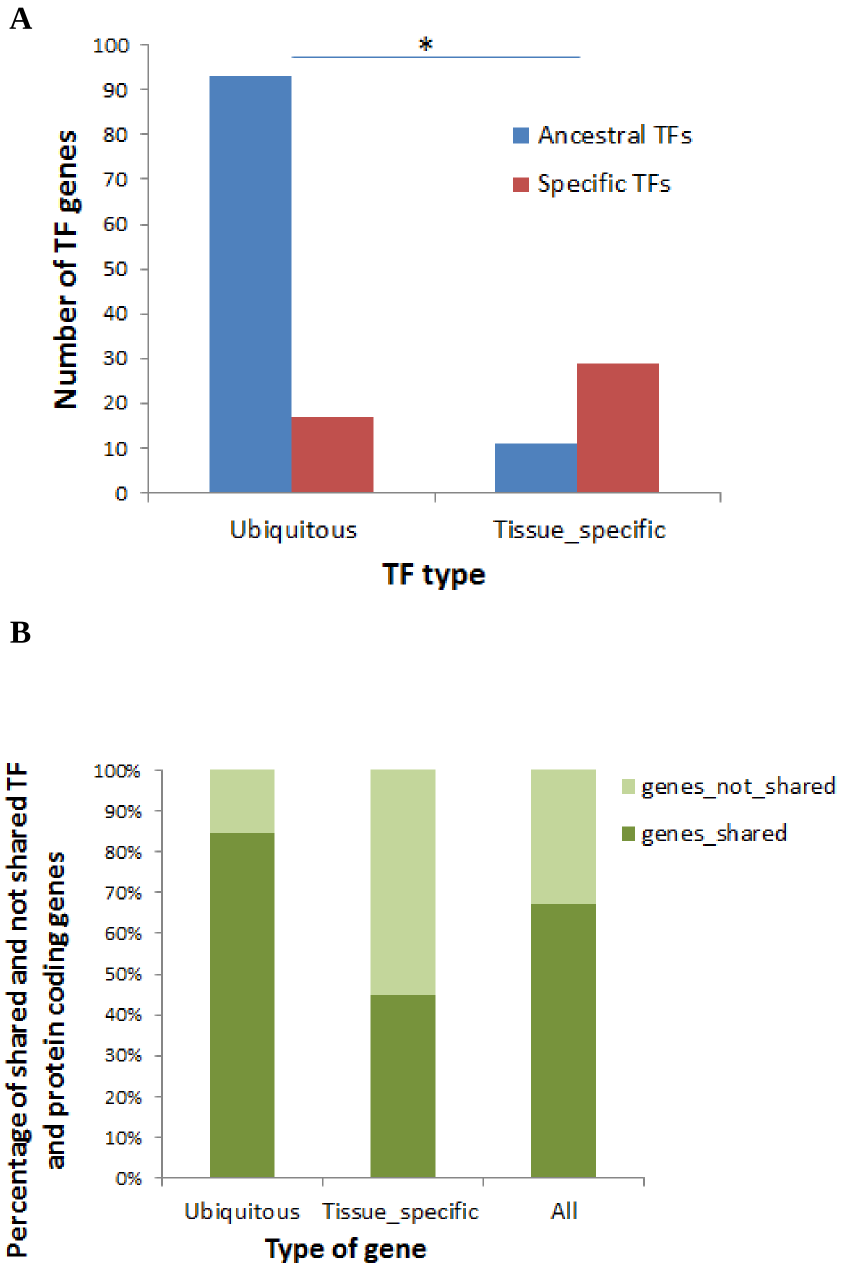
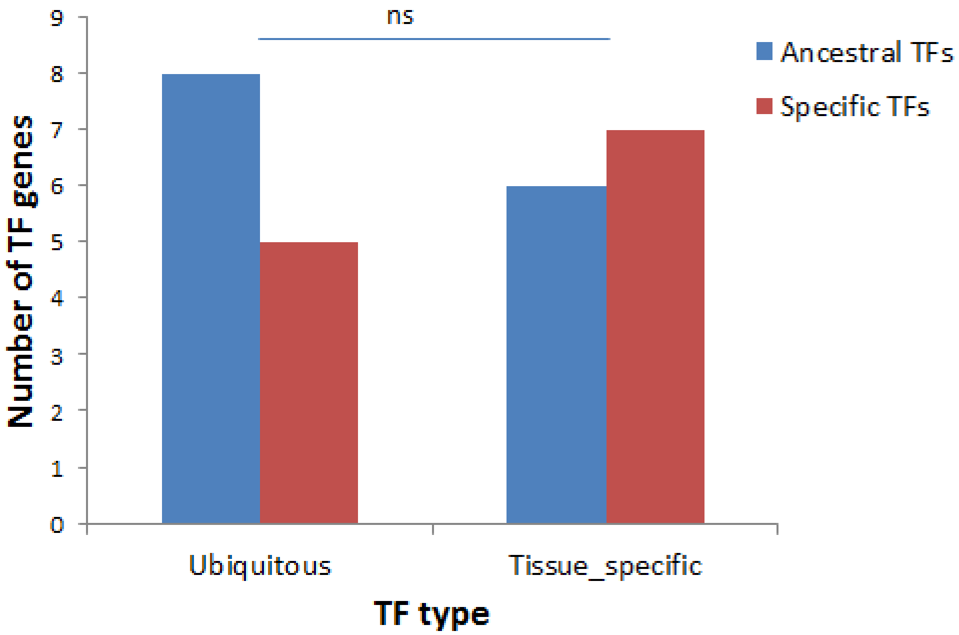


Figure 3.7 - (A) Vertebrate (*H. sapiens*) tissue specific TFs are less shared in other lineages. The vertebrate TF genes (150) that are shared in *A. thaliana, O. sativa, and C. briggsae* are designated as ancestral whereas TF genes which are not shared with respect to homology are designated as specific TFs. The difference between the ancestral and specific TFs is significant at p

= 0.05 (fisher exact test). **(B) Comparison of shared ubiquitous and tissue specific TFs with all protein coding genes of human conserved in *A. thaliana, O. sativa, and C. briggsae.*** In comparison with the random expectation as shown in 3rd bar (All) of the histogram the ubiquitous TFs are significantly more conserved while the tissue specific TFs are significantly less conserved

A


ns
Ancestral TFs
Specific TFs
Number of TF genes
0
1
2
3
4
5
6
7
8
9
Ubiquitous
Tissue_specific
TF type

B

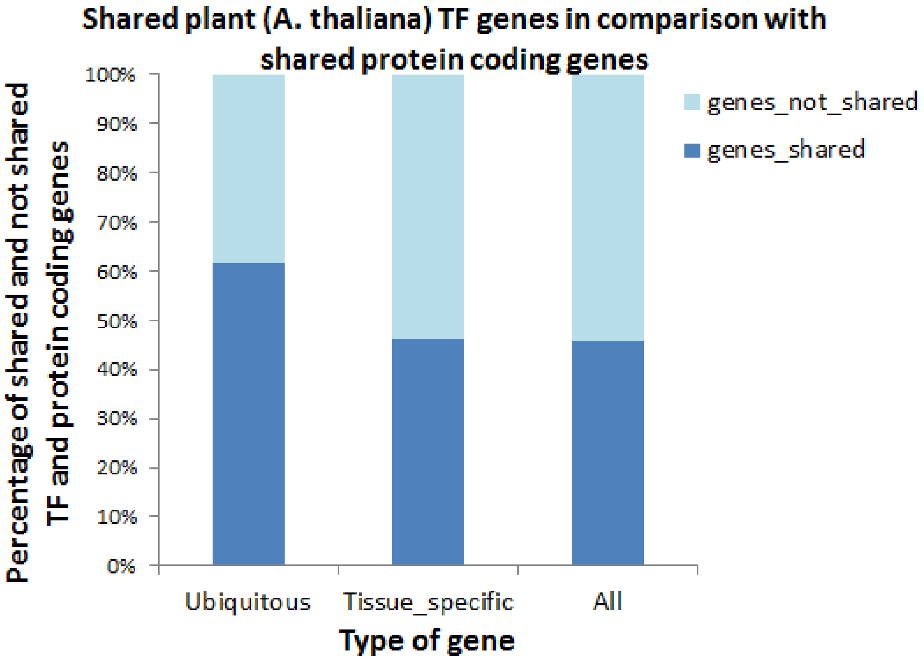

Shared plant (A. thaliana) TF genes in comparison with shared protein coding genes
genes_not_shared
genes_shared
Percentage of shared and not shared TF and protein coding genes
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Ubiquitous
Tissue_specific
All
Type of gene

Figure 3.8 - (A) Plants do not show any significant difference in sharing of tissue specific and ubiquitous TF binding sites with regards to other species (*H. sapiens, G. gallus, and T. rubripes*).The plant (A. thaliana) TF genes (26) that are shared in *H. sapiens, G. gallus, and T. rubripes* are designated as ancestral whereas TF genes which are not shared with respect to homology are designated as specific TFs. The difference between the ancestral and specific TFs was statistically tested with Fisher's exact test. **(B) Comparison of shared ubiquitous and tissue specific TFs with all protein coding genes of human conserved in *H. sapiens, G. gallus, and T. rubripes.*** In comparison with the random expectation as shown in 3rd bar (All) of the histogram the ubiquitous TFs are significantly more conserved while the tissue specific TFs in plants show no significant difference from the random expectation.

3.4 Discussion

I identified lineage-common conserved noncoding regions for fungi, invertebrate and non-mammalian vertebrate genomes that are shared among organisms in a particular lineage. The GC contents for the CNSs differed among lineages. The fungal and invertebrate CNSs were generally GC rich, whereas non-mammalian vertebrate CNSs were GC poor. In my previous study (Hettiarachchi et al. 2014) I found that plant CNSs are also GC rich, showing similar characteristic as fungal and invertebrate CNSs. However Babarinde and Saitou (2013) reported that mammalian CNSs were GC poor and their result showed similarity to the non-mammalian vertebrate CNSs I identified in this analysis. This low GC content in CNSs therefore appears to be a general feature that is shared by vertebrates. This result suggests that there seems to be a sudden transition of GC content preference in CNSs or in other words potential regulatory elements from plants, fungi and invertebrates to vertebrates. Next question I addressed was what could be the plausible reason for this observed transition. To this end I tried to determine the sequence properties of different transcription factors. I discovered that in vertebrates the transcription factor binding sites related to tissue specific expression, transcription and development are GC poor and binding sites for ubiquitous transcription factors such as SP1, NRF1, and E2F6 are GC rich.

The above mentioned pattern I observed for GC content for transcription factor binding sites in vertebrates was not observed when I examined the transcription factors of plants. The ubiquitous transcription factors for plants seemed to be GC poor whereas the tissue specific and plant development and transcription regulation associated transcription factors are GC rich. This goes well in line with my observation for plant CNSs (Hettiarachchi et al. 2014) being GC rich and the identified CNSs in my previous analysis showed a highly enriched GO for transcription regulation and development.

In order to explain the origin of heterogeneity I decided to look into the evolutionary dynamics of the transcription factors. After examining the TF binding site data for vertebrates (human as the reference) I found that ubiquitous TF binding sites have higher GC compared to tissue specific binding sites with respect to the genomic GC content. In contrast, plant TF binding sites followed an opposite pattern where the tissue specific binding sites were GC rich and the ubiquitous binding sites were found to be GC poor. And this heterogeneity in GC content with regards to the CNSs in lineages might be attributable to the tissue specific TFs, and in fact I found that vertebrate tissue specific TFs are more lineage specific than the ubiquitous ones. and that the tissue specific TFs were significantly less shared than the ubiquitous TFs. The lineage specific features should come from tissue specific TFs that are not shared across lineages. At this point it becomes evident that since vertebrates evolved more tissue specific TFs that are lineage specific which are GC poor and more binding sites that are GC poor, the overall CNS GC content is also low.

Even though I expected that the high GC lineages to have evolved more high GC TFs with high GC binding sites, the reality was different. After testing *A. thaliana* TF genes against human, chicken and fugu for all annotated genes, I found no significant difference in conservation of ubiquitous and tissue specific TFs with regards to plants. This means that underrepresentation of tissue specific TFs among conserved TFs is a specific feature to the vertebrate lineage.

I also found that the GC content of the CNSs had a direct relation to the location of the CNSs in the genome. The low GC CNSs showed a higher probability to be located in open chromatin regions whereas high GC CNSs tend to be located in clearly positioned nucleosomes. The location of the CNSs is important since the CNSs located in open chromatin regions are easier to be accessed by the transcription factors, whereas the ones with high nucleosome occupancy are harder to be accessed due to the coiled nature of the region. This structural architecture of the CNSs should have a direct impact on the binding of the proteins to that region. Several studies have found that some binding sites actually require being located in coiled nucleosome regions for its proper regulation. For example, Cirillo and Zaret (1999) reported that HNF3 liver enriched transcription factor binds to albumin gene enhancer region which is clearly located inside a nucleosome region. Binding of HNF3 stabilizes the nucleosome position which results in a very stable binding complex. Similarly TP53 transcription factor which is known to be a roadblock against cancer also binds to a high nucleosome occupancy region (Nili et al. 2010). Experimental evidence is needed to determine the function and the binding properties of the CNSs.

The GO analysis for nematode, Diptera and bird CNSs target genes showed a high enrichment for regulation of transcription and development, transcription factor activity, and DNA binding, among others. This goes in line with numerous findings that are already documented and experimentally verified that CNSs are related to transcription regulation and development (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Babarinde and Saitou 2013; Haudry et al. 2013; Hettiarachchi et al. 2014).

In conclusion, I observed a GC content heterogeneity of the CNSs belonging to diverse lineages. Further I discovered that the CNS GC content transition occurred from GC rich state of plants, fungi and invertebrates to GC poor state in the vertebrate lineage due to GC poor binding sites that are lineage specific in vertebrates. Also I found that the GC poor vertebrate CNSs are specifically located in stretches of GC rich isochore-like regions. The CNS GC content is closely related with the location of the CNSs in the genome. The GC poor CNSs have a tendency to be located in open chromatin regions where as the high GC CNSs have a higher probability to be located in heterochromatin regions. Also the substitution patterns of CNSs are generally from GC to AT in other lineages such as invertebrate and non-mammalian vertebrates except for fungi. The Diptera CNSs behaved similar to vertebrate CNSs with respect to GC content. This is one enigmatic aspect in my findings which still needs further investigation.

Chapter 4

Conclusion

The conserved noncoding sequences have been extensively studied over twenty years for its importance in potential gene regulatory functions. Comparative genomics approaches which are guided by extensive computation have paved way to identify such potential regulatory regions prior to experimental verification. The functional aspects of CNSs have also been experimentally verified upon identification by computational methods (Lee et al. 2010; Clarke et al. 2012). Different thresholds have been considered in identifying and classifying CNSs. Bejerano et al. (2004) identified ultraconserved elements and later on numerous groups identified highly conserved noncoding elements and ultraconserved like elements including in plant genomes (Krtisas et al. 2012). As we know organisms in different lineages carry diverse morphological and physiological characters which sometimes are signature features of the clade. For example incisors of rodents, placenta formation in eutherian mammals, single cotyledon formation in monocots and double cotyledon in eudiocots are some of the lineage specific features that we observe in nature. Also we are in the understanding that most protein coding genes are conserved across lineages and these lineage specific features in organisms are therefore unlikely to have risen from coding regions that are shared across lineages. Therefore lineage specificity should rise from genomic elements that are not shared across lineages but are specific. These genomic elements that govern lineage specificity are regulatory elements that are restricted to different lineages (Sun et al. 2006; Sumiyama et al. 2012) .There have been studies that specifically dealt with lineage specificity with regards to conserved noncoding sequences. Takahashi and Saitou (2012) reported primate and rodent CNSs and Babarinde and Saitou (2013) identified lineage specific noncoding elements of primates, rodents, carnivores and cetariodactyls. Even though there had been several studies on plant CNSs, plant noncoding sequences are not been extensively studied. My first study, already published as Hettiarachchi et al. (2014), was an attempt to identify and characterize lineage specific CNSs in plant genomes for the first time in plant comparative genomics.

In this study I have addressed two aspects regarding CNSs. First aspect was the lineage specific CNSs in plant genomes and the latter was regarding the GC content heterogeneity of CNSs in different lineages.

My first study covered in chapter 2 dealt with lineage specific CNSs in plant genomes. This is one aspect that had not been studied before on plant genomes and this study covered the lineage specificity with regards to conserved noncoding sequences in several plant groups, namely eudicots, monocots, grasses, angiosperms, land plants and plants. The protein coding genes are found to be conserved even across lineages that are highly diverged and have varying morphological and physiological features. Thus the morphological and physiological diversity in different species could not have risen from shared genomic elements. Regulatory elements is the governing factor Here I identified 27 eudicot, 6536 grass, 204 monocot, 19 angiosperm, 2 vascular plant specific CNSs that are likely to have originated in their respective common ancestors. In this analysis I found that generally grasses and monocots have more CNSs than eudicots. This comparatively high number of lineage specific CNSs in grasses and monocots is not due to the difference in number of species considered for the two lineages, evolutionary rate differences or relatively short divergence time in grass species used in the analysis. Even with reanalyzes considering the same number of species and similar divergence times of eudicots and monocots, the number of eudicot specific CNSs remained less than the monocot specific CNSs. Therefore the difference in the number of lineage specific CNSs is governed by a factor which is other than the above tested factors. Also the lineage specific loss of ancestral CNSs analysis showed that eudicot common ancestor has lost twice as much ancestral CNSs than the monocot or grass common ancestors. In order to test if the evolutionary rate could have been the reason for the difference in number of CNSs in lineages, I determined the lineage specific genes, with the expectation that if evolutionary rate had been a major governing factor both lineage specific genes and CNSs would follow the same pattern. Surprisingly I found that lineage specific genes followed the opposite pattern to lineage specific CNSs, that is eudicots had more lineage specific genes than CNSs and conversely grasses had less lineage specific genes and more lineage specific CNSs. This finding gives clarity that other than evolutionary rate there should be other factors that contribute to difference in the number of CNSs in lineages.

Examination of the location of the CNSs in the genome showed that the grass specific and monocot specific CNSs located in the UTR are significantly higher than the expectation from the genomic coverage for UTR regions. The eudicot and angiosperm CNSs also followed the same pattern irrespective of their significantly low number.

The gene enrichment analysis for the predicted likely target genes for the lineage specific CNSs indicated that these genes are predominantly associated with regulation of transcription and development. The most underrepresented GO terms for likely target genes of CNSs were related to housekeeping and defense related genes. Conversely the most prominent GO term for lineage specific genes were plant defense related. The finding suggests that lineage specific CNSs and genes are working in two arenas for proper regulation and development.

In order to identify special characteristics of the flanking regions of CNSs with regards to nucleotide composition I tested the AT content distribution of CNSs and their flanking regions. There was a statistically significant difference between the flanking regions and the CNSs with respect to the A+T content. A decline in the A+T content was observed near the immediate border of the grass and monocot specific CNSs. This pattern had been reported before for animal CNS studies (Walter et al. 2005; Vavouri et al. 2007) but had not been reported for lineage specific CNSs regarding plant genomes. Similar AT drop was observed for the eudicot specific CNSs with a significant difference between the flanking regions and the inside of CNSs.

The AT poor regions have high propensity to form nucleosome regions (Jansen and Verstrepen 2011) and also low AT content is related with recombination hot spots (Spencer et al. 2006). Here I found that the eudicot specific CNSs do not overlap with recombination hotspots as expected and the nucleosome occupancy prediction analysis showed that grass and eudicot specific CNSs to be located in well positioned nucleosomes.

In chapter three I tried to clarify and investigate aspects of GC content heterogeneity of CNSs in different lineages. In my previous study I determined the plant lineage specific CNSs and found that plant CNSs are GC rich. Similarly Babarinde and Saitou (2013) found that mammalian CNSs are GC poor. This GC content heterogeneity and where the GC poor CNSs originated in line of evolution is the next question I tried to address. In this study I considered several lineages namely fungi, invertebrates and non-mammalian vertebrates to obtain a eukaryote wide perspective of the CNSs. The fungi and invertebrate CNSs were predominantly GC rich similar to plant CNSs whereas non-mammalian vertebrate CNSs were GC poor similar to mammalian CNSs identified by

Babarinde and Saitou (2013). This GC poor character with regards to CNSs appears to be shared by vertebrates in general. In order to explain this transition in GC content of CNSs I examined the evolutionary dynamics of transcription factor binding sites of low GC vertebrates and high GC plants. In vertebrates the tissue specific binding sites were GC poor whereas ubiquitous binding sites were GC rich. The opposite followed for plant transcription factor binding sites, where the tissue specific binding sites were GC rich and the ubiquitous binding sites were GC poor. I found that the vertebrate tissue specific transcription factor binding sites are more lineage specific and are significantly less shared than ubiquitous ones. For the plant TF genes I found no significant difference in conservation of ubiquitous and tissue specific TFs implying that underrepresentation of tissue specific TFs among conserved TFs is a specific feature to vertebrate lineage. This low GC feature therefore has risen in vertebrate lineage due to enrollment of more tissue specific TFs that are GC poor thus more binding sites that are GC poor resulting in overall low GC CNSs.

Another unique feature for non-mammalian vertebrate CNSs observed was that they are located in GC rich isochore-like regions and the CNSs were the only GC poor region in a long stretch of genomic DNA.

Also related to the GC content I found that the low GC CNSs are located in open chromatin regions and high GC CNSs to have higher propensity to be located in heterochromatin regions. The location of the binding sites is important as open chromatin binding sites are supposedly easier to be accessed, whereas binding sites in heterochromatin regions are not easily accessible. But it is known that some binding sites such as HNF3 and TP53 (Cirillo and Zaret 2002; Nili et al. 2010) require restricted accessibility there they require to be located in heterochromatin regions for proper functionality.

I also tried to determine the histone modification signals related with CNSs, as it is considered as a signature of functionality for the CNSs. Nematode and teleost CNSs showed an overrepresentation for H3K4Me1, H3K27ac at early development stage compared to the random expectation. These modification signals that are known to be related with active enhancer regions imply these CNSs may act as enhancers during early development of the organism. Also the predicted target genes of Diptera, nematode and teleost fish CNSs were found to be related to transcription and development.

One important aspect that needs further clarification and further analyses is the similarity between vertebrate and Diptera CNSs with regards to GC content. This enigmatic feature needs further investigation.

References

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

Anderson CL, Bremer K, Friis EM. 2005. Dating phylogenetically basal eudicots using rbcl sequences and multiple fossil reference points. Am J Bot 92(10):1737-1748.

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C. et al. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes.* Proc Natl Acad Sci 92:1684-1688.

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G. et al. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences.Nucleic Acids Res40 (Web Server issue) W580–W584.

Babarinde IA, Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biol Evol doi:10.1093/gbe/evt177.

Banks JA, et al. 2011. The Selaginella genome identifies genetic changes associated with the evolution of vascular Plants. Science 332:960-963.

Bai L, Morozov AV. 2010. Gene regulation by nucleosome positioning. Trends Genet 26(11):476-483.

Bailey TL, et al. 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res 37(2):202-208.

Baxter L, Jironkin A, Hickman R, Moore J, Barrington C. et al. 2012. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell. 24: 3949–3965.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ. et al. 2004. Ultraconserved elements in the human genome. Science 304(5675):1321-1325.

Bennetzen JL, et al. 2012. Reference genome sequence of the model plant Setaria.Nat Biotechnol 13:36(6):555-61. Doi:10.1038/nbt.2196

Bilichak A, Iinystkyy Y, Hollunder J, Kovalchuk I. 2011. The progeny of Arabidopsis thaliana plants exposed to salt exhibit changes in DNA methylation, histone modifications and gene expression. PLoS One 7(1): e30515. doi:10.1371/journal.pone.0030515.

Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. 2009. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes and Dev 23:2484-2489.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10:1213–1218.

Buske FA, Bodén M, Bauer DC, Bailey TL. 2009. Assigning roles to DNA regulatory motifs using comparative genomics. Bioinformatics 26 (7): 860-866.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila.* Mol Biol Evol 24:2222-2234.

Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M. et al. 2002. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell 9(2):279-289

Clarke SL, VanderMeer JE, Wegner AM, Schaar BT, Ahituv N. et al. 2012. Human Developmental Enhancers Conserved between Deuterostomes and Protostomes. PLoS Genet DOI: 10.1371/journal.pgen.1002852.

Clark AG. 2001. The search for meaning in noncoding DNA. Genome Res 11:1319-1320.

Cokus SJ. et al. 2008. Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452(7184):215-219.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. Genome Res 16:536-541.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW. et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts development state. Proc Natl Acad Sci 107:21931–21936.

Dennis G, et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4(5): P3.

D’Hont A, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213-217.

Drake JA, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 28:223-227.

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

Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7: 399–406

Elgar G. 2009. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. Brief Funct Genomic Proteomic 8(4):256-265.

Elgar G, Vavouri T. 2008.Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 24(7):344-352.

Farré D, Bellora N, Mularoni L, Messeguer X, Mar Albà M. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biology 8:R140 doi:10.1186/gb-2007-8-7-r140.

Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K. et al. 2012. Controls of nucleosome positioning in the human genome. PLoS Genet 8(11):e1003036. doi: 10.1371/journal.pgen.1003036.

Geiman TM, Muegge K. 2010. DNA methylation in early development. Mol Reprod Dev 22(2):105-113.

Guo H, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell 15:1143-1158.

Guo WJ, Ling J, Li P. 2008. Consensus features of microsatellite distribution: Microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. Genomics 93:323– 331.

Gutierrez-Arcelus et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife 2:e00523.

Grunewald W, Parizot B, Inze D, Gheysen G, Beeckman T. 2007. Developmental biology of roots: one common pathway for all angiosperms? Int J Plant Dev Biol I(2):212-225.

Grzybowska EA, Wilczynska A, Siedlecki JA. 2001. Regulatory functions of 3'UTRs. Biochem Biophyl Commun 288:291-295.

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

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M. et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genet 45:891-898.

Hebbes TR, Clayton AL, Thorne AW, Crane-Robinson C. 1994. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken β-globin chromosomal domain. EMBO J 13:1823–1830.

Heckman DS, et al. 2001. Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129-1133.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A. et al. 2009. Histone modifications athuman enhancers reflect global cell-type-specific gene expression. Nature 459:108–112.

Hettiarachchi N, Kryukov K, Sumiyama K, Saitou N. 2014. Lineage-Specific Conserved Noncoding Sequences of Plant Genomes: Their Possible Role in Nucleosome Positioning. Genome Biol Evol 6(9):2527-2542.

Heyndrickx KS, Vandepoele K. 2012. Systematic identification of functional plant modules through the integration of complementary data sources. Plant Physiol 159: 884-901.

Hiller M, Schaar BT, Bejerano G. 2012. Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res 1-14.

Horton MW, et al. 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet 44:212-216.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4(1):44-57.

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1-13.

Huh SU, Kim KJ, Paek KH. 2012. Capsicum annuum basic transcription factor 3 (CaBtf3) regulates transcription of pathogenesis-related genes during hypersensitive response upon Tobacco mosaic virus infection. Biochem Biophys Res Commun 417(2):910-7. doi: 10.1016/j.bbrc.2011.12.074.

Inada DC, et al. 2003. Conserved noncoding sequences in the grasses. Genome Res 13:2030-2041.

Janes DE, et al. 2010. Reptiles and Mammals Have Differentially Retained Long Conserved Noncoding Sequences from the Amniote Ancestor. Genome Biol Evol 3:102-113.

Jansen A, Verstrepen KJ. 2011. Nucleosome positioning in *Saccharomyces cerevisiae.* Microbial Mol Biol 75:301-320.

Jareborg N, Birrney E, Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 9: 815–824.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10:161-172.

Jürgens G. 1992. Pattern formation in the flowering plant embryo. Curr. Opin. Genet. Dev. 2: 567–570.

Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. 2005. Regulated chromatin domain comprising cluster of co- expressed genes in Drosophila melanogaster. Nucleic Acids Res 33:1435–1444.

Kannan MB, Solovieva V, Blank V. 2012. The small MAF transcription factors MAFF, MAFG and MAFK: Current knowledge and perspectives. Biochem Biophys Acta 1823:1841–1846.

Kaplan N, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Natue 458(7236):362-366.

Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. Proc Nat Acad Sci 99: 6147–6151.

Kataoka K. 2007. Multiple mechanisms and functions of maf transcription factors in the regulation of tissue-specific genes. J Biochem 141(6):775-81

Kerstetter RA, Hake S. 1997. Shoot meristem formation in vegetative development. Plant Cell 9: 1001–1010.

Kritsas K, et al. 2012. Computational analysis and characterization of ULE-like elements (ULEs) in plant genomes. Genome Res 22(12):2455-2466.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D. et al. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environmentat regulatory elements. Genome Res 22:1735-1747.

Lee Ap, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2010. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol 28(3):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(14):1725-1735.

Lipman DJ. 1997. Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res 25: 3580–3583.

Mabon SA, Misteli T. 2005. Differential recruitment of Pre-mRNA splicing factors to alternatively spliced transcripts. PLoS Biol 3(11):e374.

Marchais A, Naville M, Bohn C, Bouloc P, Gautheret D. 2009. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. Genome Res 19:1084–1092.

Marinotti O, Cerqueira GC, De Almeida LG, Ferro MI, Loreto EL. et al. 2013. The genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Res 41(15):387-400.

Matsunami M, Saitou N. 2012. Vertebrate Paralogous Conserved Noncoding Sequences May Be Related to Gene Expressions in Brain. Genome Biol Evol 5(1):140-150.

McEwen GK, et al. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet 5:1-9.

Nili EL, Field Y, Lubling Y, Widom J, Oren M. et al. 2010. p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. Genome Res 20(10):1361–1368.

Nishida H. 2012. Nucleosome positioning. ISRN Molecular Biology doi:10.5402/2012/245706.

Okubo K, Itoh K, Fukushima A, Yoshii J, Matsubara K. 1995. Monitoring cell physiology by expression profiles and discovering cell type-specific genes by compiled expression profiles. Genomics 30(2):178-86.

Parry G, Ward S, Cernac A, Dharmasiri S, Estelle M. 2006. The Arabidopsis suppressor of auxin resisitance proteins are nucleoporins with an important role in hormone signaling and development. The plant cell 18(7):1590-1603.

Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109.

Phillips T. 2008. The role of methylation in gene expression. Nature Education 1(1).

Polychronopoulos D, Sellis D, Almirantis Y. 2014. Conserved Noncoding Elements Follow Power-Law-Like Distributions in Several Genomes as a Result of Genome Dynamics. PLoS One DOI: 10.1371/journal.pone.0095437.

Reddy ASN, Day IS, Göhring J, Barta A. 2012. Localization and Dynamics of Nuclear Speckles in Plants. Plant Physiol 158(1):67-77.

Rensing SA, et al. 2007. The *Physcomitrella* Genome reveals evolutionary insights into the conquest of land by plants. Science 319:64-69.

Saisawang C, Ketterman AJ. 2014. Micro-plasticity of genomes as illustrated by the evolution of glutathione transferases in 12 Drosophila species. PLoS One 9(10):e109518 doi: 10.1371/journal.pone.0109518.

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

Sandelin A, et al. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99. doi:10.1186/1471-2164-5-99.

Seridi L, Ryu T, Ravasi T. 2014. Dynamic Epigenetic Control of Highly Conserved Noncoding Elements. PLoS One DOI: 10.1371/journal.pone.0109326

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

Shin JT, et al. 2005. Human- zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res 33(17):5437-45.

Spencer CCA, et al. 2006. The Influence of Recombination on Human Genetic Diversity. PLoS Genet 2(9):e148. doi:10.1371/journal.pgen.0020148.

Sussex IM. 1989. Developmental programming of the shoot meristem. Cell 56: 225–229.

Takahashi M, Saitou N. 2012. Identification and characterization of lineage-specific highly conserved noncoding sequences in Mammalian genomes. Genome Biol Evol 4(5):641-657.

Tang H, et al. 2008. Synteny and collinearity of plant genomes. Science 320(5875):486-488.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526.

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.

Tillo D, et al. 2010. High Nucleosome occupancy is encoded at human regulatory sequences. PLoS ONE 5(2): e9129. doi:10.1371/journal.pone.0009129.

Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res 18(7):1084-1091.

Tseral L, Kolde R, Rebane A, Kisand K, Org T. et al. 2010. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. BMC genomics 11:642 doi:10.1186/1471-2164-11-642.

The angiosperm phylogeny group 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society 141:399–436. doi: 10.1046/j.1095- 8339.2003.t01-1-00158.x.

Tyler L, et al. 2004. DELLA proteins and gibberellin-regulated seed germination and floral development in Arabidopsis. Plant Physio 135(2):1008-1019.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, et al. 2011. Determinants of nucleosome organization in primary human cells. Nature 474: 516–520.

Vavouri T, Mcewen GK, Woolfe A, Gilks WR, Elgar G. 2005. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet 22:5-10.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biology 8(2):R15.doiI:10.1186/gb-2007-8-2-r15.

Washietl S, Machne R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. Trends Genetv24:583–587.

Waterhouse PM, Wang MB, Lough T. 2001. Gene silencing as an adaptive defense against viruses. Nature 411:834-842.

Warnecke T, Batada NN, Hurst LD. 2008. The Impact of the Nucleosome Code on Protein-Coding Sequence Evolution in Yeast. PLoS Genet DOI: 10.1371/journal.pgen.1000250.

Woodhouse MR, Pedersen B, Freeling M. 2010. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Biol 6(5):e1000949. doi:10.1371/journal.pgen.1000949.

Woolfe A, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3(1): e7. doi:10.1371/journal.pbio.0030007.

Yang L, Zou M, Fu B, He S. 2013. Genome-wide identification, charachterization, and expression analysis of lineage-specific genes within zebrafish. BMC Genomics 14:65. doi:10.1186/1471-2164-14-65.

Zimmermann R, Werr W. 2005. Pattern formation in the monocot embryo as revealed by NAM and CUC3 orthologues from Zea mays L. Plant Mol Bio 58:669-685.

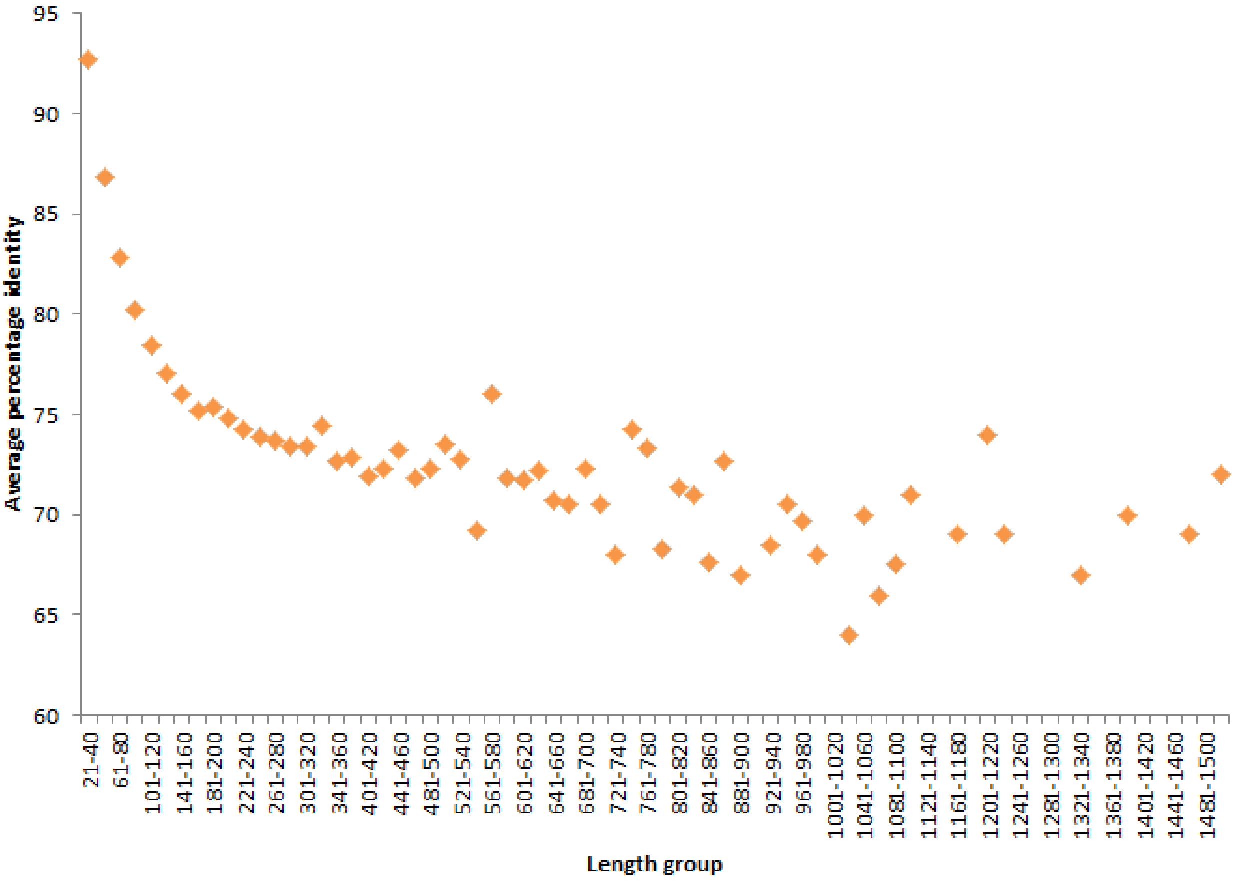
Zimmermann R, Werr W. 2007. Transcription of the putative maize ortho-logue of the Arabidopsis DORNROSCHEN gene marks early asymmetry in the proembryo and during leaf initiation in the shoot apical meristem. Gene Expression Patterns 7:158-164.

Appendix

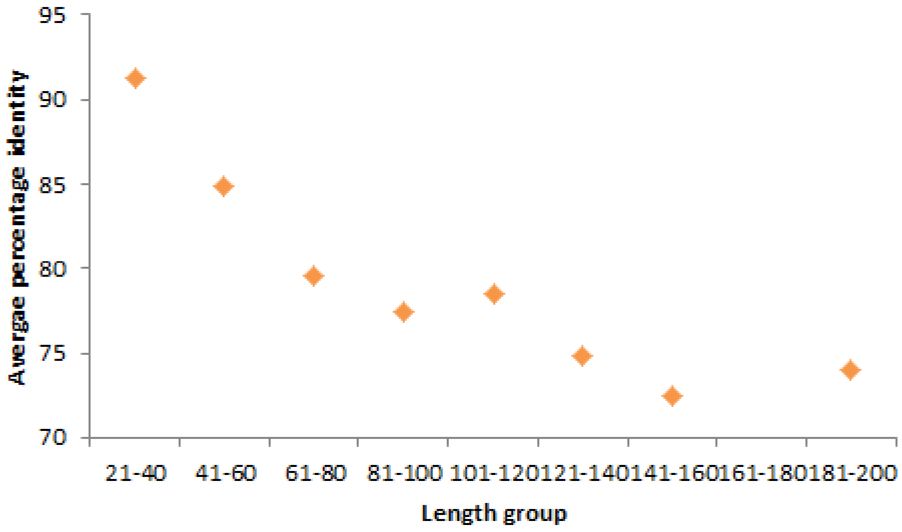
Appendix A1:

Length and conservation level relationship for lineage specific CNSs. (A) Length and conservation level for grass specific CNSs. (B) Length and conservation level for monocot specific CNSs. (C) Length and conservation level for dicot specific CNSs. (D) Length and conservation level for angiosperm specific CNSs.

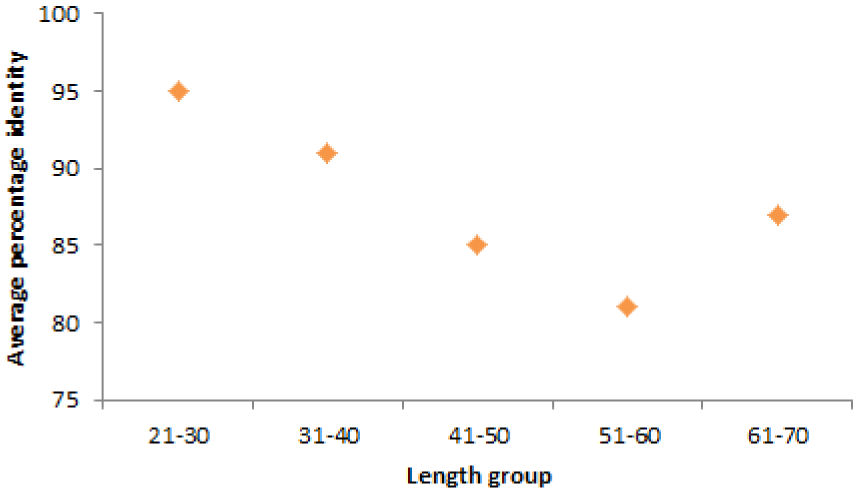
(A)



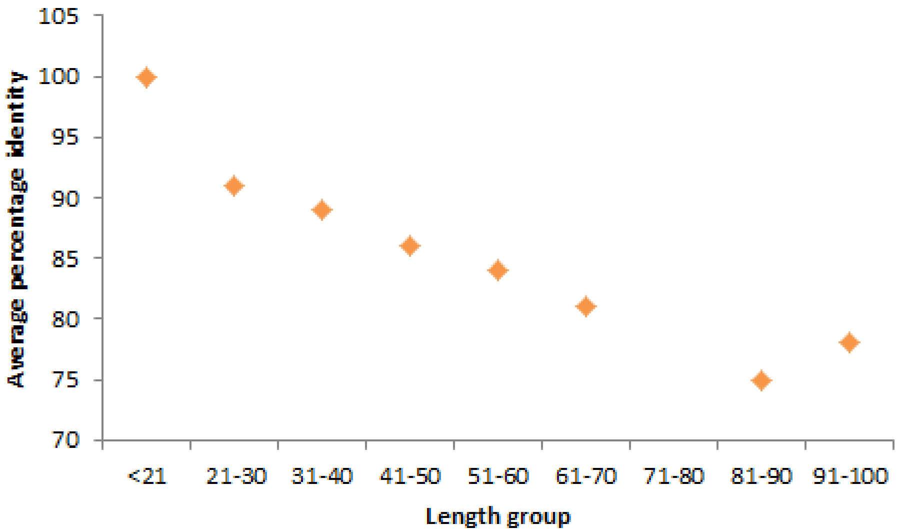
(B)


95
90
85
80
75
70
Awergae percentage identity
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
Length group

(C)

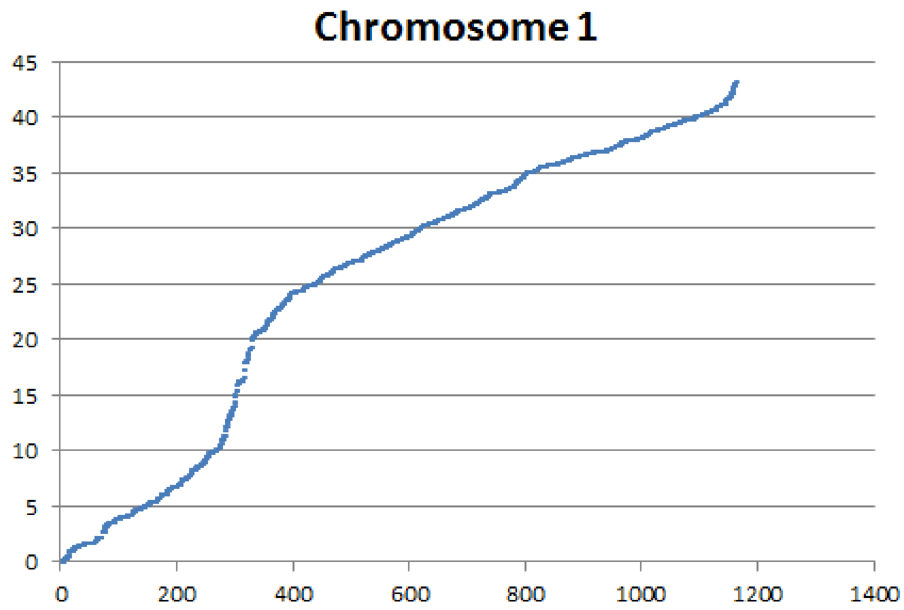

100
95
90
85
80
75
Average percentage identity
21-30
31-40
41-50
51-60
61-70
Length group

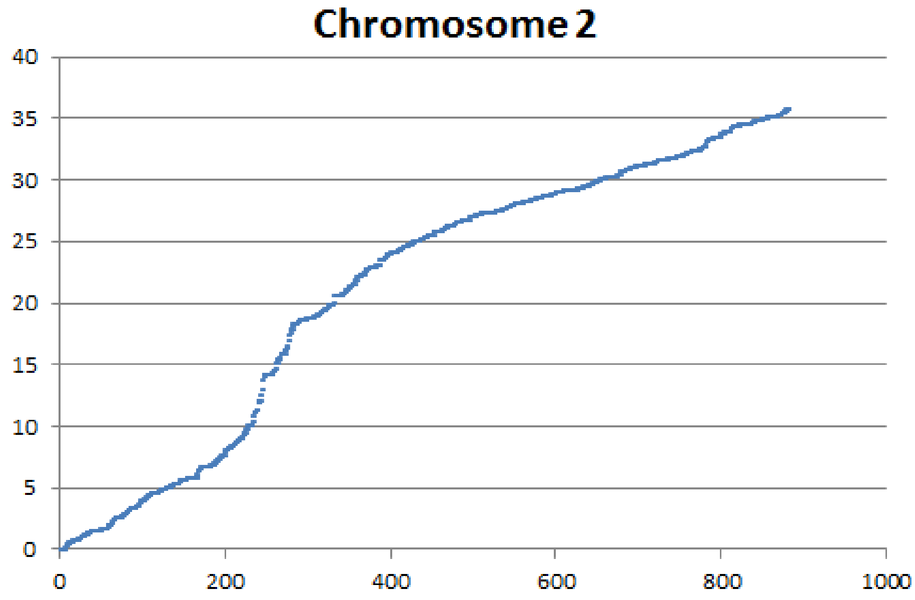
(D)

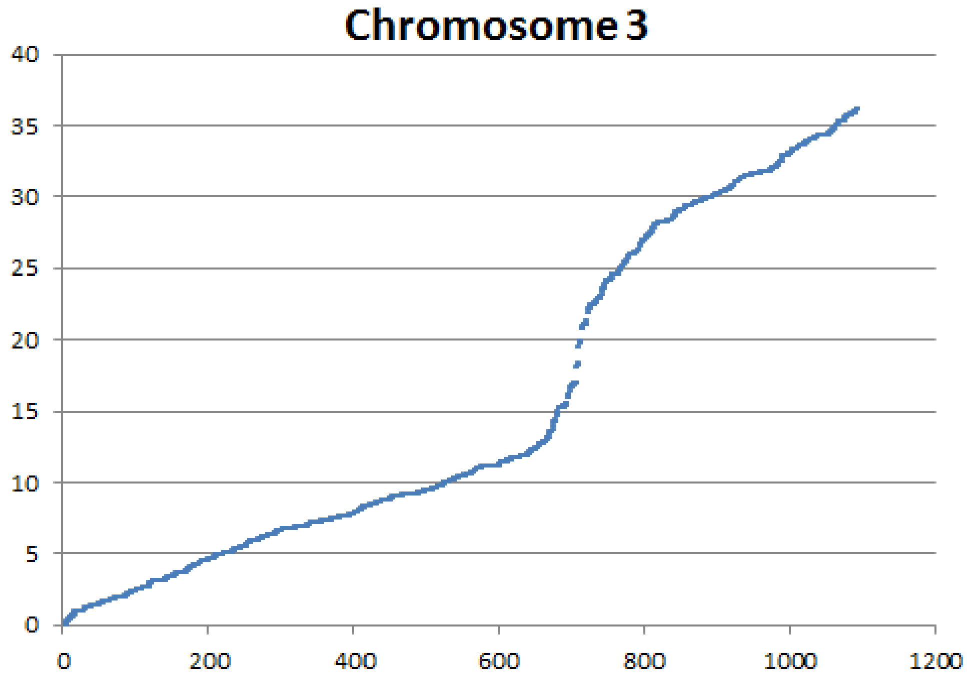

Average percentage identity
105
100
95
90
85
80
75
70
<21
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
Length group

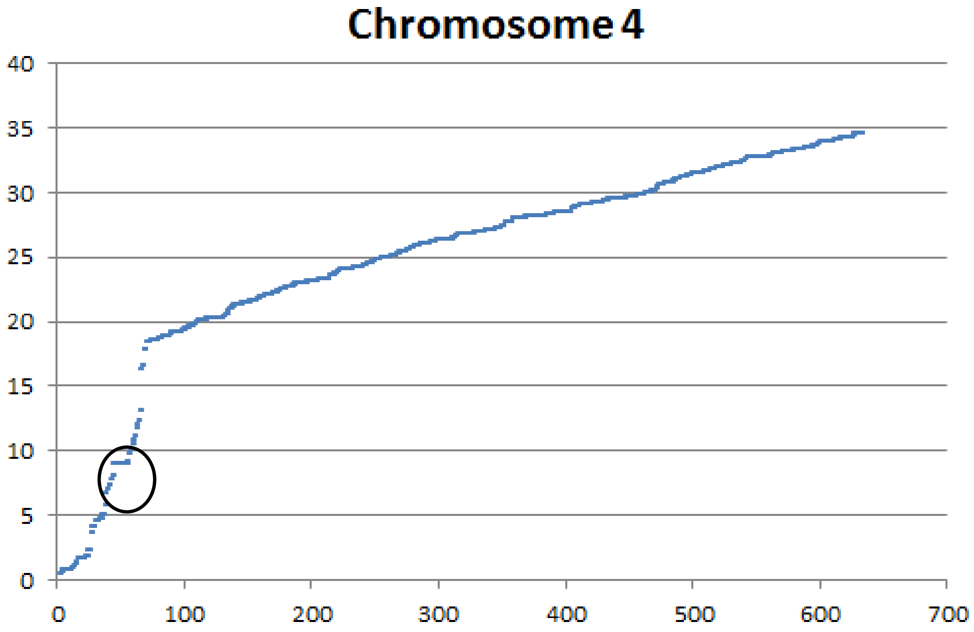
Appendix A2:

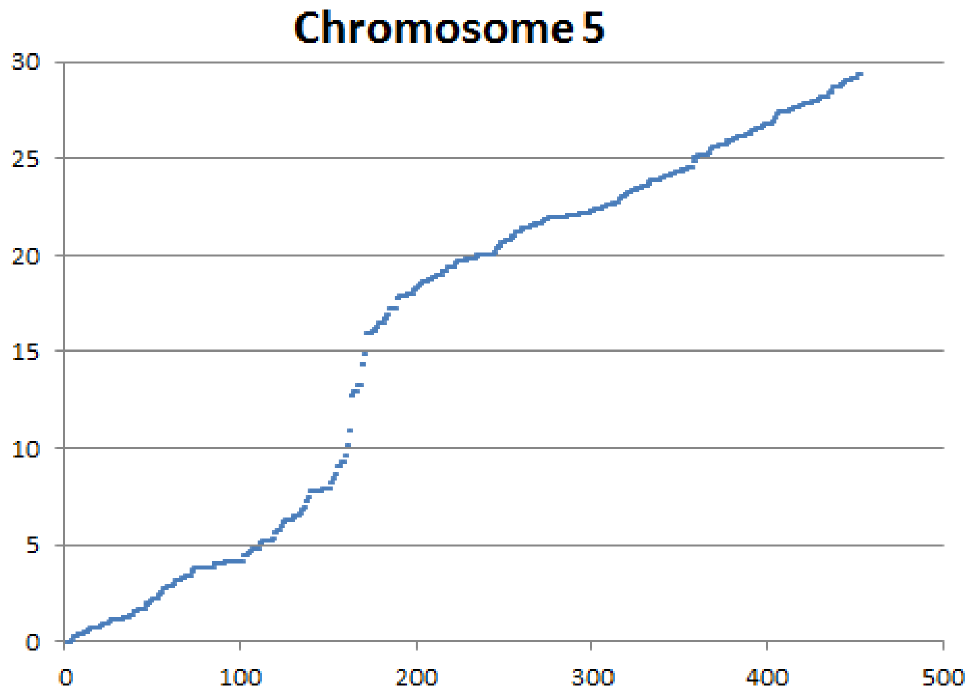
(A) Distribution of grass specific CNSs on *O.sativa japonica* chromosomes. Y axis – genomic location of the CNSs (Mb), X axis – cumulative frequency of the CNSs. The clearly visible clusters of CNSs are encircled.

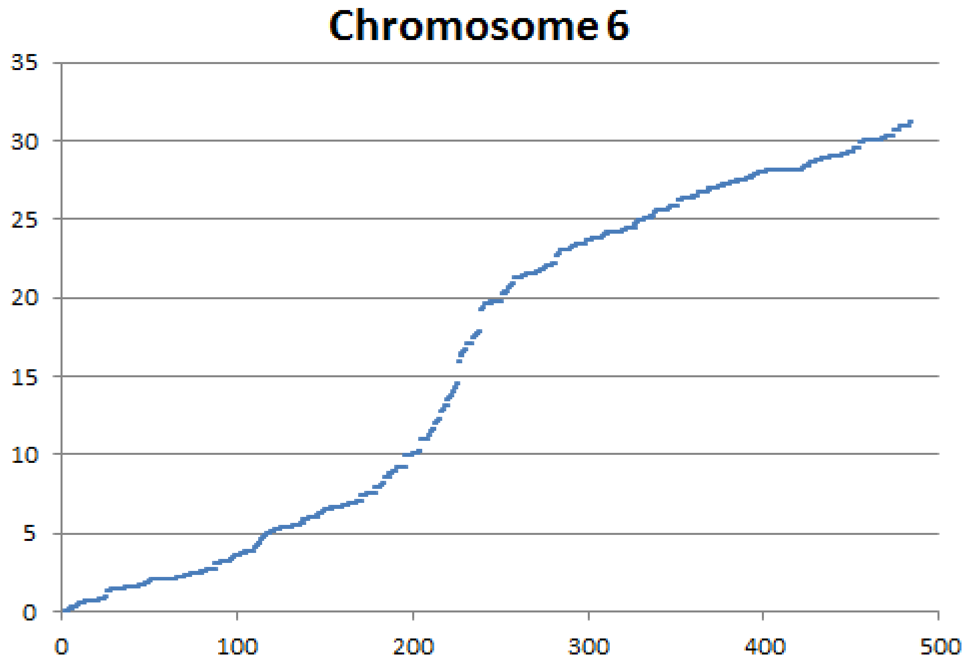


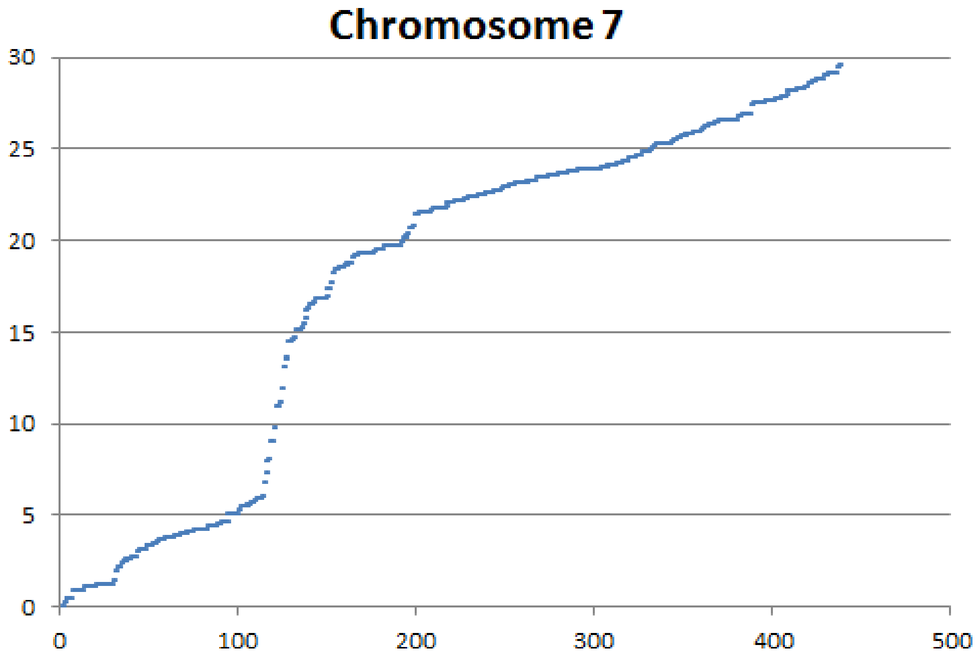


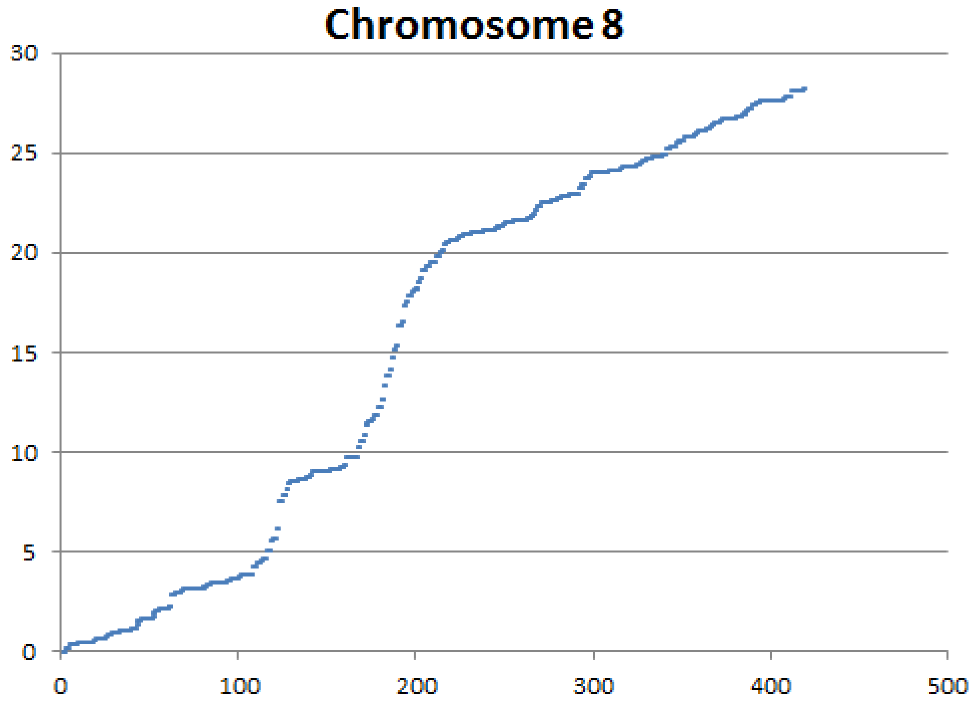

Chromosome 3
40
35
30
25
20
15
10
5
0
0
200
400
600
800
1000
1200

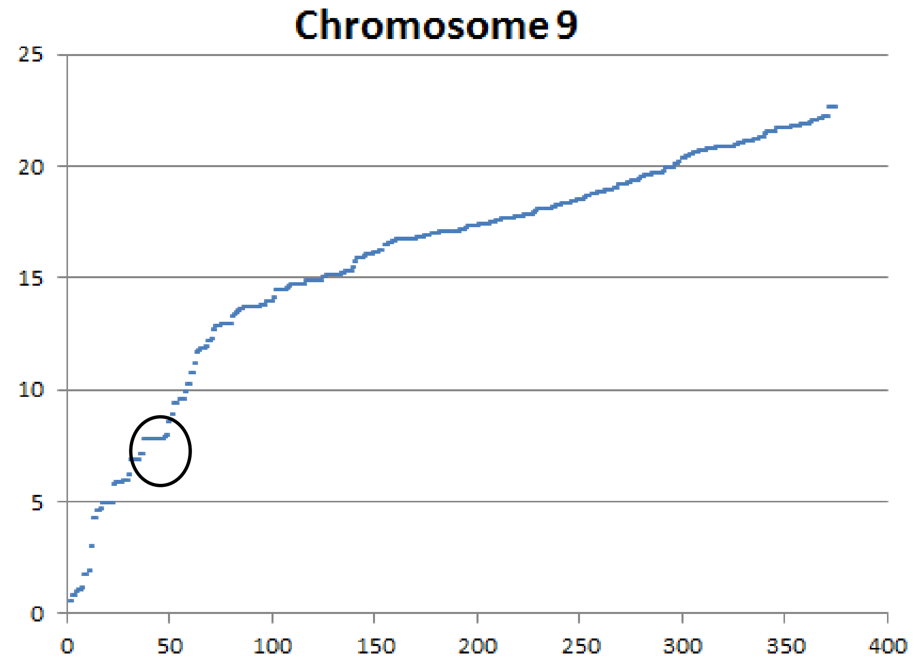

Chromosome 4
40
35
30
25
20
15
10
5
0
0
100
200
300
400
500
600
700

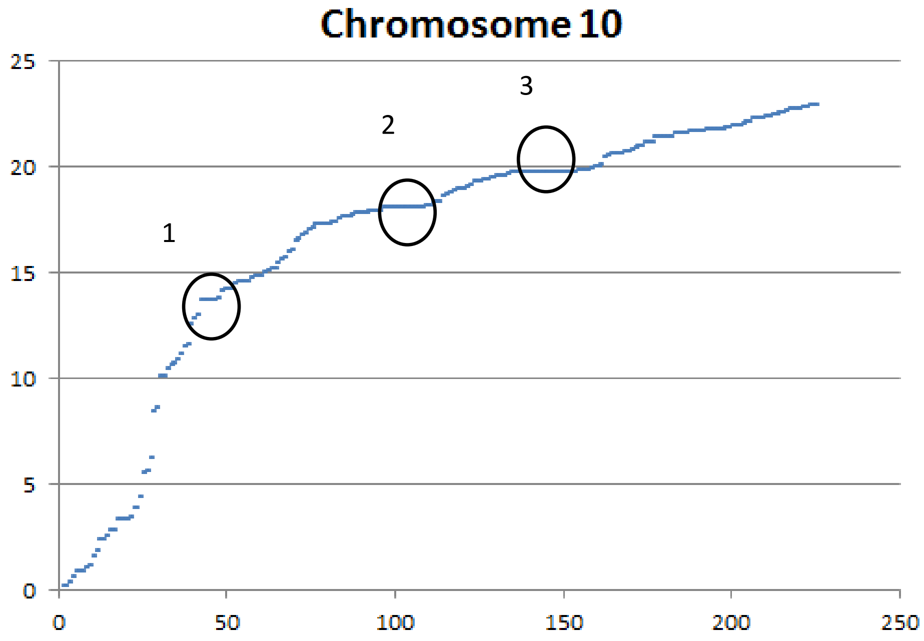

Chromosome 5
0
5
10
15
20
25
30
0
100
200
300
400
500

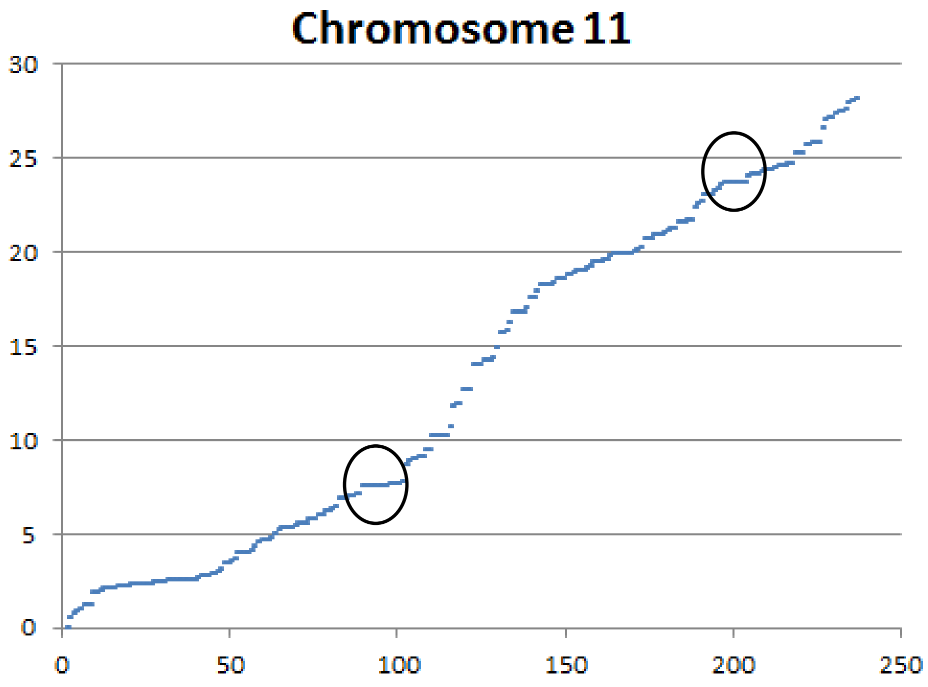

Chromosome 6
0
5
10
15
20
25
30
35
0
100
200
300
400
500

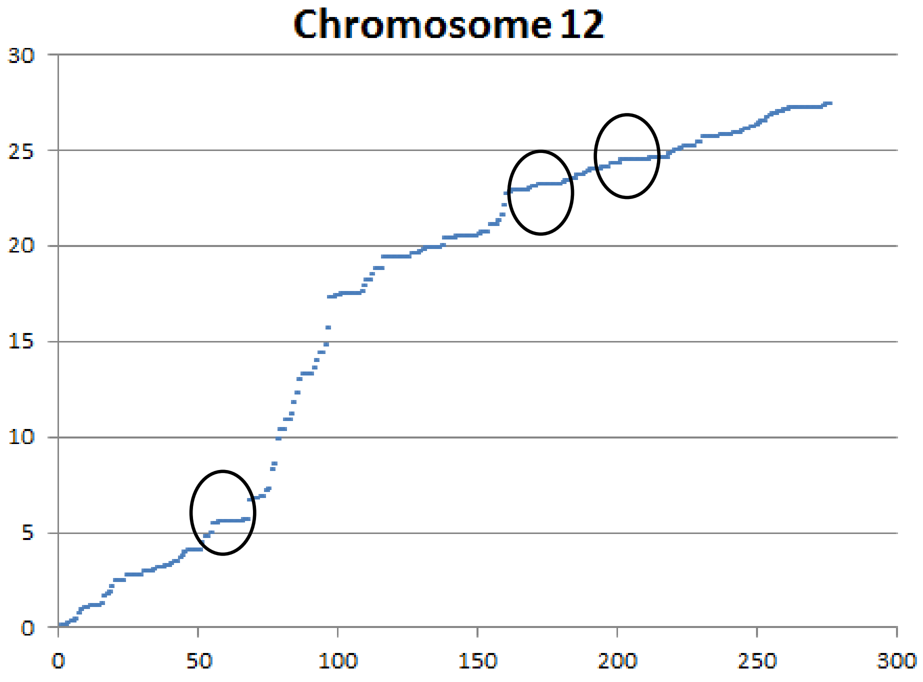

Chromosome 7
30
25
20
15
10
5
0
0
100
200
300
400
500


Chromosome 8
30
25
20
15
10
5
0
0
100
200
300
400
500

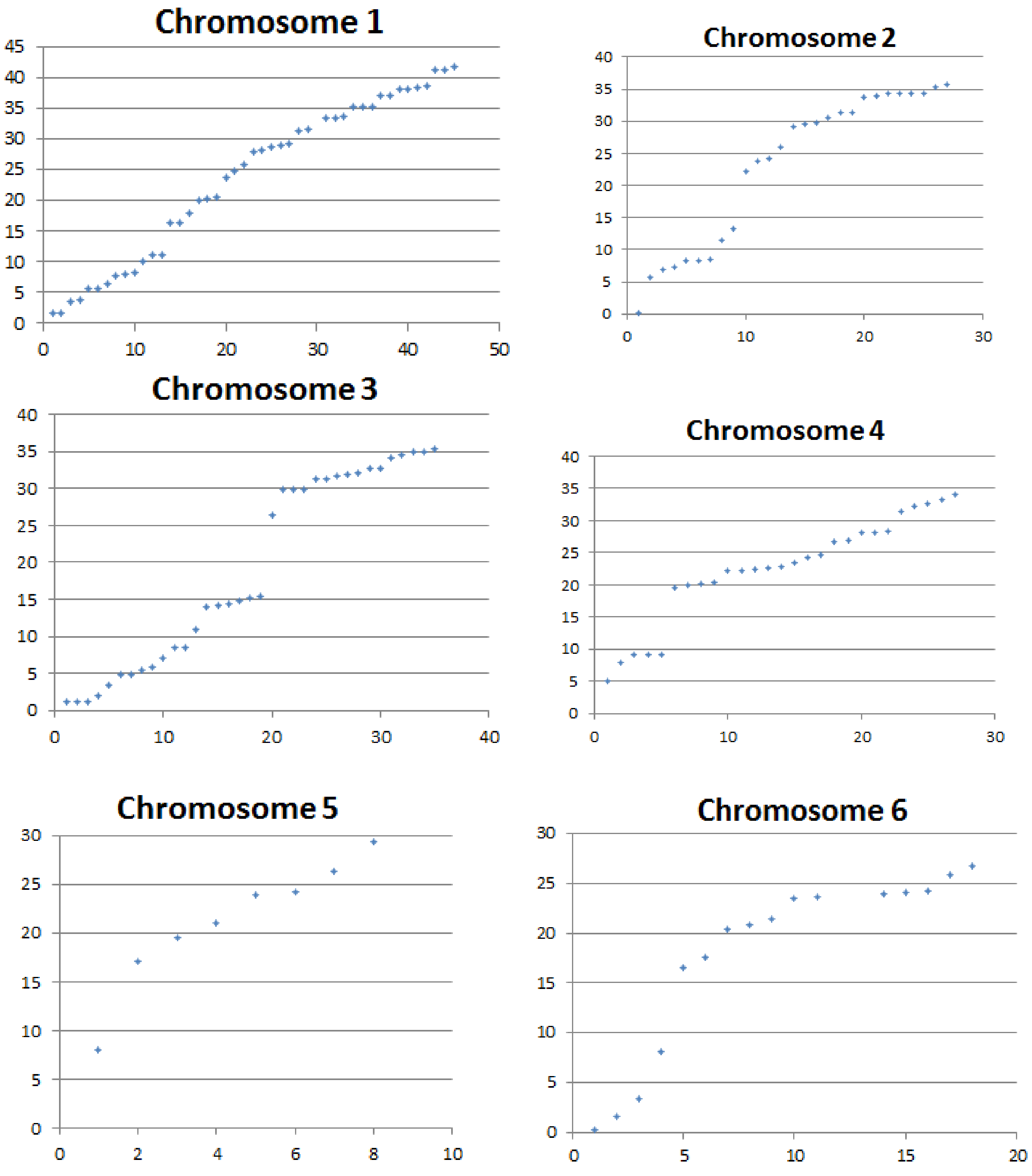

Chromosome 9
0
5
10
15
20
25
0
50
100
150
200
250
300
350
400

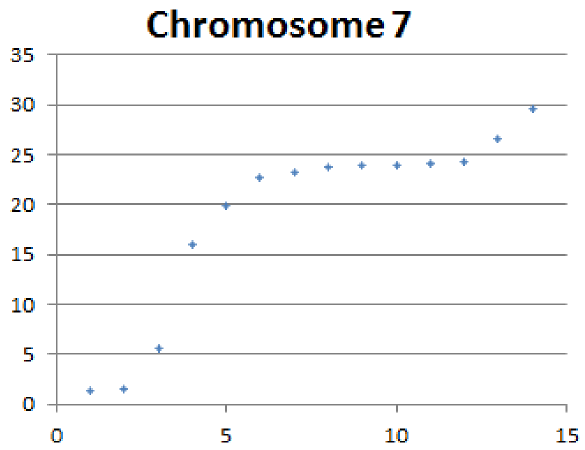

Chromosome 10
1
2
3
0
5
10
15
20
25
0
50
100
150
200
250

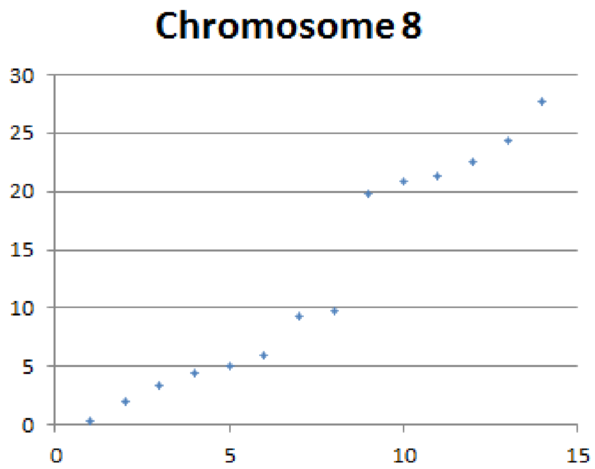

Chromosome 11
30
25
20
15
10
5
0
0
50
100
150
200
250

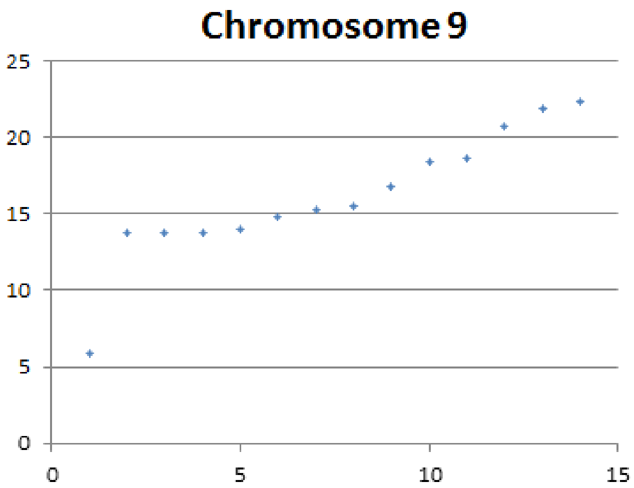

Chromosome 12
30
25
20
15
10
5
0
0
50
100
150
200
250
300

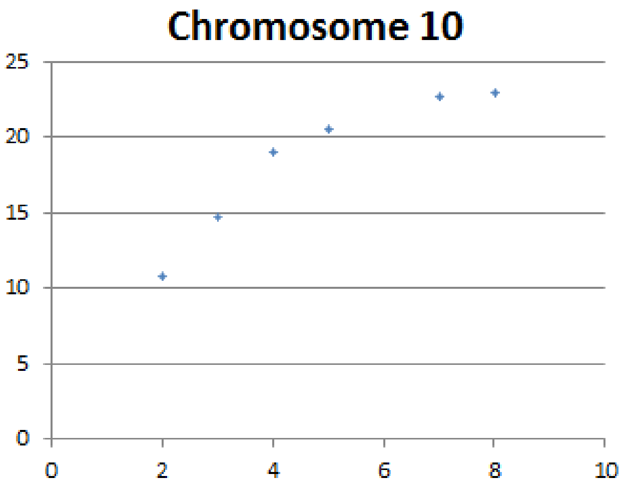
(B) Distribution of monocot specific CNSs on *O.sativa japonica* chromosomes. Y axis – genomic location of the CNSs (Mb), X axis – cumulative frequency of the CNSs.

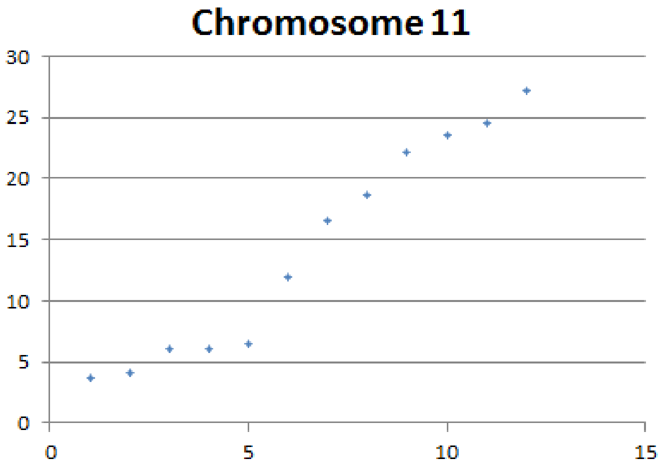


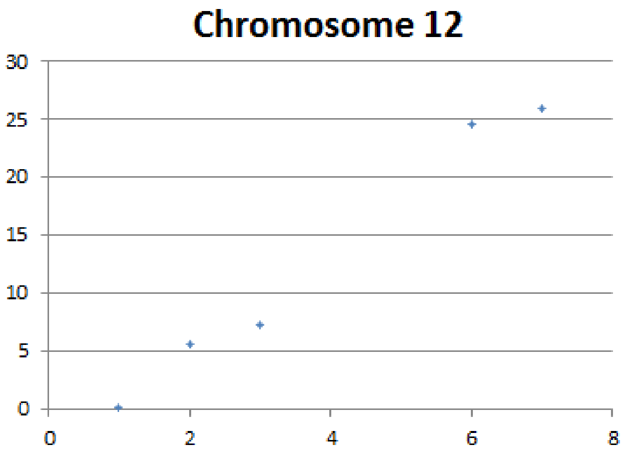

Chromosome 7
35
30
25
20
15
10
5
0
0
5
10
15


Chromosome 8
30
25
20
15
10
5
0
0
5
10
15


Chromosome 9
25
20
15
10
5
0
0
5
10
15

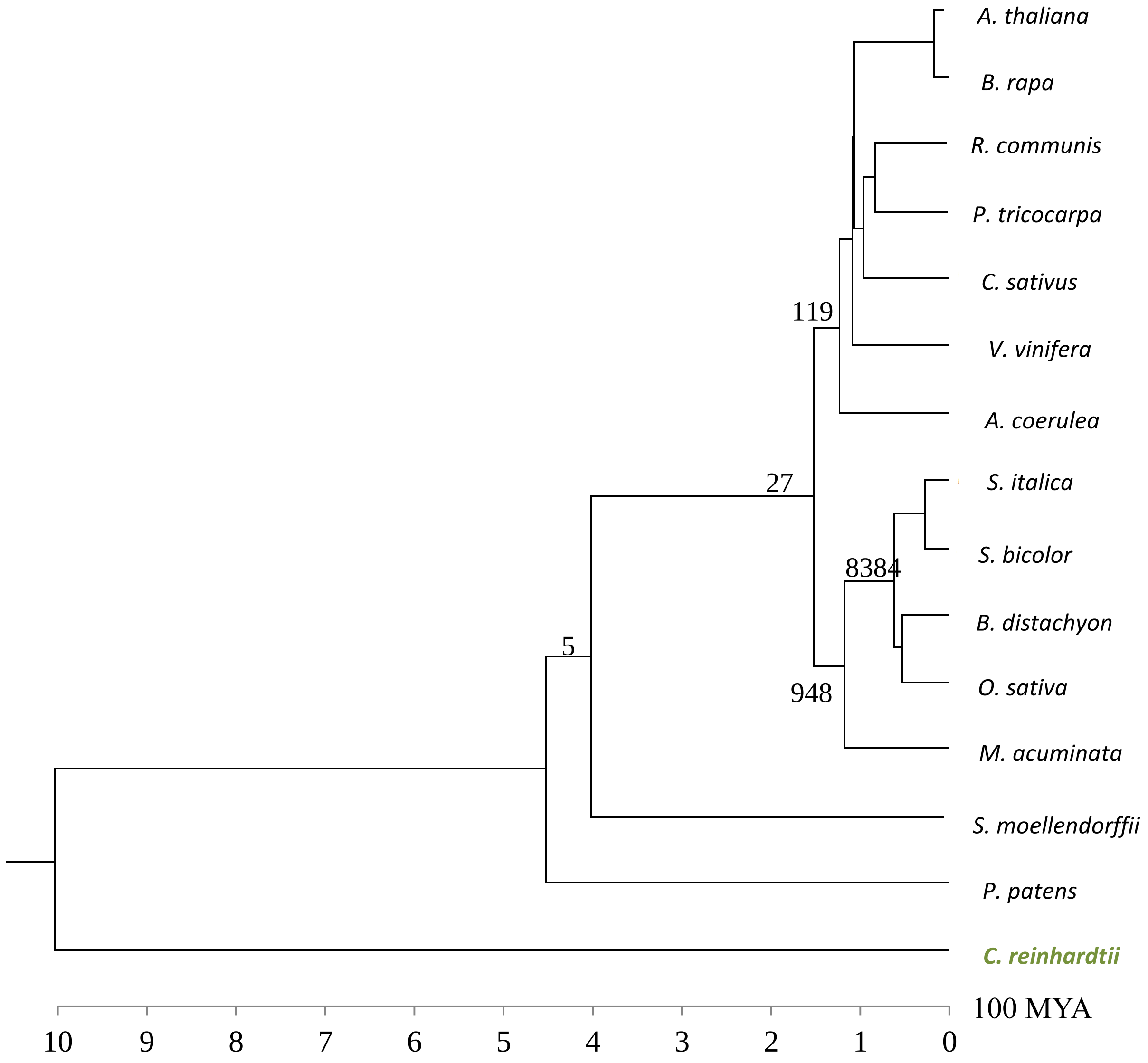

Chromosome 10
25
20
15
10
5
0
0
2
4
6
8
10


Chromosome 11
30
25
20
15
10
5
0
0
5
10
15


Chromosome 12
30
25
20
15
10
5
0
0
2
4
6
8

Appendix A3:

The phylogenetic tree for lineage common CNSs. The numbers on branches represent the CNSs that are common to each lineage.



Appendix A4

(A)

Gene enrichment analysis for 6536 random genes of *O. sativa japonica*

Functional group	Percentage of genes in the group	P-value
Root hair cell differentiation	100.0	1.7E-8
Tricoblast maturation	100.0	1.7E-8
Cell maturation	100.0	1.7E-8
Trichoblast differentiation	100.0	2.2E-8
Root epidermal cell differentiation	100.0	3.3E-8
Developmental maturation	100.0	3.8E-8
Epidermal cell differentiation	100.0	5.1E-7
Cell growth	75.0	6.9E-4
Regulation of cell size	75.0	7.7E-4
Growth	75.0	9.1E-4
Cell tip growth	50.0	1.5E-2
Leaf morphogenesis	50.0	2.3E-2
Shoot morphogenesis	50.0	3.2E-2
Leaf development	50.0	3.9E-2
Glycoprotein	50.0	4.0E-2
Phyllome development	50.0	4.3E-2
Shoot development	50.0	5.9E-2
Post embryonic development	25.0	1.0E-0
GPI-anchor	25.0	1.0E-0
Endomembrane system	25.0	1.0E-0

(B)

(I) Under-represented GO terms related to catabolic processes. (II) Under-represented GO terms related to enzymatic and structural processes.

(I)

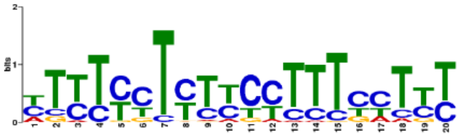
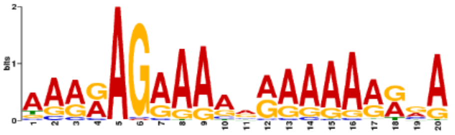
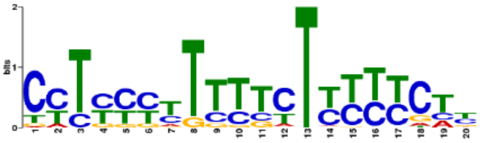
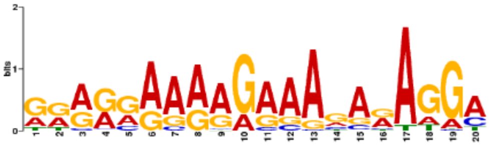
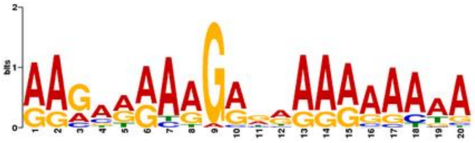
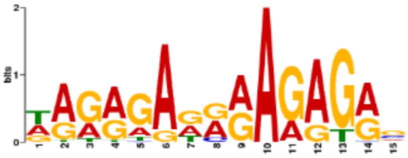
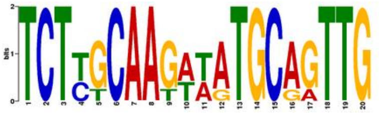
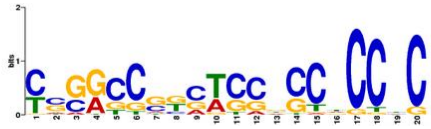
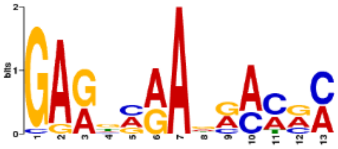
Functional group	Percentage of genes in the group	P-value
Modification dependent protein catabolic process	98.5	2.8E-92
Macromolecule catabolic process	98.5	2.8E-92
Cellular protein catabolic process	98.5	5.1E-92
Ubl conjugation pathway	81.8	6.5E-90
Proteolysis	98.5	6.8E-74
Ubiquitin-protein ligase activity	59.1	3.5E-53
Amino-acid ligase activity	59.1	2.3E-51
Ubiquitin-ligase-complex	37.9	9.2E-23
U box domain	25.8	1.9E-27
Protein ubiquitination	28.8	3.0E-22

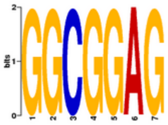
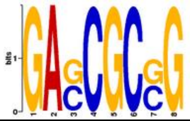
(II)

Functional group	Percentage of genes in the group	P-value
Glycoside hydrolase	65.7	9.8E-40
Glycosidase	60.0	2.3E-33
Signaling pathway	57.1	9.0E-18
Glycoprotein	48.6	6.2E-16
Related to extra cellular region	48.6	6.5E-9
Cation binding	65.7	1.1E-7
Cell wall	31.4	1.4E-6
External encapsulating structure	31.4	1.6E-6
Apoplast	22.9	3.4E-6
Cell wall biogenesis/degradation	17.1	8.5E-6

Appendix A5:

Motif overrepresentation data for lineage specific CNSs. This table gives the motifs with highest statistical significance. Motifs identified by MEME for dicot specific CNSs have a higher E-value.

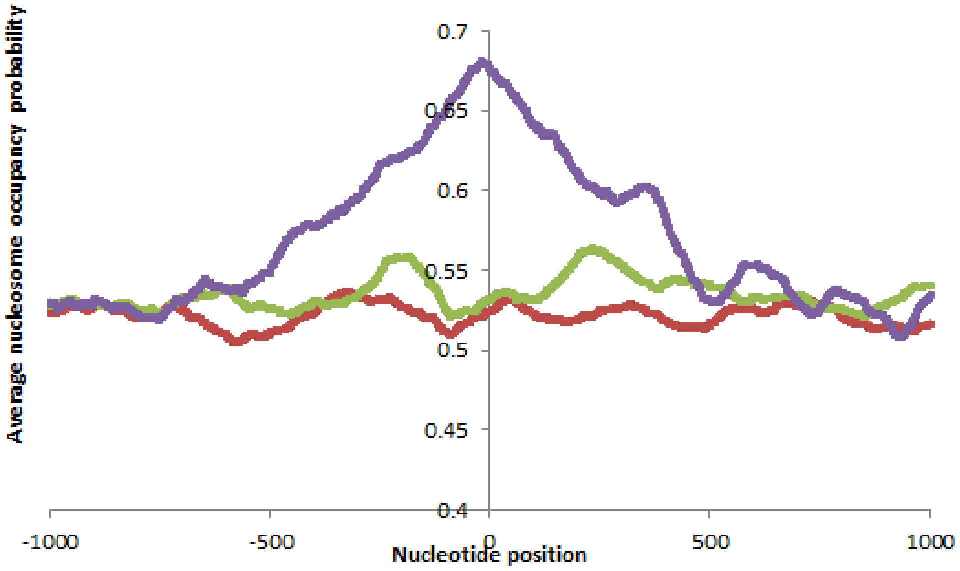
	Motif logo	E-value	GO term
Grass specific		4.2e-024	TF activity
		4.5e-023	TF activity
		2.3e-023	TF activity
Monocot specific		6.2e-013	DNA binding
		3.3e-015	TF activity
		8.5e-023	TF activity
		1.5e-001	GO term unknown
		2.8e-001	DNA binding
Dicot specific		1.8e+002	GO unknown

	7.8e+003	GO unknown
	2.4e+004	GO unkonwn

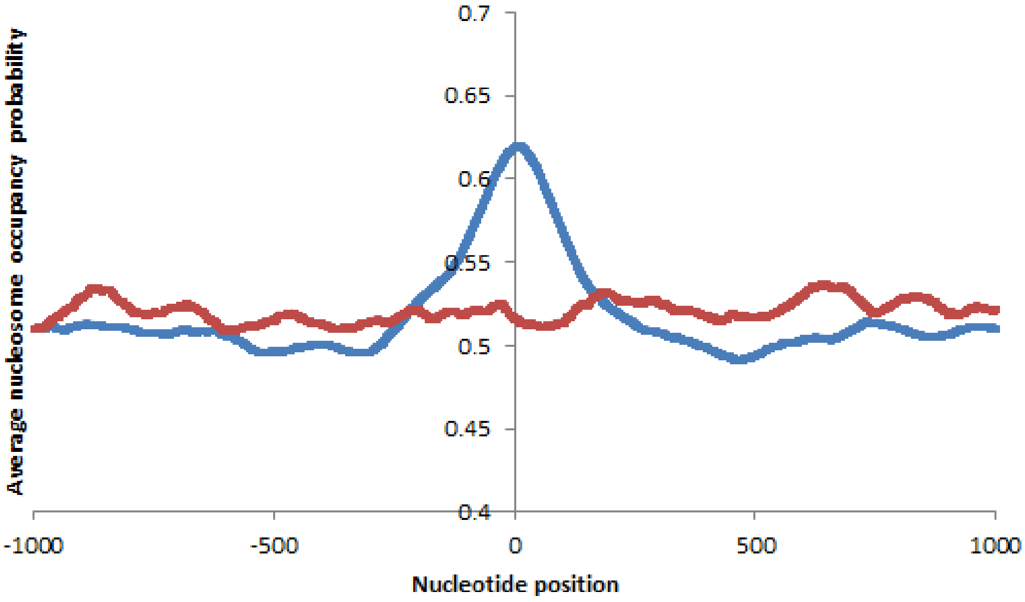
Appendix A6:

Nucleosome occupancy probability for grass specific UTR, Intergenic and Intronic CNSs. 0th nucleotide position represent the center of the CNSs. (A) Nucleosome occupancy probability for grass specific UTR CNSs. Purple, green and red graphs respectively show nucleosome occupancy probabilities of the 5' UTR CNSs, 3' UTR CNSs and random samples with same AT content as CNSs. (B) Nucleosome occupancy probability for grass specific intergenic CNSs. Blue and red graphs show the nucleosome probabilities for intergenic CNSs and random samples. (C) Nucleosome occupancy probability for grass specific intronic CNSs. Blue and red graphs show the nucleosome probabilities for intronic CNSs and random samples.

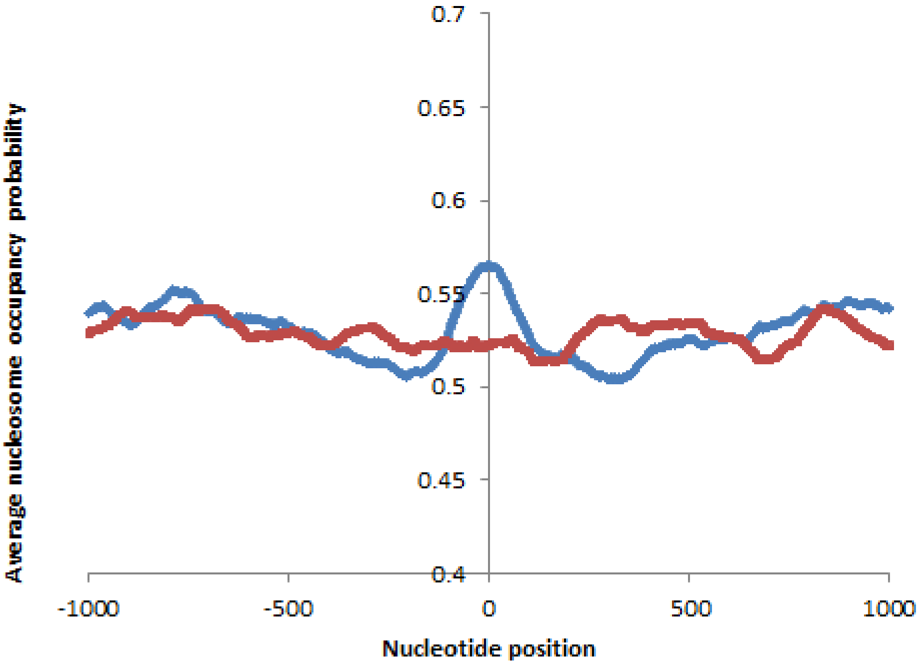
(A)



(B)


0.7
0.65
0.6
0.55
0.5
0.45
0.4
-1000
-500
0
500
1000
Nucleotide position
Average nucleosome occupancy probability

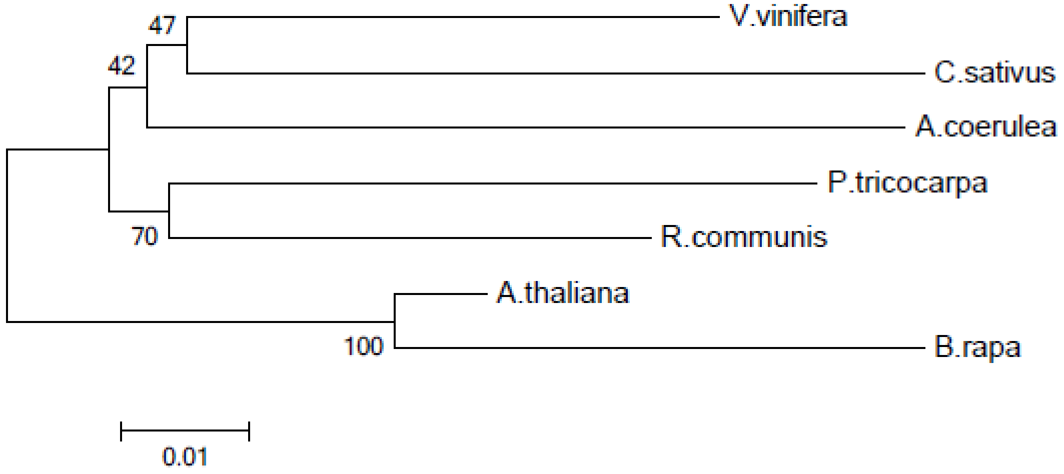
(C)


0.7
0.65
0.6
0.55
0.5
0.45
0.4
-1000
-500
0
500
1000
Nucleotide position
Average nucleosome occupancy probability

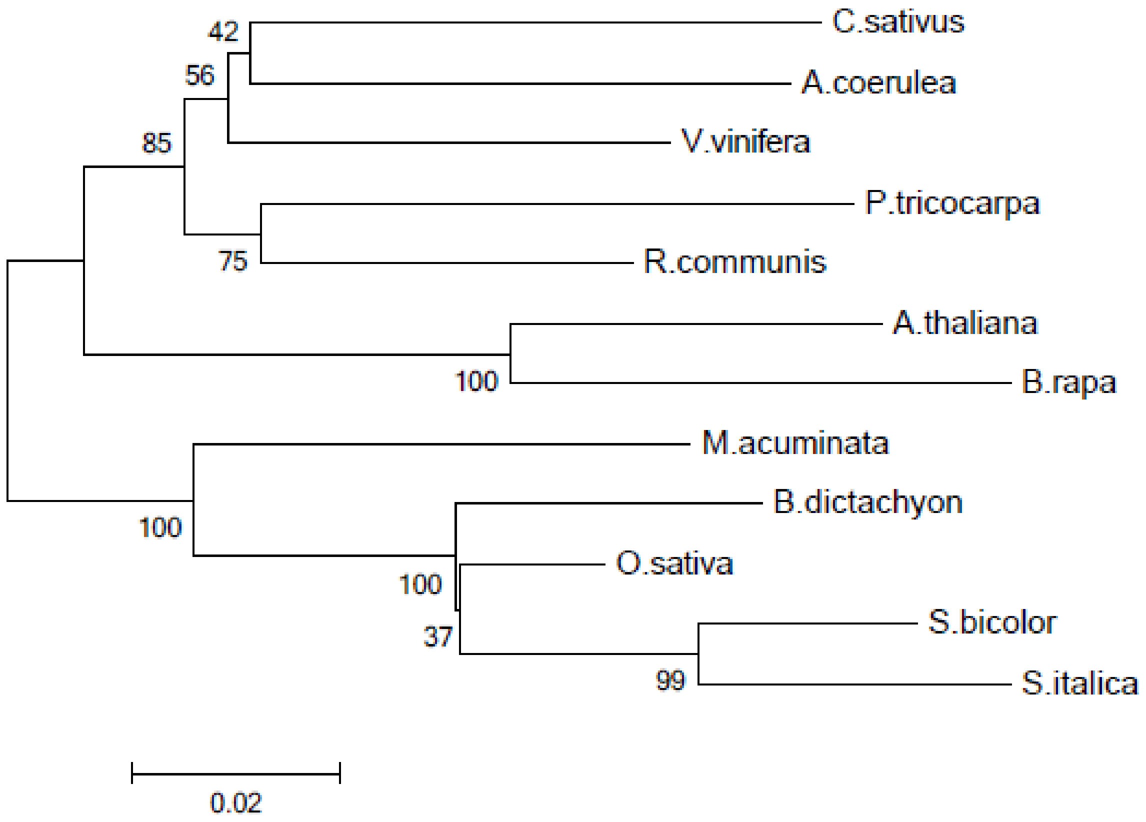
Appendix A7:

The phylogenetic trees constructed for eudiot and angiosperm CNSs. The concatenated lineage specific CNSs were used to construct the phylogenetic trees with neighbor-joining method. (A) Phylogenetic tree constructed for dicots in the study with dicot specific CNSs. (B) Phylogenetic tree constructed for all angiosperms in the study with angiosperm specific CNSs.

(A)



(B)



Appendix A8:

Thresholds for CNSs determined based on protein conservation level based on CDSs of species used in the analysis. (A) Fungi (B) Invertebrates (C) Non-mammalian vertebrates

A

Species	Average percentage identity
A. oryzae – A. nidulans	73
B. fuckeliana – S. sclerotiorum	81
G. zeae – T.reesei	73
M. oryzae – M. poae	75
P. graminis – P. triticina	79
P. nodorum – P.teres	74
A. gossypii – S. cerevisiae	71
S. octosporus – S. pombe	72
S. reilianum – U. maydis	77

B

Species	Average percentage identity
D. plexippus– B. mori	74
D. melanogaster – A. darlingi	85
A. cephalotes– N. vitripennis	73
C. briggsae – C. japonica	75

C

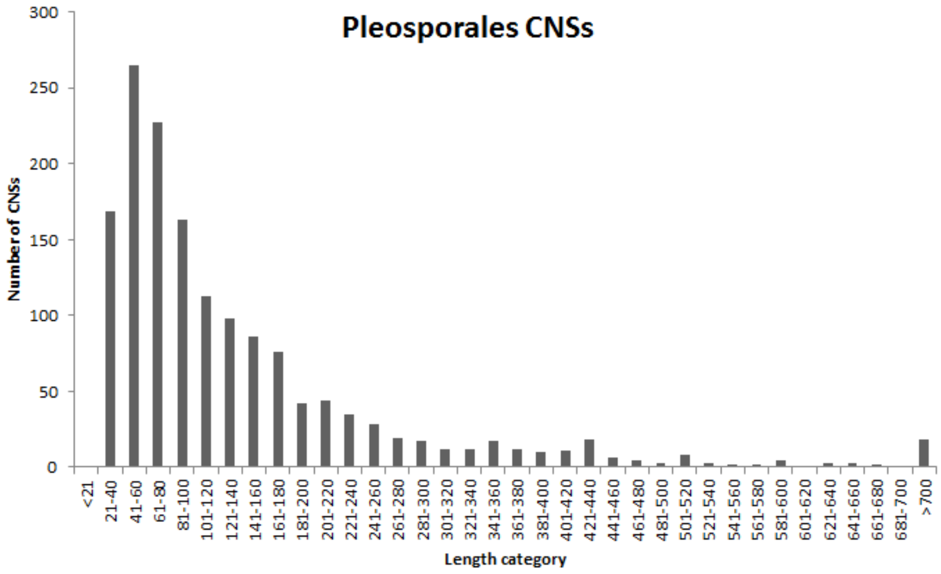
Species	Average percentage identity
T. nigroviridis– T.rubripes	86
G. gallus – T. guttata	85
G.gallus – P. sinensis	82
G. gallus – A. carolinensis	78
G. gallus – X. tropicalis	75

Appendix A9:

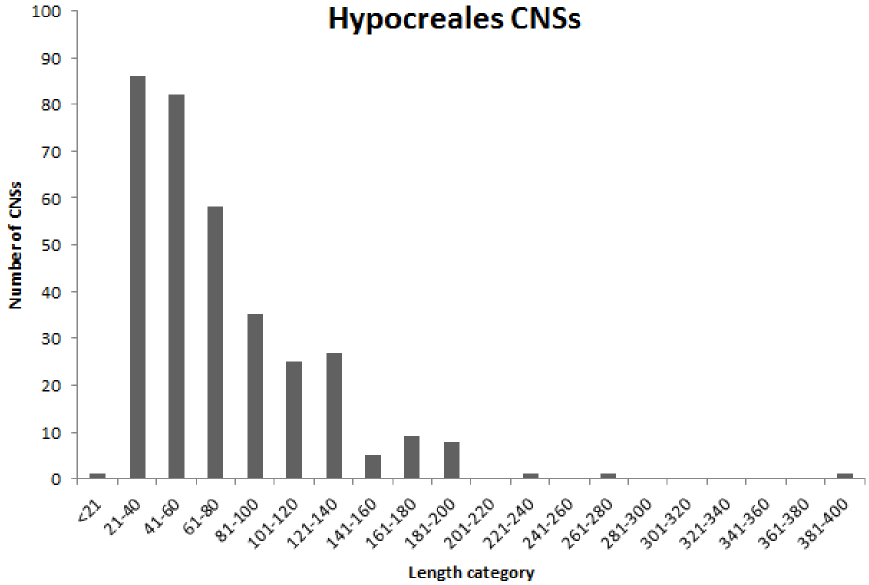
The length distributions for fungi, invertebrate and non-mammalian vertebrate CNSs.

(A) Length distributions for fungi common CNSs. (I) Pleosporales (II) Hypocreales (III) Schizosaccharomycetales (IV) Eurotiales show length distributions of each group for CNSs respectively.

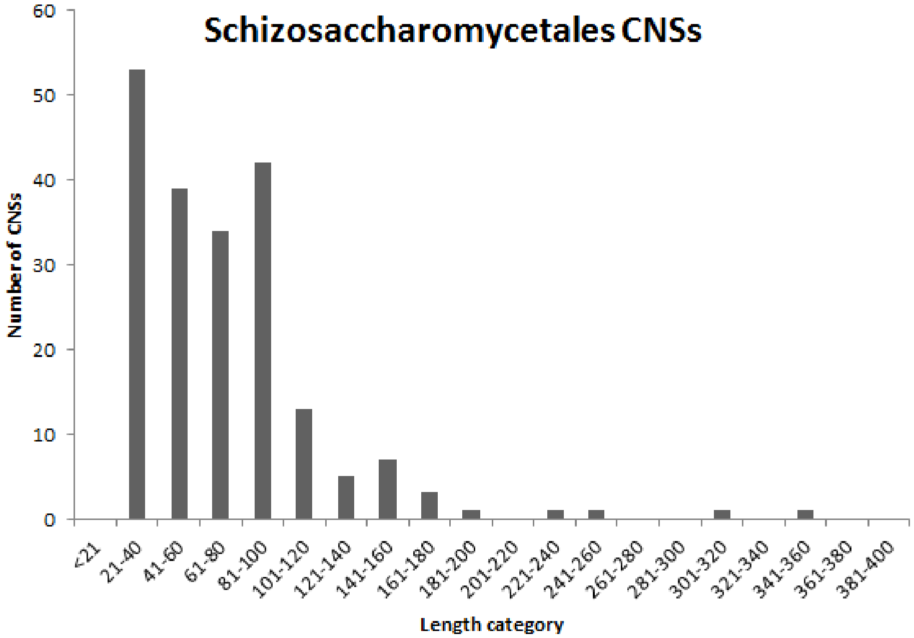
I



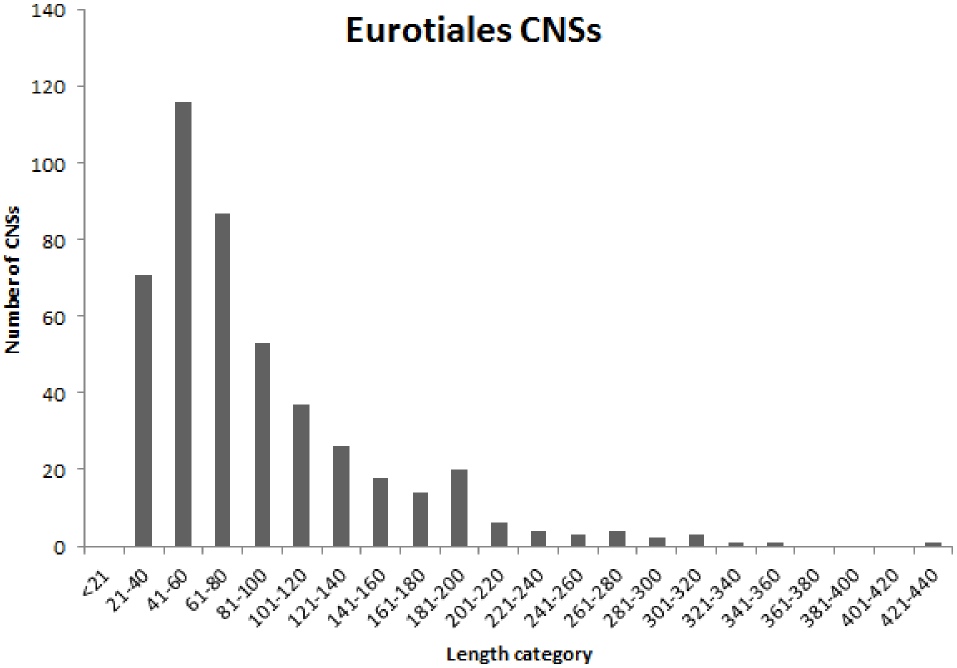
II



III

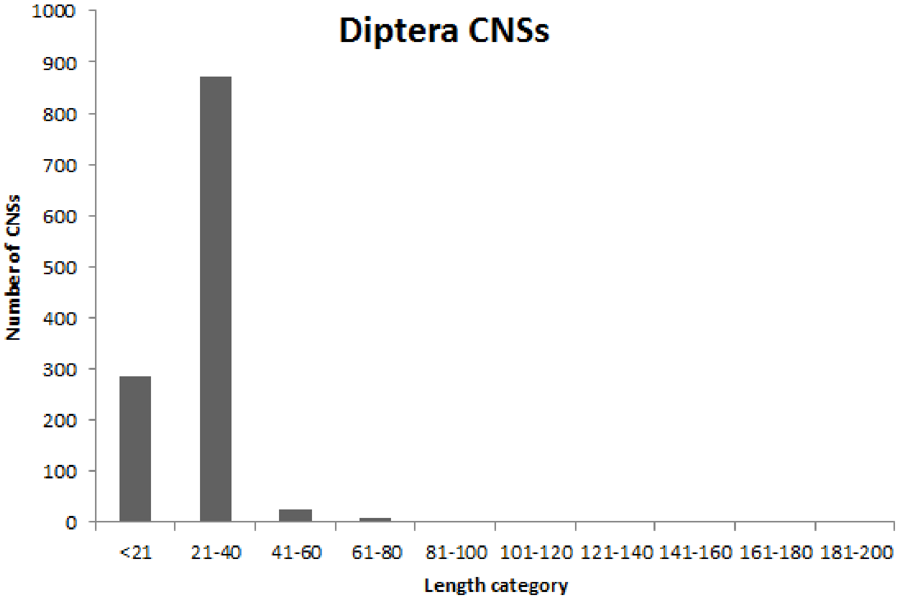

Schizosaccharomycetales CNSs
Number of CNSs
0
10
20
30
40
50
60
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
Length category

IV

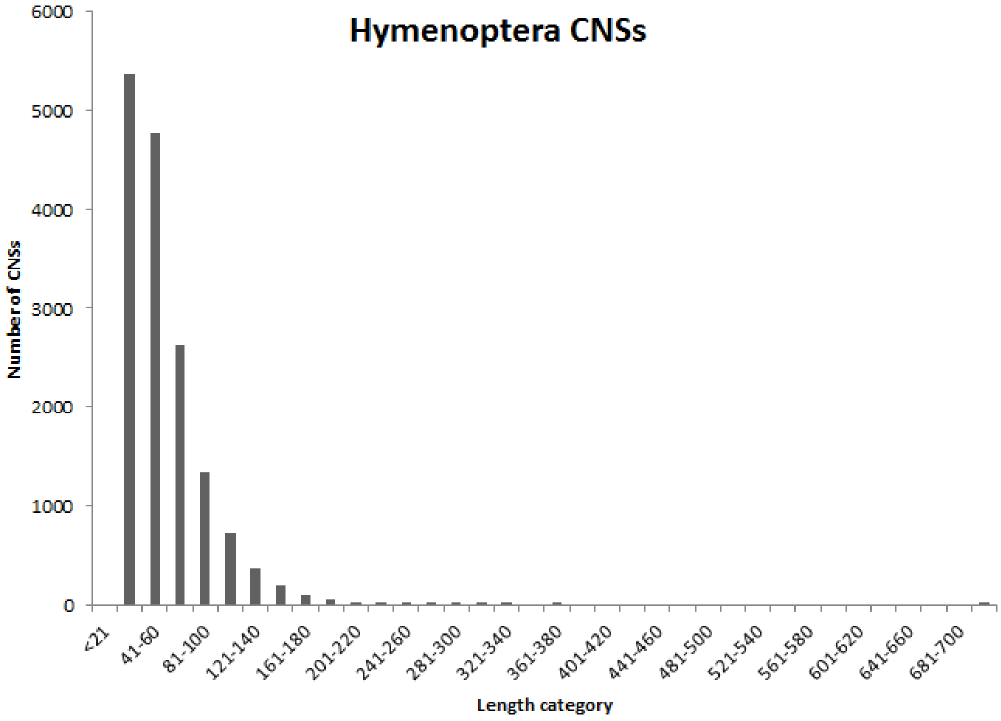

Eurotiales CNSs
Number of CNSs
0
20
40
60
80
100
120
140
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
401-420
421-440
Length category

(B) Length distributions for Invertebrate common CNSs. Figure (I) Diptera (II) Hymenoptera (III) Lepidoptera (IV) Nematode gives length distributions for each order.

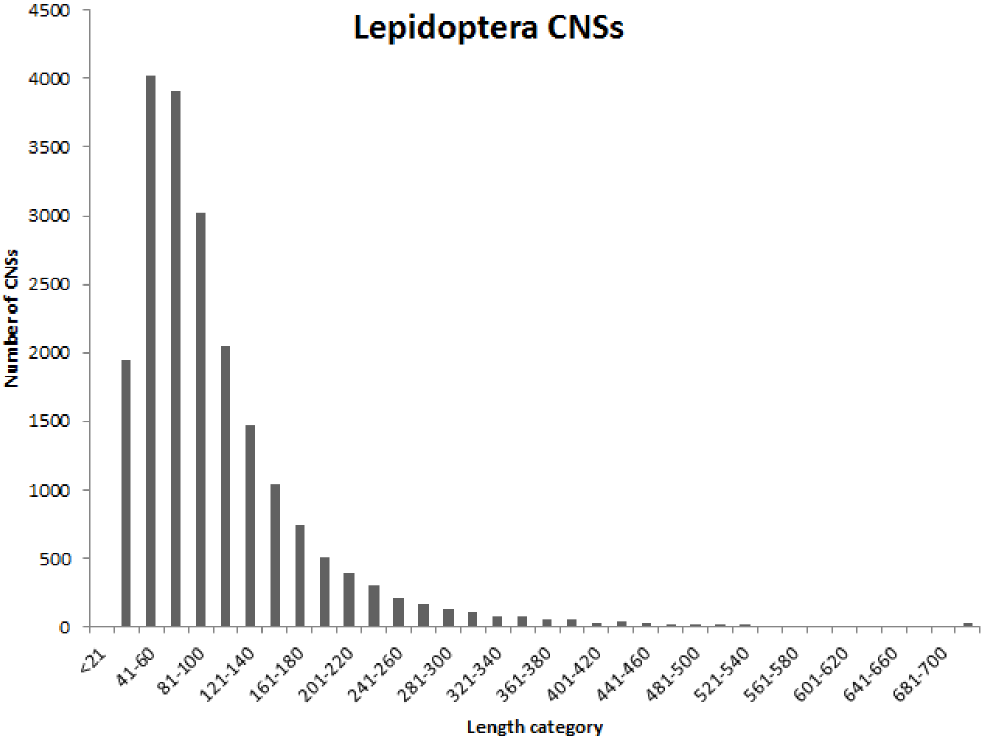
I



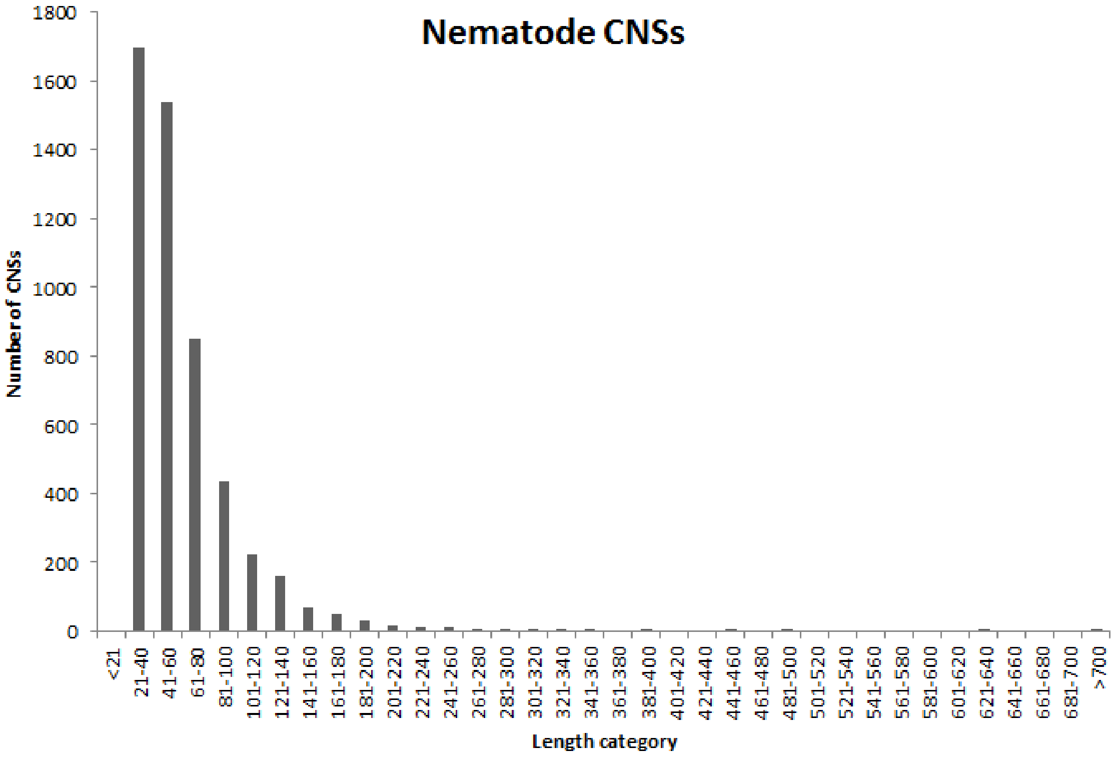
II



III

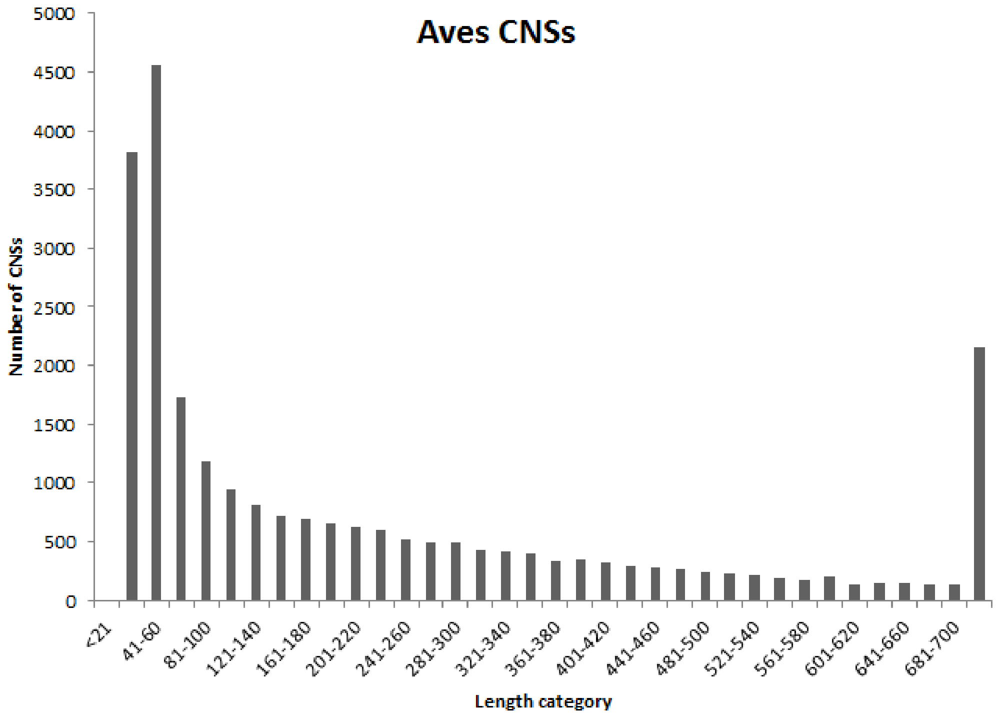

Lepidoptera CNSs
Number of CNSs
4500
4000
3500
3000
2500
2000
1500
1000
500
0
<21
41-60
81-100
121-140
161-180
201-220
241-260
281-300
321-340
361-380
401-420
441-460
481-500
521-540
561-580
601-620
641-660
681-700
Length category

IV

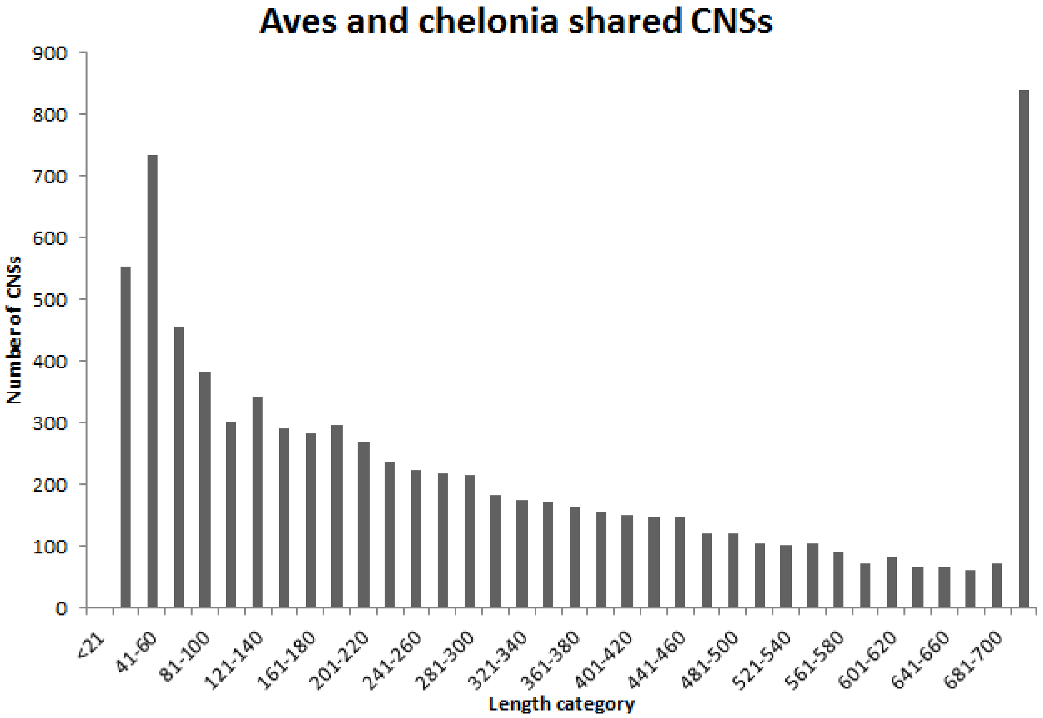

Nematode CNSs
Number of CNSs
1800
1600
1400
1200
1000
800
600
400
200
0
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
401-420
421-440
441-460
461-480
481-500
501-520
521-540
541-560
561-580
581-600
601-620
621-640
641-660
661-680
681-700
>700
Length category

(C) Length distributions for non-mammalian vertebrate common CNSs. Figures I, II, III, IV ,V gives the length distribution for Aves, Aves and Chelonian shared, Reptilian, Reptilian and Amphibian shared and teleost fish CNSs respectively.

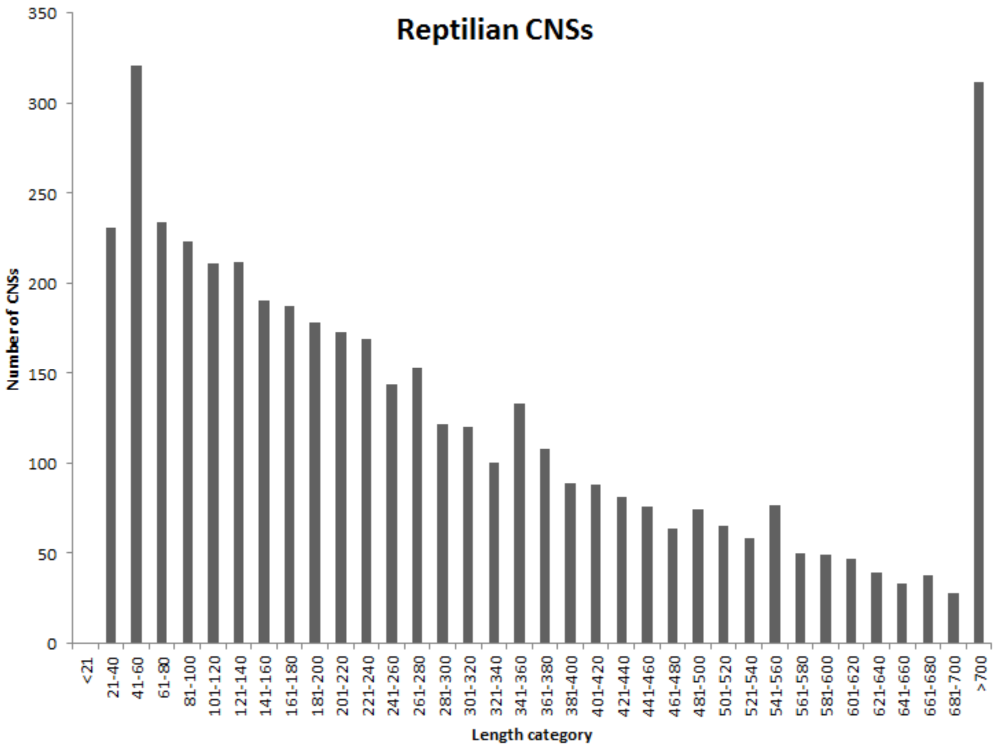
I



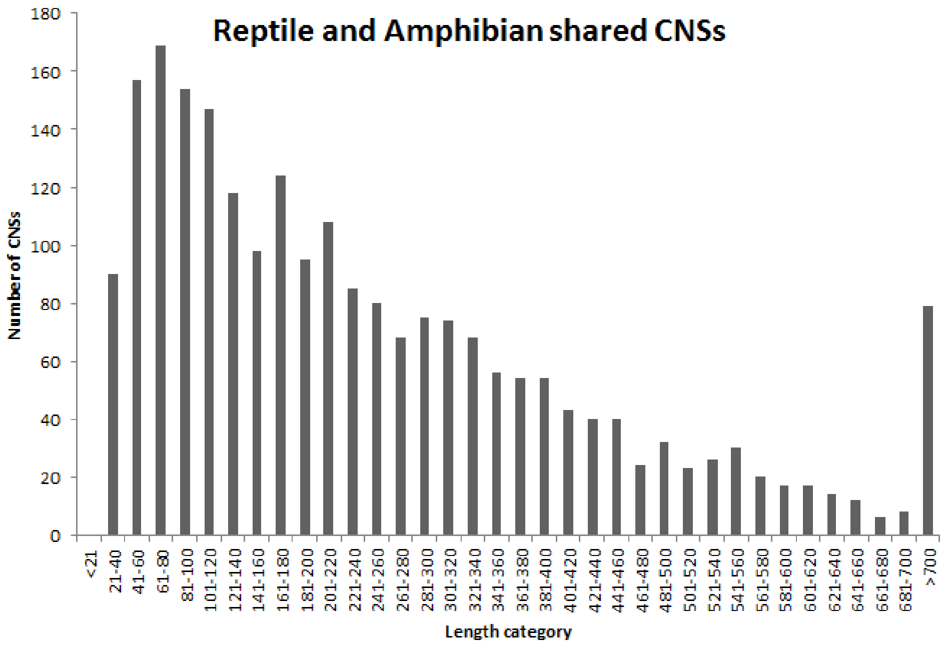
II



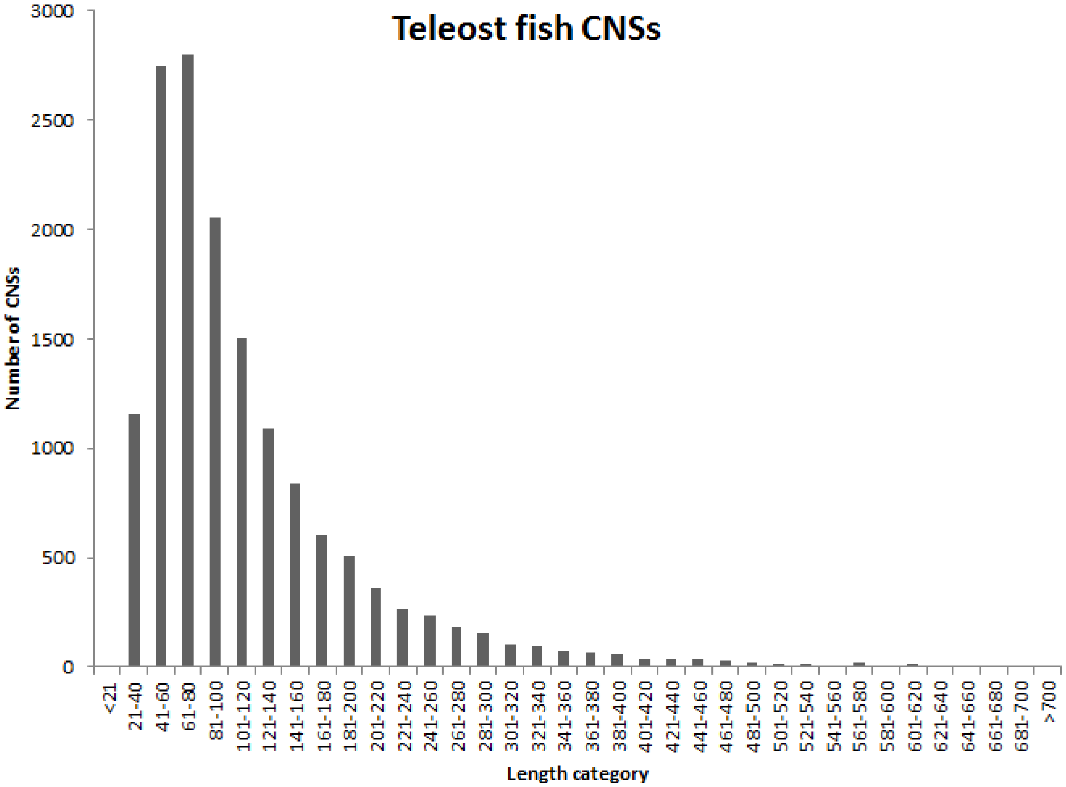
III


Reptilian CNSs
Number of CNSs
0
50
100
150
200
250
300
350
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
401-420
421-440
441-460
461-480
481-500
501-520
521-540
541-560
561-580
581-600
601-620
621-640
641-660
661-680
681-700
>700
Length category

IV


Reptile and Amphibian shared CNSs
Number of CNSs
0
20
40
60
80
100
120
140
160
180
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
401-420
421-440
441-460
461-480
481-500
501-520
521-540
541-560
561-580
581-600
601-620
621-640
641-660
661-680
681-700
>700
Length category

V


Teleost fish CNSs
Number of CNSs
3000
2500
2000
1500
1000
500
0
<21
21-40
41-60
61-80
81-100
101-120
121-140
141-160
161-180
181-200
201-220
221-240
241-260
261-280
281-300
301-320
321-340
341-360
361-380
381-400
401-420
421-440
441-460
461-480
481-500
501-520
521-540
541-560
561-580
581-600
601-620
621-640
641-660
661-680
681-700
>700
Length category

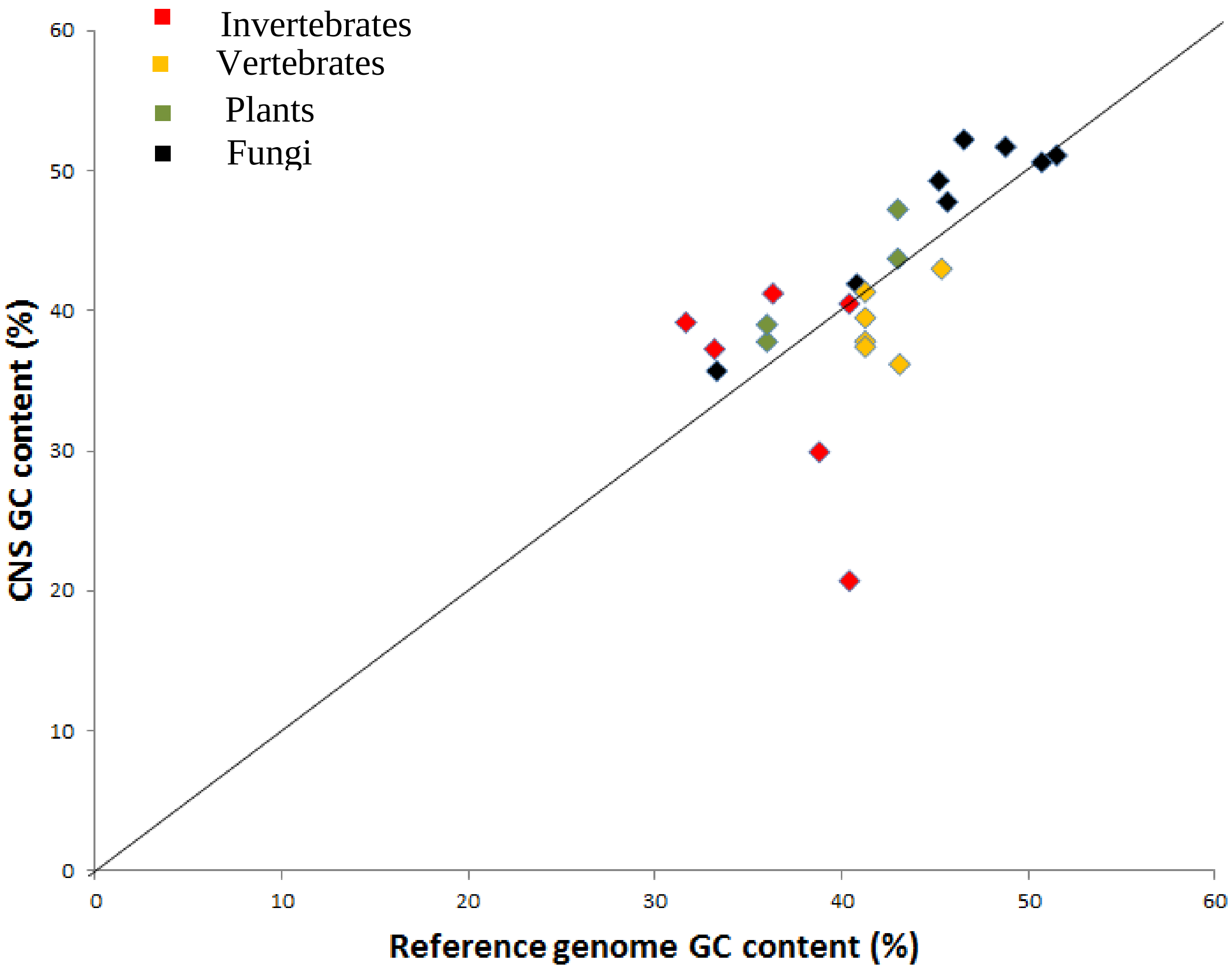
Appendix A10:

Substitution patterns for CNSs determined by MEGA 6.0.

Group	Lineage	Substitution pattern for CNSs
Plants	Eudicots Grasses Monocots Angiosperms	GC=>AT
Mammals (Babarinde and Saitou 2013)	Tetrapod Primates	GC=>AT
Non-mammalian vertebrates	Birds Reptiles Teleost fish	GC=>AT
Fungi	Eurotilaes	AT=>GC
	Schizosaccharomtcetales	GC=>AT
	Hypocreales	AT=>GC
	Pleoaporales	AT=>GC

Appendix A11:

The average GC content of CNSs in different lineages with respect to the reference genome GC content. The drosophids and mosquito groups were considered separately for this analysis. Mammals and non-mammalian vertebrates are termed as vertebrates in general.



Appendix A12:

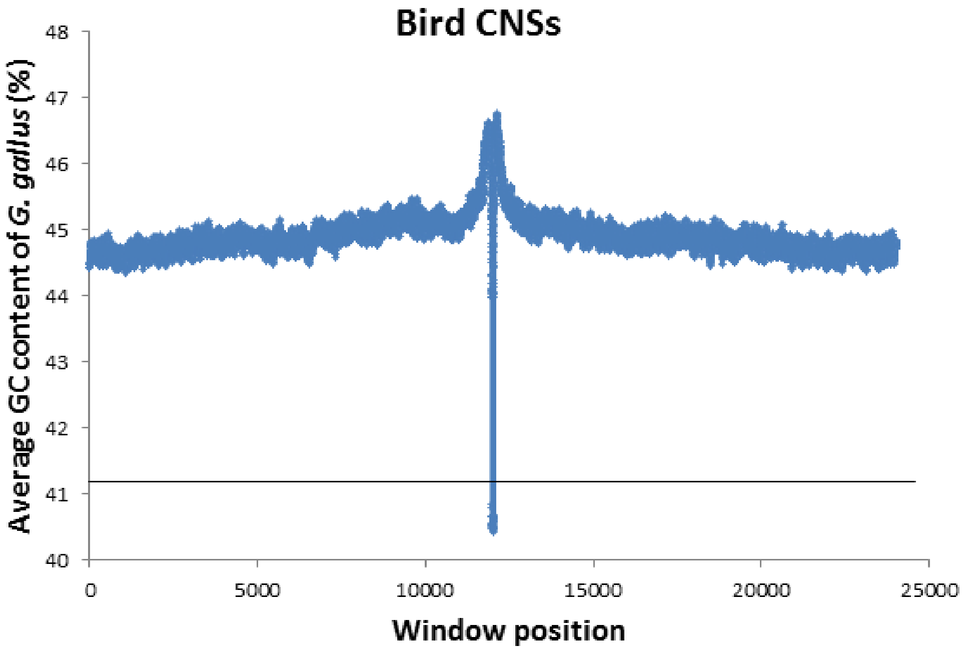
Statistical significance of CNS GC content compared to the reference genome noncoding GC contents for fungi, invertebrate and non-mammalian vertebrate CNSs.

Lineage	Groups used in the analysis	Reference genome GC content (%)	CNS GC content (%)	P-value
Fungi	Eurotiales	46.55	52.21	0.004
	Pleosporales	48.78	51.70	1.352E-05
	Hypocreales	45.21	49.28	0.035
	Schizosaccharomycetales	33.35	35.70	0.001
	Sclerotiniaceae	40.81	41.72	NS
	Pucciniales	46.67	47.77	NS
Invertebrates	Diptera	40.41	20.72	5.910E-193
	Hymenoptera	33.19	37.26	7.700E-72
	Lepidoptera	31.65	39.17	0.000
	Nematoda	36.31	41.23	3.20E-06
Non-mammalian vertebrates	Birds	41.26	41.33	NS
	Birds and chelonia	41.26	39.50	4.100E-32
	Reptiles	41.26	37.81	6.120E-184
	Reptiles and amphibian	41.26	37.42	1.600E-110
	Teleost fish	45.35	42.99	2.960E-94

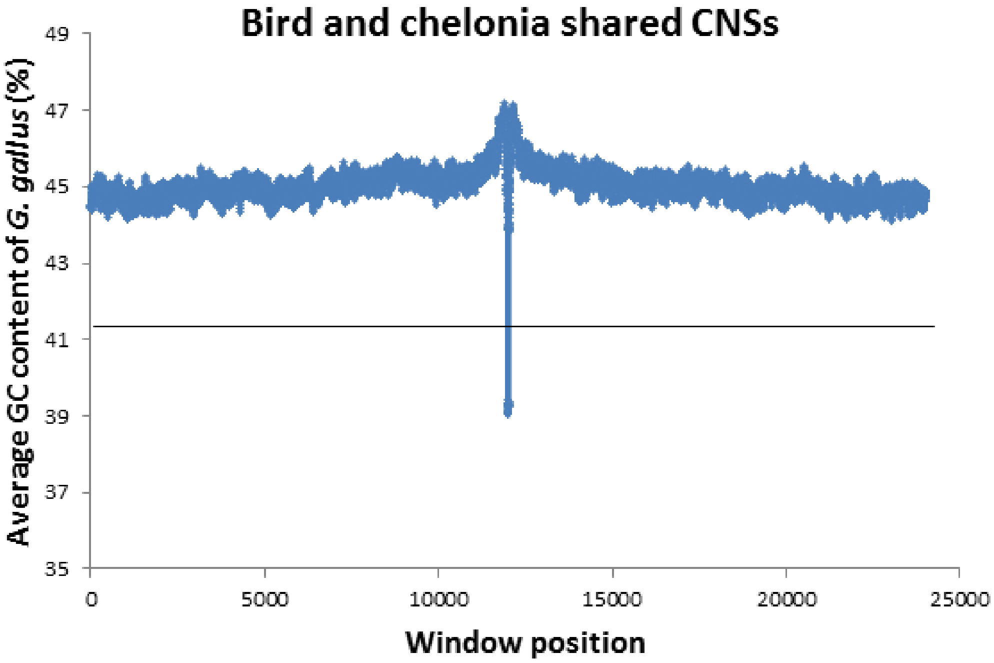
Appendix A13:

GC content distribution pattern for extended (12kb) flanking regions of non-mammalian vertebrates. Moving window analysis for GC content distribution for (A) bird, (B) bird and chelonian shared (C) reptilian (D) reptilian and amphibian shared CNSs (E) Mammalian common CNSs (Babarinde and Saitou (2013)).

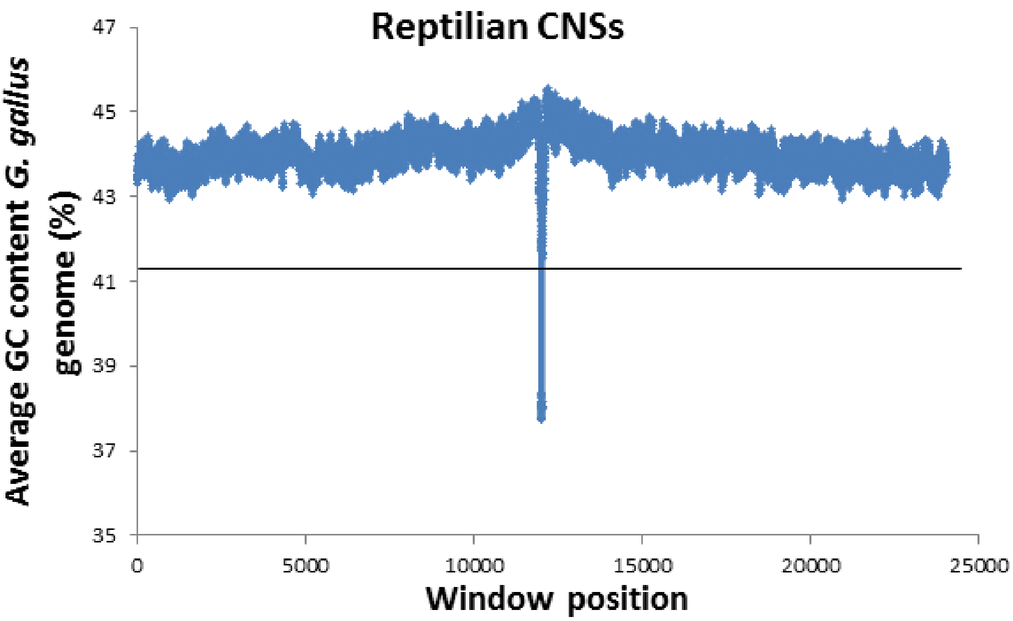
A



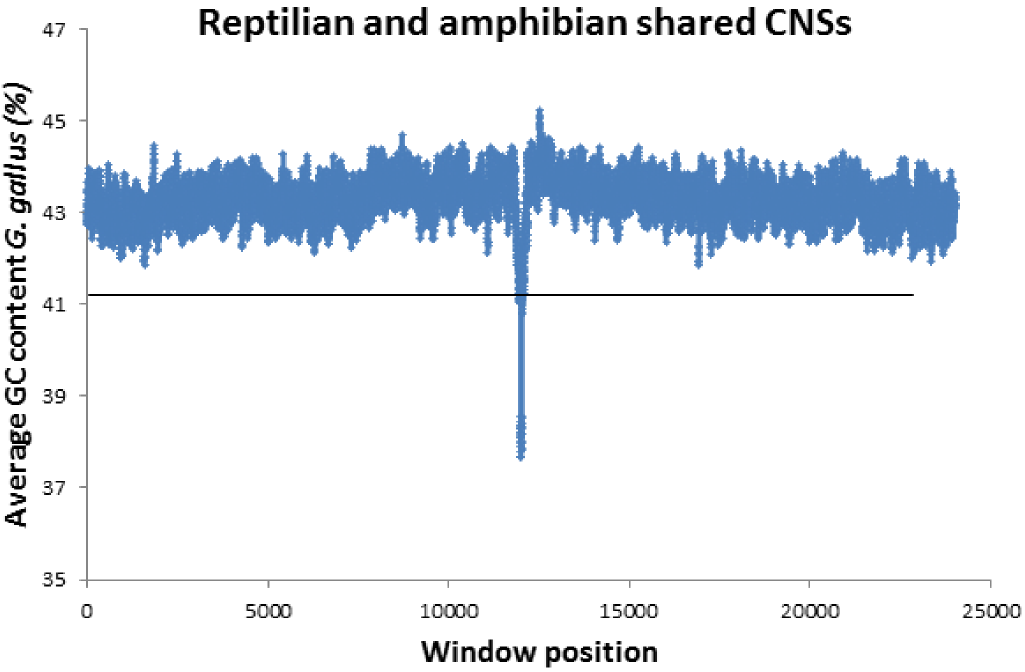
B



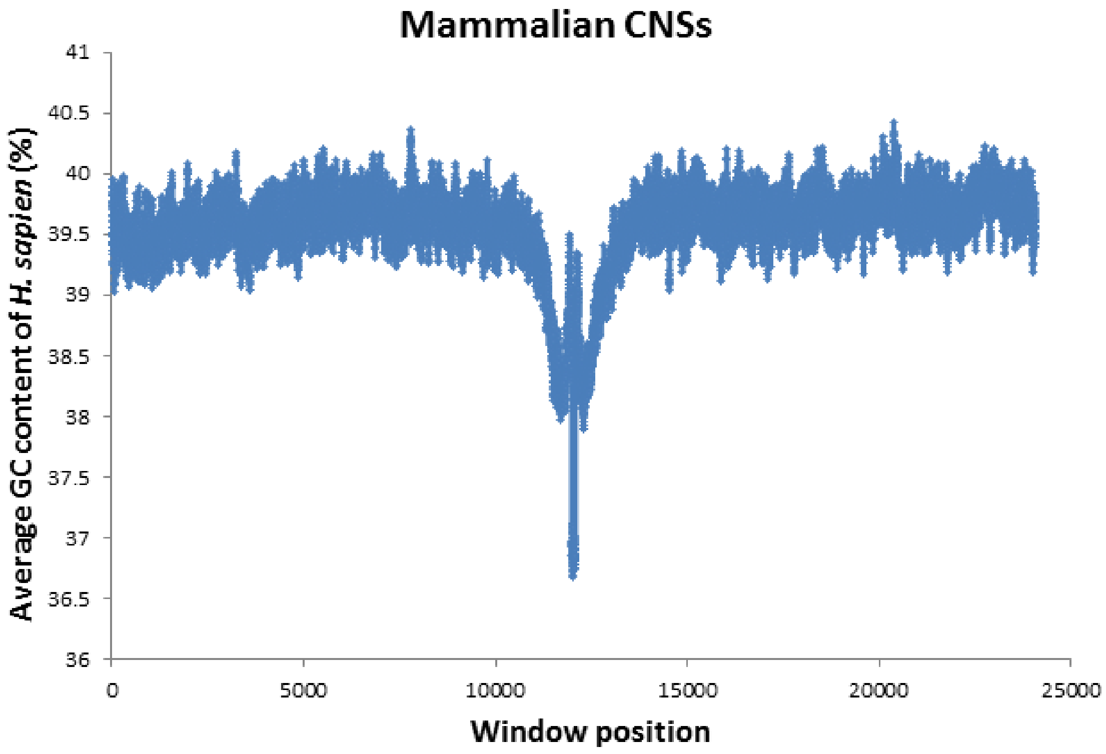
C


Reptilian CNSs
Average GC content G. gallus genome (%)
47
45
43
41
39
37
35
0
5000
10000
15000
20000
25000
Window position

D


Reptilian and amphibian shared CNSs
Average GC content G. gallus (%)
47
45
43
41
39
37
35
0
5000
10000
15000
20000
25000
Window position

E

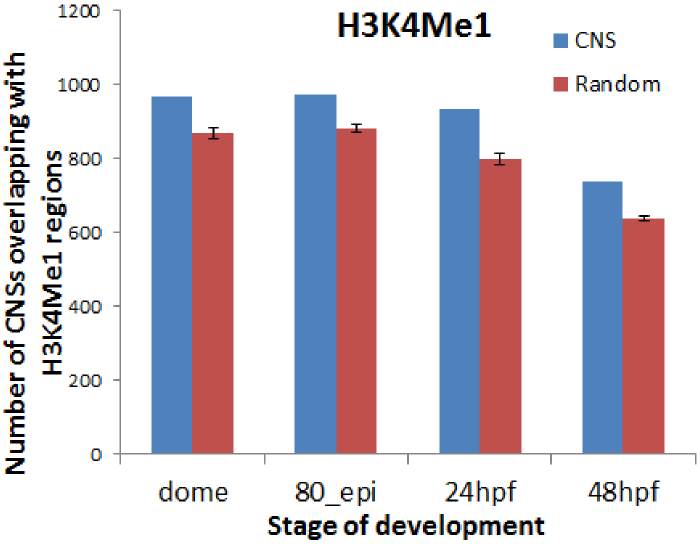

Mammalian CNSs
Average GC content of *H. sapien* (%)
41
40.5
40
39.5
39
38.5
38
37.5
37
36.5
36
0
5000
10000
15000
20000
25000
Window position

Appendix A14:

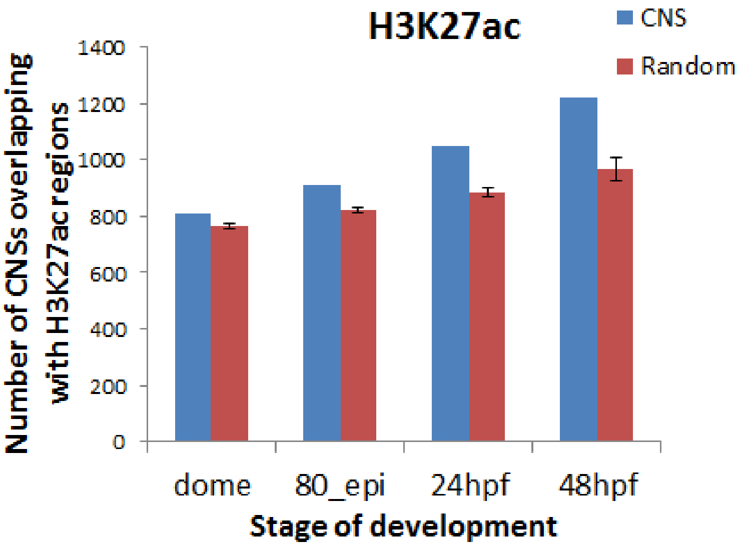
Chromatin modification signals overlapping with CNSs

(A) Chromatin modification signals overlapping with teleost fish CNSs at different development stages. (I) CNSs overlapping with H3K4Me1 regions. (II) CNSs overlapping with H3K27ac regions. (III) CNSs overlapping with H3K4Me3 regions.

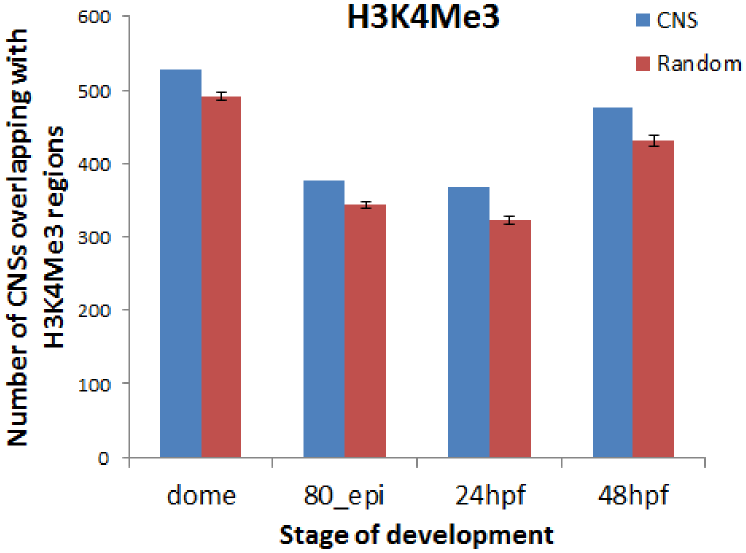
(I)



(II)

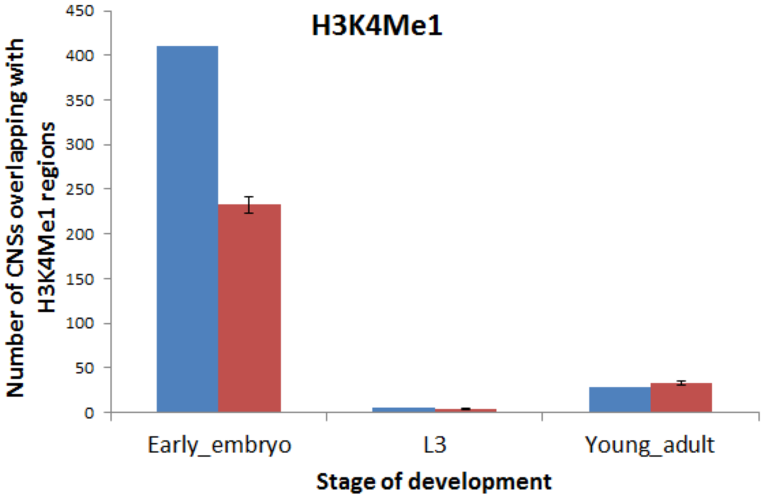


(III)

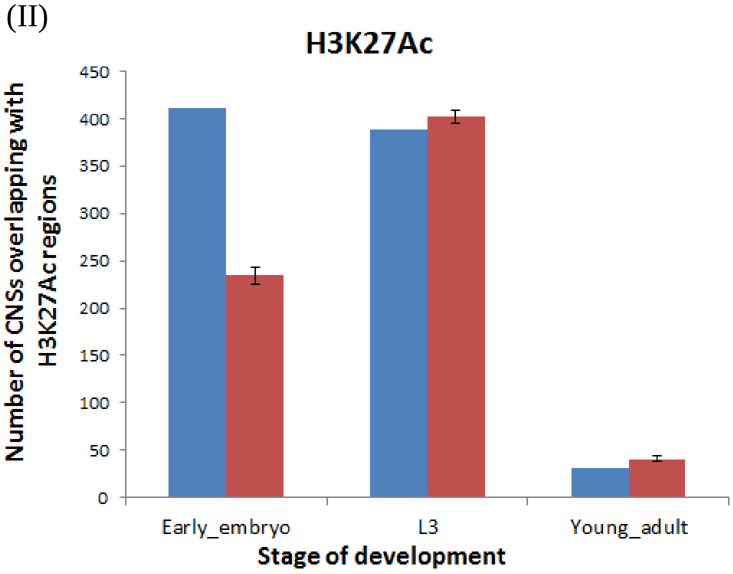


(B) Chromatin modification signals overlapping with Nematode CNSs at different development stages. (I) CNSs overlapping with H3K4Me1 regions. (II) CNSs overlapping with H3K27ac regions. (III) CNSs overlapping with H3K4Me3 regions.

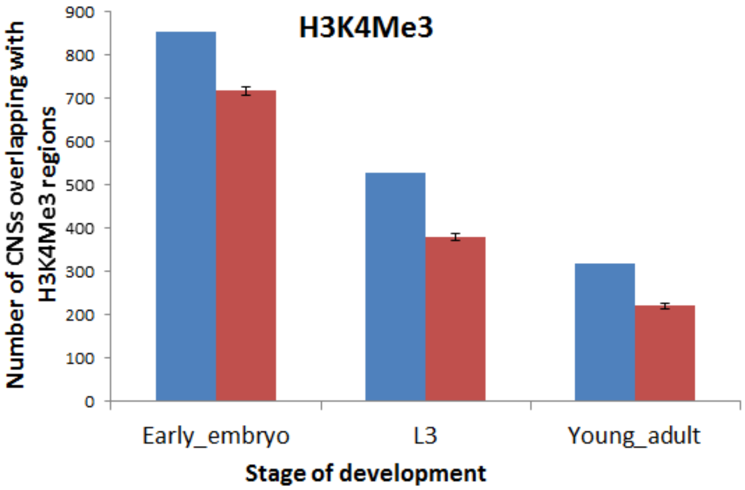
(I)



(II)

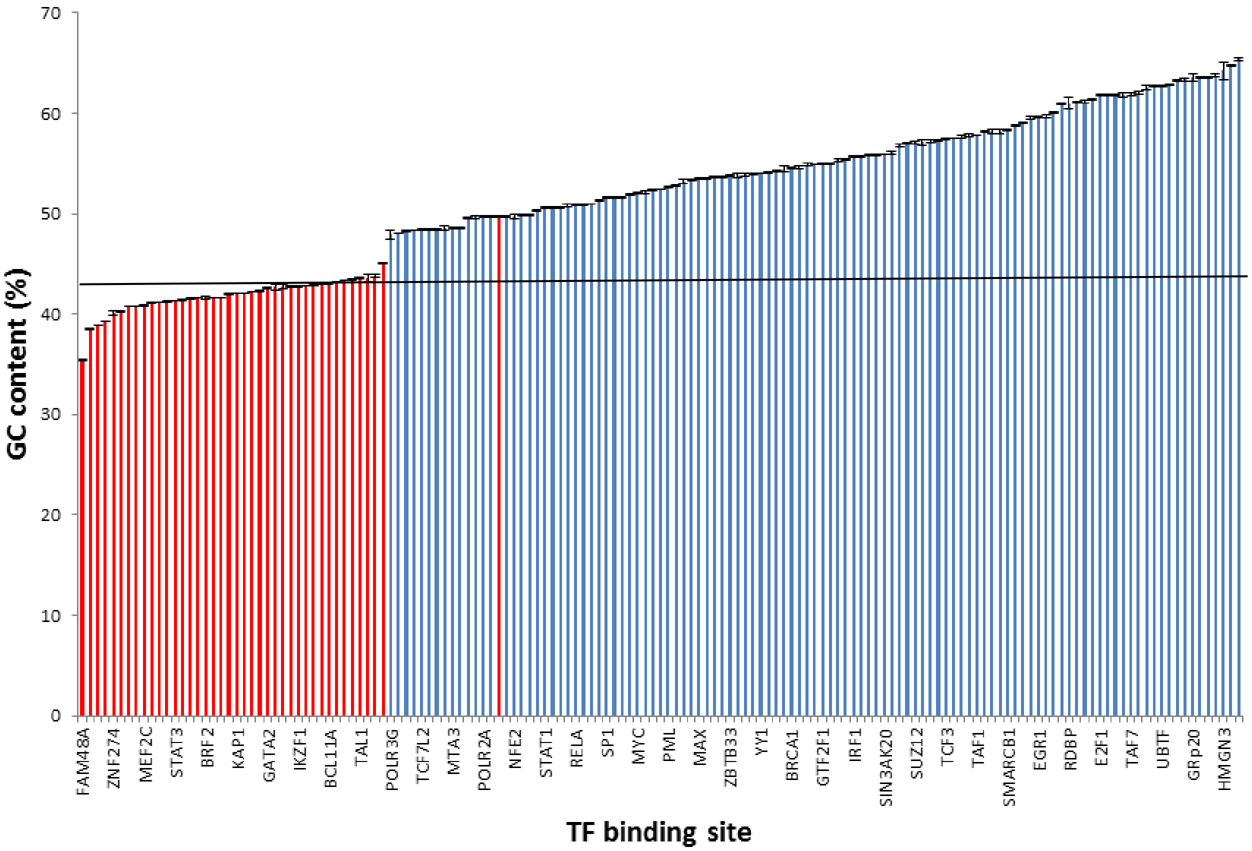


(III)



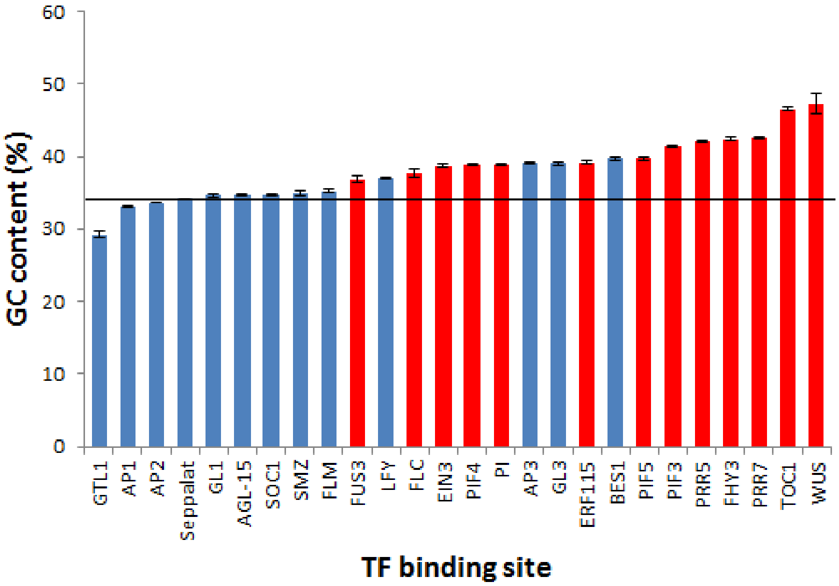
Appendix A15:

The GC content distribution for vertebrate (human) TF binding sites. The ubiquitous binding sites are shown in blue bars and tissue specific binding sites are represented in red color. The black horizontal line represents the noncoding genomic GC content for human genome.



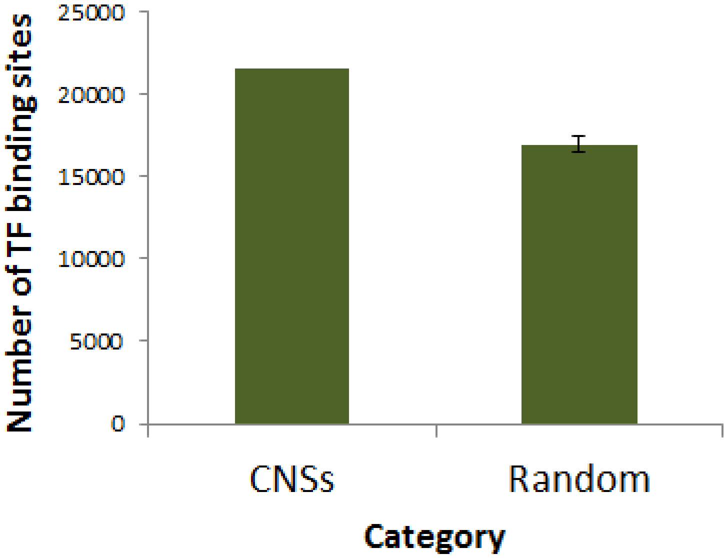
Appendix A16:

GC content distribution for plant (*A. thaliana*) TF binding sites. Ubiquitous binding sites are presented in blue color bars, whereas the tissue specific binding sites are provided in red color. The black horizontal line represents the noncoding genomic GC content for human genome.



Appendix A17:

CNSs are overrepresented with TF binding bites sites. The avian CNSs with homology to human genome were tested for TF binding site overrepresentation.


Number of TF binding sites
CNSs
Random
Category