The Pennsylvania State University

The Graduate School

Graduate Program in Cell and Developmental Biology

UNDERSTANDING EVOLUTIONARY HISTORY USING MOLECULAR PHYLOGENETICS: FROM GENES TO GENOMES

A Dissertation in

Cell and Developmental Biology

by

Xiaofan Zhou

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2011

The dissertation of Xiaofan Zhou was reviewed and approved* by the following:

Hong Ma
Distinguished Professor of Biology
Dissertation Advisor
Chair of Committee

Claude dePamphilis
Professor of Biology

Zhi-Chun Lai
Associate Professor of Biology, Biochemistry and Molecular Biology
Program Chair for Cell and Developmental Biology

David Geiser
Professor of Biology

* Signatures are on file in the Graduate School

ABSTRACT

Molecular phylogenetics is a powerful tool for deciphering the history of life on earth. The development of molecular phylogenetics and the increasing availability of molecular data have enabled unprecedented understanding of evolution at levels from genes to genomes. In this dissertation, I demonstrate the use of molecular phylogenetics in studying various evolutionary problems, including the phylogeny of individual gene families, the role of gene and genome duplication in organismal evolution, and the reconstruction of the tree of life. In Chapter 2, to study the relationship between gene function and the pattern of gene birth and death, I analyzed the evolutionary history of histone demethylase families as a case study. I found that the two histone demethylase families, KDM1 and JmjC, exhibit distinct evolutionary patterns, which might be explained by the differences in their functions. In Chapter 3, to understand the contribution of gene duplication to the evolution of early eukaryotes, I performed genome-scale analyses of gene family phylogenies to identify gene duplication occurred before the split of animals, fungi and plants. I identified more than 300 early eukaryotic duplications, which possibly resulted from whole genome or segmental duplication(s). The proposed large-scale duplication(s) might provide a genomic basis for the successful radiation of early eukaryotes. In Chapter 4, to identify phylogenetic makers for eukaryotic phylogeny, I systematically identified ~1000 genes that are widely distributed

in eukaryotes and have reasonable orthology. From these genes, I further selected ~30 genes that are highly conserved and single-copy in most sequenced eukaryotic genomes. In addition, I demonstrated the performance of these genes in resolving relationships within and among eukaryotic lineages, including some challenging examples such as the placement of Microsporidia and the monophyly of Excavata. The genes I identified will serve as useful tools in future phylogenomic studies and taxon-rich analyses of the eukaryotic tree of life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Hong Ma, without whom this dissertation would not have been possible. I feel very lucky and privileged to have him as my advisor. To me, he is a role model in many ways. He is a great mentor, an outstanding scientist, and a wonderful person. He always stands by my side, continuously supports and encourages me, and leads me to the right direction. He teaches me that a good scientist should be hard-working and open-minded, and always think positively about the results. What I have learned from him in the past five years will be the treasure of my life. I am also grateful to Dr. Claude dePamphilis, Dr. David Geiser, and Dr. Zhi-Chun Lai, for serving on my committee and providing valuable advices on my research projects and dissertation.

I would also like to thank my colleagues and friends: Dr. Yujin Sun, Dr. Zhenguo Lin, Dr. Yiben Peng, Dr. Zhao Su, Dr. Pingli Lu, Xinwei Han, Xuan Ma, Dihong Lu, Liye Zhang, Yazhou Sun, and Liyana Sukiran, Yi Hu, and Jiong Wang. I greatly appreciate their help and support in many aspects, both within and outside the lab. My thanks also go to Yuannian Jiao for the highly enjoyable collaboration and friendship.

I would like to thank the University Graduate Fellowship, the J. Ben and Helen D. Hill Memorial Fund Award, the Pennsylvania State University Braddock Graduate Fellowship, and the IMEG Travel Grant for providing financial supports during my

doctoral study.

Finally, I must sincerely acknowledge my parents Daochun Zhou and Fengyuan Xie, and my wife Yue Zhao. Their endless love and support has been my source of power, and I am full of gratitude beyond words.

CHAPTER 1

INTRODUCTION: MOLECULAR EVOLUTION OF GENES, GENOMES AND ORGANISMS

1.1 MOLECULAR EVOLUTION OF GENES AND GENOMES

1.1.1 Overview of gene duplication

Genes are the basic units of inheritance and function in all living organisms. Gene duplication generates extra copies of existing genes in the genome, thus directly contributing to the complexity of genomes and organisms. More importantly, the additional gene copies generated by gene duplication provide a major source of raw genetic material for functional innovation and diversification. Therefore, as Susumu Ohno argued in his famous book "Evolution by Gene Duplication" [1], gene duplication has long been considered one of the most important driving forces in evolution.

Almost a century ago, long before any molecular information was available, the potential role of gene duplication in evolution was recognized based on the studies of chromosomal and morphological variations in maize [2] and *Drosophila melanogaster* (in 1930s) [3]. In the decades that followed, numerous observations of gene duplication accumulated from cytological studies, and putative relationships between gene duplication and phenotypic variations were also proposed, providing hints about the evolutionary significance of gene duplication [4]. Subsequently, electrophoretic studies of duplicated isozyme loci provided further evidence for gene duplication (e.g. [5]) and

revealed expression divergence among duplicates (e.g. [6]). Fueled by rapid advances in DNA sequencing technology, our understanding of gene duplication increased at an unprecedented pace in recent years: sequence-based analysis has revealed insights about the tempo, pattern and mechanisms of gene duplication and the subsequent divergence of duplicated genes [4]. In the genomics era, the rapidly increased availability of sequence data also allows thorough investigation of gene duplication in various organisms throughout the evolutionary tree of life [7].

Gene duplication is a ubiquitous phenomenon in all three domains of cellular life including eubacteria, archaebacteria, and eukaryotes. It represents large fractions of genes in eukaryotic and prokaryotic genomes (e.g. ~65% in the *Arabidopsis thaliana* genome) are produced by gene duplication [7]. In addition, gene duplication occurs at a relatively high rate similar to the nucleotide mutation rate; the rate of gene duplication was estimated to be about 0.01 per gene per million years based on genome sequences of selected model organisms in animals, fungi, and plants [8]. In other words, on average, each gene in a genome is duplicated once every 100 million years.

Given the prevalence of gene duplication, the subsequent divergence of duplicates is considered as one major source of adaptive novelties in evolution [9]. For instance, increases in genotypic and phenotypic complexity during eukaryotic evolution are usually associated with the expansion of gene families. It was shown that the diversification of gene families involved in cell differentiation and cell-cell communication contributed to

the origin(s) of multicellularity [10]. In addition, duplications of genes encoding important developmental regulators (e.g. MADS-box genes in plants [11] and HOX genes in animals [12]) have played important roles in the evolution of morphological complexity. Other well-known examples include the rapid duplication of olfactory receptor genes in mammals [13] .

1.1.2 Mechanisms of gene duplication

Several molecular mechanisms of gene duplication have been elucidated and the major ones include tandem duplication, retrotransposition, and large-scale duplication [7]. Tandem duplication is generally thought to result from unequal cross-over and the duplicated genes usually reside in neighboring locations in the genome, until genome rearrangement processes interrupt their adjacency. This mechanism has significant contribution to the evolution of gene families, such as the expansion of the Kelch Repeat-containing F-box family [14] and bHLH transcription factor family [15]. Furthermore, due to the unique spatial arrangement of tandem duplicates, it is suggested that tandem duplication has an important role in the origin and evolution of proteins with repetitive motifs/domains (e.g. β-propeller structure [16]).

Retrotransposition is the process by which the mRNA molecule of a gene is reverse-transcribed and then inserted into a position in the genome that is different from

the location of the original copy. Because they derive from mRNAs, relatively recent duplicates arising from retrotransposition are usually intronless and sometimes have identifiable remnant 3' poly-A sequences and short direct repeats in their flanking regions [7]. It is believed that most duplicates derived from retrotransposition are inserted into genomic regions without promoters and thus not expressed or functional [17]. Therefore, the role of retrotransposition in the evolution of gene families is less well understood compared to the other two types of gene duplication. Increasingly, retrotransposition-based gene duplication events have been documented and the importance of this mechanism has become better appreciated at both gene family and genome levels [18, 19]. In particular, recent studies revealed that retrotransposition contributed to the generation of chimeric genes that have domains from both the retrotransposed gene and the gene near the insertion site [20, 21].

Both tandem duplication and retrotransposition are usually recognized as small-scale duplication events, in which only one or a small number of genes are involved. Large-scale duplications also occur, involving large portions of chromosomes or even entire genomes. Large-scale duplication might result from multiple mechanisms including replication accidents that lead to the duplication of chromosomal segments, failures of chromosomal separation in meiosis that lead to gametes with abnormal numbers of chromosomes (aneuploidy) or completely unreduced gametes (autopolyploidy), and chromosomal doubling in hybrids of closely related species

(allopolyploidy). Large-scale duplications can be studied by analyzing syntenic genomic regions [22, 23], the distribution of synonymous substitution rates (dS) among paralogous genes [24], or phylogenies of a large number of gene families [25].

1.1.3 Significance of large-scale duplication

As a rather dramatic type of genomic change, large-scale duplication is less frequent than small-scale duplication. However, once established, large-scale duplication, for example whole-genome duplication (WGD), will have prominent impacts on evolution at levels from genes to organisms. Large-scale duplication is the major mechanism responsible for the expansion of many gene families [26]. Also, large-scale duplication is of special interest because it allows the generation of multiple new functional modules with many genes that are unrelated at the sequence level [27]. Under the "Gene Balance" hypothesis, the duplication of genes in a functional module (e.g. genes that encode subunits of a protein complex) is constrained by dosage balance effects; only the simultaneous duplication of all the genes in the module, presumably in a large-scale duplication event, is retained [27]. This type of feat of large-scale duplication is not possible for other duplication mechanisms to accomplish.

The numerous duplicated genes generated by large-scale duplications not only allow the evolution of new functions, but also play an important role in promoting speciation

[8]. After large-scale duplications, the different fates of duplicated genes in different populations could generate the genetic diversity that then allows both reproductive isolation/speciation and environmental adaptation [28, 29]. There is increasing evidence that large-scale duplications have occurred in multiple eukaryotic lineages, including animals, fungi, plants, and ciliates [22, 23, 30-34]. For example, it is well accepted that two rounds of WGD occurred in early vertebrates and it is estimated that more than 60% of flowering plants experienced polyploidization in their histories [35]. These large-scale duplications often coincid with rapid species diversification [36, 37], suggesting a possible causal relationship between large-scale duplications and speciation.

1.1.4 Evolutionary fates of duplicated genes

While gene duplication produces extra gene copies and provides a genomic basis for evolutionary change, in most cases this process alone does not create new gene functions (certain exceptions exist, such as the gain of new regulatory elements and/or functional domains during retrotransposition). Instead, the subsequent maintenance and divergence of duplicated genes are essential for the evolution of functional novelties. It is widely accepted that most duplicated-gene pairs are short-lived and one duplicate is usually lost (nonfunctionalization) [1, 8]. Due to functional redundancy following gene duplication, one of the duplicated genes is relieved from selective pressure and can accumulate

mutations. Because most mutations are deleterious, it is very likely that the duplicate will soon lose its function and become a pseudogene (Fig. 1.1).

Many models have been proposed to describe possible evolutionary fates of those pairs of duplicates that are preserved [9]. The classical neofunctionalization model posits that, while one of the duplicates maintains its original function, the other copy can acquire new functions due to adaptive mutations (Fig. 1.1) [1]. According to Ohno, neofunctionalization is responsible for the retention of most duplicated genes [1]. Although a few examples of neofunctionalization have been found [38-40], the primary role of neofunctionalization in preserving duplicates is questioned. On the one hand, neofunctionalization is expected to be rare because the mutational space is dominated by deleterious mutations. On the other hand, the rate of retention after gene duplication is found to be high in many cases [41, 42].

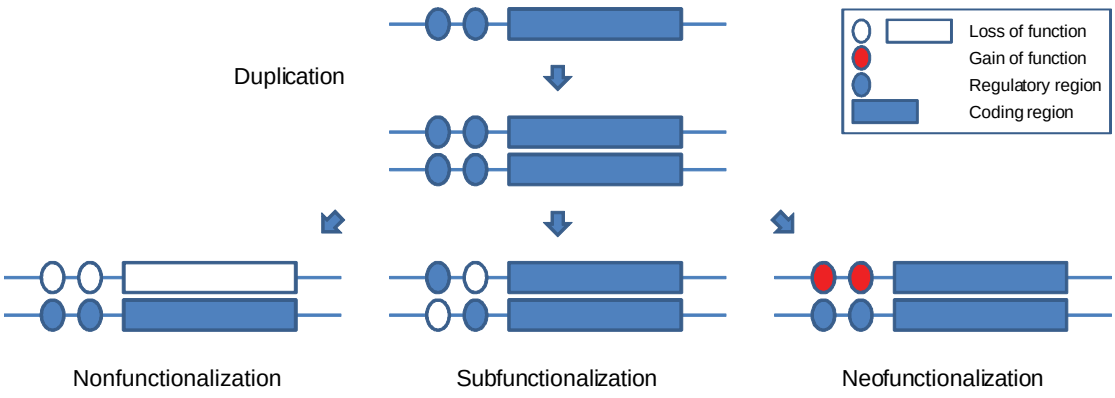


Figure 1.1 Schematic showing evolutionary fates of duplicated genes. Subfunctionalization and neofunctionalization can also occur in coding regions.

Therefore, subfunctionalization is proposed as an alternative mechanism to better explain the preservation of duplicated genes. In their model called "Duplication-Degeneration-Complementation" (DDC), Force et al. described a process in which the two duplicates accumulate different degenerative mutations and each retains a complementary subset of the ancestral functions (Fig. 1.1) [43]. Compared to neofunctionalization, only degenerative mutations are involved in the DDC model. In this case, the duplicate preservation is achieved because both copies are needed to fulfill the functions of the ancestral gene. In certain situations, subfunctionalization can also provide adaptive advantages. For example, in the "Escape from Adaptive Conflict" (EAC) model [44], adaptive mutations are prohibited in the multifunctional ancestral gene as the improvement of one function would compromise the others. After gene duplication, such constraints are removed and individual functions can be optimized in different duplicates [44, 45]. In both DDC and EAC models, subfunctionalization can apply to both expression profiles (in regulatory regions) and protein functions (in coding regions).

It is worth noting that neofunctionalization and subfunctionalization are not mutually exclusive; in fact, it is found that the two processes usually accompany each other and a model called "sub-neofunctionalization" has been formulated accordingly [46]. In addition, there are several alternative views on the evolution of duplicated genes. For example, the preservation of duplicated genes will be favored if: 1) increased gene expression caused by extra copies is beneficial; 2) the duplicated gene is born with a new

function; or 3) the extra copy serves as a genetic backup and masks the negative effect of deleterious mutations [9].

1.1.5 Evolution of multigene families

A direct consequence of gene duplication is the widespread existence of multigene families [1, 17]. A multigene family consists of homologous genes that are derived from the same ancestral gene. Depending on how two homologous genes diverge from each other, their relationship can be classified as either orthologous, if they separate via a speciation event, or paralogous, if they originate through a gene duplication event [47]. Such distinctions are important because they are not only relevant to our understanding of evolutionary relationships among homologous genes, but also helpful in the prediction of gene function [48]. For example, although not always true, it is common to assume that orthologs rather than paralogs perform the same or analogous function(s) in different organisms [49].

In recent years, active efforts have been made to understand the evolution of multigene families [13]. It is now clear that most eukaryotic multigene families are subject to birth-and-death evolution, in which new family members are generated by gene duplication, while some existing members become pseudogenes and are eventually eliminated from the genome [13]. In addition, phylogenetic studies of individual

multigene families have revealed a broad spectrum of rates of gene birth-and-death [14, 18, 50-55]. A relationship between evolutionary patterns and functions of multigene families has also been observed in some gene families.

On one end of the spectrum are multigene families that have strikingly low birth-and-death rates throughout eukaryotic history. Among these families are the *recA/RAD51* family [50], the *MutL* family [51], the *MutS* family [51], the *SMC* family [52], the *MCM* family [53], the *CCT* family [54], and others. These families experienced gene duplications before the divergence of major eukaryotic lineages and maintained highly stable copy numbers afterwards. Interestingly, these families all perform functions in fundamental processes such as DNA replication (*MCM*), DNA repair and recombination (*recA/RAD51*, *MutL* and *MutS*), maintenance of chromosome structure (*SMC*), and protein folding (*CCT*). It is possible that the maintenance of a stable copy number in these gene families is partly due to their functional essentiality.

In contrast, many families that have regulatory functions in physiology and biochemistry are found to have very high birth-and-death rates. For example, the *F-box* family and the *SKP1* family encode subunits of the SCF ubiquitin ligase complex [56], which is important for post-translational regulation of cellular proteins. Both families experienced rapid expansion in angiosperms [14, 18, 57]. F-box and SKP1 proteins determine the substrate specificity of SCF complexes. It is possible that the large number of new *F-box* and *SKP1* genes allows functional divergence (e.g. the recognition of new

or more specialized targets) and leads to increased regulatory complexity in plants. Interestingly, the rates of copy number change also vary considerably among different *F-box* subfamilies [57]. Subfamilies with more conserved functions tend to have stable copy numbers, while subfamilies with more divergent functions exhibit rapid gene gain. This again suggests a relationship between evolutionary pattern and function of multigene families.

1.2 CLASSIFICATION AND EVOLUTION OF ORGANISMS

1.2.1 Current understanding of eukaryotic tree of life

To understand the evolutionary relationships among organisms is a central goal of evolutionary biology. It has obvious significance in cataloging biodiversity and is of fundamental importance to almost all aspects of evolutionary biology. Examples include, but not limited to, reconciliation of gene trees with species trees, reconstruction of ancestral status of characters, inference of timing of evolutionary events, and measurement of mode and tempo of evolution [58]. Furthermore, the phylogenetic inference also provides benefits to many aspects of biological studies. For instance, the resolution of relationships between important crop plants and their close relatives can provide valuable information for crop improvement [59]. Phylogeny can also reveal co-evolutionary relationships between pathogens and hosts, which in turn provides beneficial information for disease diagnosis and treatment [60]. As the amount of genomic data increases at an explosive rate, it is also of great importance to have accurate organismal phylogenies to guide the analyses of available information.

Traditionally, organisms were classified mainly based on similarities and differences in morphological and other phenotypic characteristics [61]. In the 1920s, cellular

organisms were classified into two categories: prokaryotes, which are morphologically simple organisms that lack the nucleus; and eukaryotes, which are complex single- or multi-cellular organisms that have a nucleus and organelles (e.g. mitochondria and chloroplasts) [62]. Technical advances in the 1970s enabled people to obtain nucleic acid sequence information from organisms, and the use of this information has revolutionized the understanding of the tree of life [63]. In the early 1990s, the small subunit ribosomal RNA gene was widely used in tree of life studies because it is universally distributed and has a slow evolutionary rate [61]. Based on the phylogenetic analyses of 16S (18S) rRNA sequences, Woese and Fox introduced the three-domain system which divides all cellular organisms into three domains: eubacteria, archaea, and eukaryotes (Fig. 1.2) [64].

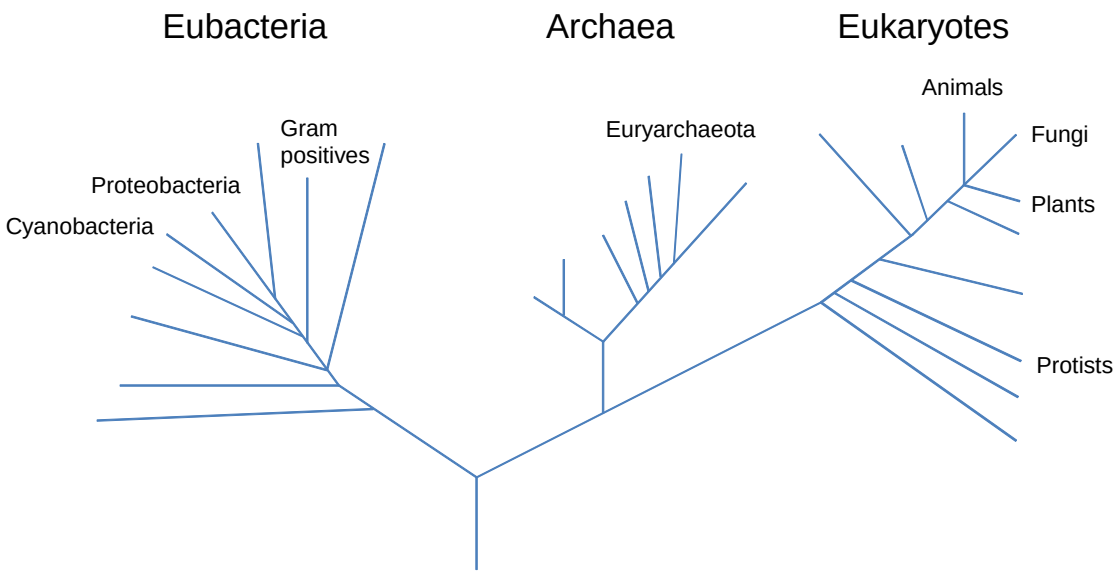


Figure 1.2 The phylogenetic tree of life based on 16S rRNA sequences. (Modified from Woese et al. [65]).

Eukaryotes consist of organisms with astonishing morphological diversity. Perhaps

because of difficulties of interpreting this vast diversity, traditional classification based on morphology has poor resolution for the relationships among major eukaryotic lineages. According to the traditional five-kingdom system, eukaryotes that have simple organizations (single-cellular, or multi-cellular but lacking tissue) but do not belong to animals, fungi or plants are all included in a single kingdom called "Protista" [66]. Similar to the case of eukaryotes-prokaryotes division, the understanding of the eukaryotic tree of life has also been greatly improved by molecular phylogenetics. In the early 1990s, the "crown-stem" model (Fig. 1.3A) of eukaryotic phylogeny was proposed based on the study of small-subunit ribosomal RNA sequences [67, 68]. This "crown-stem" model suggests that protists do not form a monophyletic group. Instead, they occupy the lowermost and middle branches on the stem of the eukaryotic tree of life. Plants, animals, and fungi are nested in a crown of the eukaryotic tree and they separated from each other more recently than early branching protists.

More recently, an alternative view of the early evolution of eukaryotes has emerged from phylogenomic studies and is increasingly accepted [69]. According to this view, eukaryotes are classified into six supergroups (Fig. 1.3B): Opisthokonta, Amoebozoa, Archaeplastida, Chromalveolata, Excavata, and Rhizaria. The Opisthokonta contains animals, fungi and several closely related protists (e.g. choanoflagellates, the closest protist relatives of animals) [70]. The Amoebozoa consists of many amoeboid species and slime molds [70]. Some famous representatives in this group are *Dictyostelium*

discoideum and *Entamoeba histolytica*. The Opisthokonta and the Amoebozoa together comprise one major clade of eukaryotes called unikonts, which are characterized by the presence of a single flagellum at some stage of their life cycles [71].

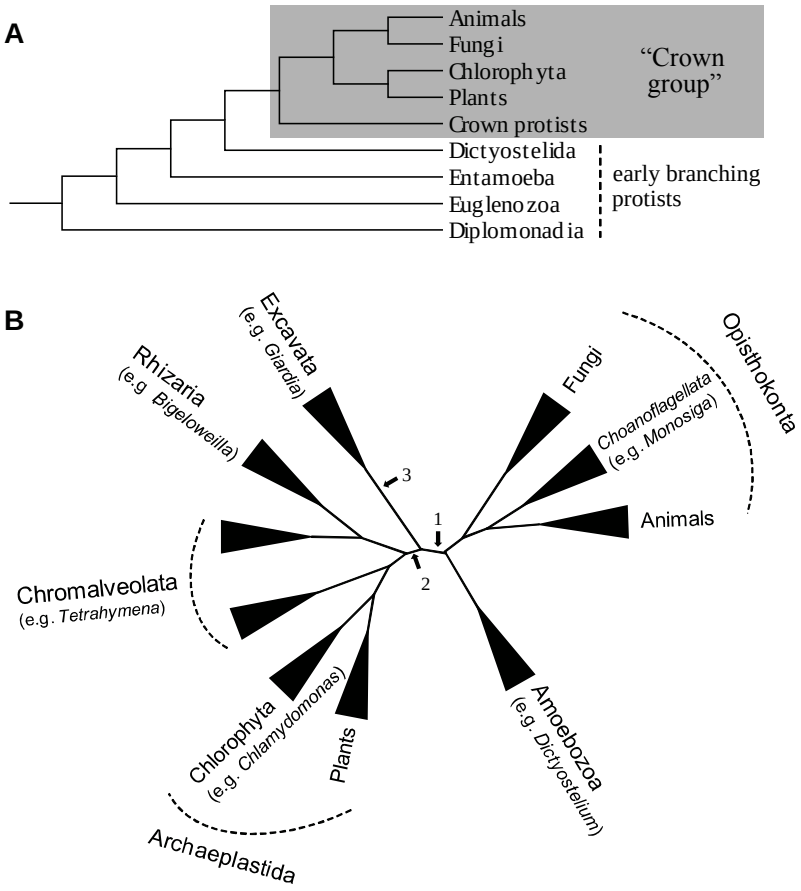


Figure 1.3 Alternative views of the eukaryotic phylogeny. (A) The "crown-stem" topology of eukaryotic phylogeny. The topology shown is adopted from Sogin [68] and Sogin et al. [72]. (B) The "six supergroups" classification of eukaryotes; the topology shown was reported by Hampl et al. [73]. Different hypotheses about the root position of the eukaryotic tree are indicated by numbered arrows: 1, the unikont–bikont hypothesis [74, 75]; 2, the photosynthetic-nonphotosynthetic scenario [76]; 3, Excavata as basal group [77]. The branch lengths are arbitrary.

The remaining four supergroups comprise the other eukaryotic clade called bikonts, members of which are thought to have two flagella in their evolutionary histories [71]. The supergroup Archaeplastida unites all the eukaryotes that acquired their plastids through primary endosymbiosis, including plants, green algae, red algae, and glaucophytes [70]. Several other groups of photosynthetic eukaryotes, including dinoflagellates, cryptophytes, haptophytes, and stramenopiles, are considered as secondary algae whose plastids originated from the secondary endosymbiosis of a red alga [70]. These secondary photosynthetic lineages and their non-photosynthetic relatives (ciliates and apicomplexa) form the supergroup Chromalveolata [70]. The Excavata is a group of ancient protists that include members with complex flagella but without functional mitochondria [70]. Many members of this supergroup, such as *Giardia* and *Trichomonas*, are important pathogens and form the most basal branches in the "crown-stem" model [67]. Last but not the least, the supergroup Rhizaria is recently proposed based on molecular data and includes single-cellular eukaryotes with considerably diversified forms [67].

These supergroups would have diverged during the early phase of eukaryotic evolution, which is sometimes described as a "Big Bang" event [78]. The time and order of the divergence of these supergroups remain controversial and different root positions of the eukaryotic tree have been proposed [74-77]. One such scenario is the already mentioned "unikonts-bikonts" division, which places the root between

Opisthokonta-Amoebozoa and other supergroups [71]. This rooting is supported by the fusion of two essential enzyme encoding genes in bikonts and a unique insertion in the myosin type II head domain in unikonts [74, 75]. Alternatively, the eukaryotic tree of life can be rooted between “photosynthetic” and “non-photosynthetic” organisms, as suggested by the study of rare genomic changes [76]. In this scenario, the Archaeplastida which acquired plastids through primary endosymbiosis of cyanobacteria are separated from all the other supergroups at the very beginning of eukaryotic evolution. It is also possible that the Excavata occupies the most basal position on the tree of eukaryotes [79].

The understanding of the eukaryotic tree of life is still developing, as more information becomes available and phylogenetic approaches improve. Recent studies further suggest that the number of supergroups may be more than six [73, 80]. In addition, increasing evidence suggests that some of the supergroups may not be monophyletic. For example, several recent studies lend support to a position of haptophytes within or close to the Archaeplastida instead of the Chromalveolata [73, 80]. Further efforts are still needed to fully elucidate the relationships among and within these eukaryotic supergroups.

1.2.2 Approaches for the estimation of organismal phylogeny

In the early days of molecular phylogenetics, the relationships among organisms

were mostly studied by using single or a few genes, such as small subunit ribosomal RNA genes [68, 81] and later, protein-coding genes such as translation elongation factor genes [82, 83]. Such single- or few-gene based studies have provided important insights into the understanding of the evolutionary histories of various organisms including the eukaryotes. However, gene-scale phylogenies usually have low resolution for deeper relationships, and different datasets often reveal conflicting topologies, mainly due to the weak or biased signal of phylogenetically informative characters carried by single genes (also known as stochastic error) [84].

The burst of sequence data in recent years allows the use of genome-scale data in the inference of organismal phylogeny, which is called the phylogenomic approach [84]. Phylogenomic studies usually include hundreds or even thousands of genes and these many genes can be analyzed in two different ways. In the supermatrix method, orthologs of each gene are retrieved and aligned separately; the alignments are concatenated to form a supermatrix, which is then analyzed by using standard methods of phylogenetic reconstruction. Alternatively, in the supertree method, individual gene phylogenies (including all homologous genes) are estimated first and then combined into a supertree. Furthermore, other methods are available involving genome-wide comparison of gene order, gene content, and distribution of DNA or protein strings. It is thought that the supertree method and whole genome feature-based methods are not sophisticated or accurate as the supermatrix method because they are relatively new and have not been

extensively evaluated [84].

Phylogenomic studies can overcome stochastic error by including sufficient data, but still might generate incorrect phylogenies due to the violation of phylogenetic assumptions (systematic error; such as long-branch attraction) [85], partly due to limited taxon sampling in such genome-scale analyses. On the other hand, traditional gene-scale analyses allow broad taxon sampling (e.g. over a thousand plant species [86]) which is helpful for reducing systematic error, but produce poorly resolved relationships due to limited phylogenetic information [85]. In fact, the relative importance of more genes or more taxa has been hotly debated but no consensus is reached yet [87, 88]. Although datasets rich in both characters and taxa are desirable, it is impractical to acquire genome-scale data for a large number of species.

To achieve better taxon sampling, a number of recent phylogenomic analyses have included EST sequencing with a small number of species of key importance (e.g. [73, 89-92]). These phylogenomic analyses rely on more than one hundred orthologous genes that are selected a priori. Usually a large fraction of these orthologous genes are covered in EST projects, allowing the inclusion of new species in phylogenomic analyses. Alternatively, broad taxon sampling combined with a moderate number of genes has recently proven an effective strategy to alleviate both stochastic and systematic errors in the reconstruction of organismal phylogenies [93, 94]. A moderate number of genes can provide much greater phylogenetic resolving power than single genes, sufficient for

constructing the tree of life [95]; in addition, the number of sequences is small enough to be more rapidly and cost-effectively obtained from many species than those required for genome-scale phylogenomic analyses.

CHAPTER 2

EVOLUTIONARY HISTORY OF HISTONE DEMETHYLASE FAMILIES: DISTINCT EVOLUTIONARY PATTERNS SUGGEST FUNCTIONAL DIVERGENCE

The work described in this chapter has been published in Zhou et al, *BMC Evol. Biol.*, 8:294**.**

2.1 SYNOPSIS

Histone methylation can dramatically affect chromatin structure and gene expression and was considered irreversible until recent discoveries of two families of histone demethylases, the KDM1 (previously LSD1) and JmjC domain-containing proteins. These two types of proteins have different functional domains and distinct substrate specificities. Although more and more KDM1 and JmjC proteins have been shown to have histone demethylase activity, our knowledge about their evolution history is limited. We performed systematic phylogenetic analysis of these histone demethylase families and uncovered different evolutionary patterns. The *KDM1* genes have been maintained with a stable low copy number in most organisms except for a few duplication events in flowering plants. In contrast, multiple genes for JmjC proteins with distinct domain architectures were present before the split of major eukaryotic groups, and experienced subsequent birth-and-death evolution. In addition, distinct evolutionary patterns can also be observed between animal and plant histone demethylases in both families. Furthermore, our results showed that some *JmjC* subfamilies contain only animal genes with specific demethylase activities, but do not have plant members. Our study improves the understanding about the evolutionary history of *KDM1* and *JmjC* genes and provides valuable insights into their functions. Based on the phylogenetic relationship, we

discussed possible histone demethylase activities for several plant JmjC proteins. Finally, we proposed that the observed differences in evolutionary pattern imply functional divergence between animal and plant histone demethylases.

2.2 INTRODUCTION

One important mechanism for eukaryotic gene regulation is the epigenetic regulation of chromatin structure. The basic unit of chromatin is the nucleosome, which consists of 146bp of DNA wrapped around an octamer of four histone proteins, H2A, H2B, H3, and H4. Histone proteins can be modified on the N-terminal tail and the modifications can disrupt the interaction between nucleosomes to prevent the packaging of chromatin into higher order structures; also the modified tails can serve as binding sites for chromatin modifiers, facilitating their functions [96]. Histone modifications, such as methylation and acetylation, have been well studied and many of the sites for the modifications are known [96]. For example, methylation can take place on several lysine residues on histone H3 and H4 (H3K4, H3K9, H3K27, H3K36, etc.) and each lysine residue can be mono-, di- or trimethylated. Histone arginine residues like H3R2 and H4R3 can also be mono- or dimethylated. According to the histone code hypothesis, different histone modifications are linked to distinct functional outcomes: H3K4 and H4K36 methylations are mainly associated with active genes while methylated H3K9 and H3K27 are markers for the repressed chromatin in general [96, 97].

As important mechanisms of gene regulation, histone modifications themselves are under precise control [1]. It is known that many histone modifications are dynamically

regulated by enzymes which add or remove the chromatin modifications, with defects in either of these two functions resulting in incorrect activation or repression [96]. However, histone methylation was considered irreversible for a long time. Although histone methylation was first reported in 1964 and the first histone methyltransferase was discovered in 2000 [98, 99], it was not until 2004 that KDM1 [histone lysine (K) demethylase 1; previously known as LSD1 (Lysine specific demethylase)] was identified as the first histone demethylase [100]. KDM1 contains a C-terminal amine oxidase (AOD) domain, which is responsible for the demethylase activity through a flavin adenine dinucleotide (FAD)-dependent mechanism, and an N-terminal SWIRM domain also found in other chromatin regulators [100]. Several studies showed that the SWIRM domain is important for the stability and chromatin targeting of KDM1 [101-103]. Since the chemical mechanism of KDM1 mediated demethylation requires a protonated nitrogen for the reaction to proceed, the substrate specificity of KDM1 is limited to mono- or dimethylated lysine residues [104]. Types of histone methylation shown by biochemical studies to be demethylated by KDM1 include H3K4me1/2, and in the presence of androgen receptor (AR), H3K9me1/2, representing a small subset of all the possible states of histone methylation [105].

Soon after the identification of KDM1, the Jumonji C (JmjC) domain-containing proteins were discovered to be another family of histone demethylases [106]. The JmjC domain is the catalytic domain and these proteins belong to the Cupin superfamily of

Fe(II) and α-ketoglutarate dependent dioxygenases [107]. Unlike KDM1, the JmjC domain-containing proteins that have been tested do not require a protonated nitrogen and are able to reverse all three states of lysine methylation [104]. Members in this family have been shown to be able to remove the methyl groups on H3K4, H3K9, H3K27 and H3K36 [108]. Furthermore, a protein in this family, the JMJD6, functions as a histone arginine demethylase through a similar chemical mechanism [109]. JmjC proteins usually contain additional domains, which are involved in the recognition of methylation (e.g. PHD and Tudor), protein-protein interaction (e.g. F-box) and DNA binding (e.g. C2H2 zinc finger), suggesting a wide range of possible functional interactions.

The number of studies of histone demethylases is increasing rapidly in recent years, with members in both families shown to have important biological functions. *KDM1* is an essential gene in mouse [110] and important for viability and fertility in *Drosophila* [111]. The *Arabidopsis* homologs of *KDM1*, including *Flowering Locus D* (*FLD*), regulate the transition to reproductive development [112-115]. Moreover, the JmjC domain-containing proteins are involved in a broad range of processes. For example, the newly identified H3K27 demethylases, UTX and JMJD3, play important roles in regulating *Hox* gene expression and the animal body development [116, 117]. In addition, JMJD3 was suggested to function in the neural stem cell differentiation [118]. Other JmjC domain-containing proteins are involved in processes such as the X-linked neural development (JARID1C) [119, 120] and embryonic stem cell self-renewal (JHDM2A and

JHDM3C) [121].

While these studies greatly advanced our understanding about the molecular and biological functions of histone demethylases, they only covered a limited fraction of the proteins in the two histone demethylase families. A large number of KDM1 and JmjC-containing proteins remain to be functionally characterized, especially in plants. There are only very few studies on plant histone demethylases. In addition to *FLD* and its two relatives, only three JmjC domain-containing proteins in *Arabidopsis* have reported functional studies [122, 123], although it is reasonable to expect that the plant histone demethylases have important functions.

Phylogenetic analyses can provide useful information about evolutionary relationship among related genes from different organisms and clues about possible functions of genes closely related to those with known functions. Furthermore, the differences in evolutionary pattern between gene families or species also suggest different evolutionary pressures and diverged functions. Homologs of both types of histone demethylase have been detected in major groups of eukaryotes [100, 107]. However, to our knowledge, there is no detailed phylogenetic analysis on the KDM1 proteins and only one report exists for JmjC domain-containing proteins from fungi and animals [124]. To gain a better understanding of the evolutionary history of these two histone demethylase families, we performed systematic phylogenetic analyses in this study including sequences from eukaryotes and bacteria.

2.3 MATREIALS AND METHODS

2.3.1 Data retrieval

The amino acid sequences of the AOD domain in reported KDM1 histone demethylases were retrieved from National Center for Biotechnology Information (NCBI). They were used as queries to search against NCBI, TAIR, TIGR and JGI databases for all possible AOD domain-containing proteins in selected eukaryotic organisms by using TBLASTN with e-value less than e^{-5} as cut-off. All the new results were used as queries to carry out a second round of BLAST search, until no new sequence was found. The collected protein sequences were then analyzed by SMART and Pfam for domain architecture. The proteins which lack the AOD domain or have an AOD domain with e-value greater than e^{-10} based on both SMART [125] and Pfam [126] results were excluded from the further analyses. The prokaryotic sequences were retrieved from NCBI database through BLASTP by using eukaryotic AOD domain-containing proteins as queries and e^{-5} as cut-off. The same procedure was followed for the retrieval of JmjC domain-containing proteins. Common names for the following species are shown in the figures: Arabidopsis, *Arabidopsis thaliana*; Poplar, *Populus trichocarpa*; Rice, *Oryza sativa*; Moss, *Physcomitrella patens*; Human, *Homo sapiens*; Cow, *Bos taurus*; Mouse, *Mus musculus*; Zebrafish, *Danio rerio*; Fruitfly,

Drosophila melanogaster; Mosquito, *Anopheles gambiae*; Honey bee, *Apis mellifera*; Beetle, *Tribolium castaneum*; Sea squirt, *Ciona intestinalis*; Sea urchin, *Strongylocentrotus purpuratus*; and Sea anemone, *Nematostella vectensis*.

2.3.2 Sequence alignment

A preliminary multiple sequences alignment (MSA) was generated using MUSCLE 3.6 [127] with the default settings and a Neighbor-Joining (NJ) tree was constructed using MEGA 4.0 [128] based on the MSA. According to the tree topology, the sequences were divided into several subgroups. Each subgroup of sequences was aligned by MUSCLE 3.6 separately followed by manual adjustment using GeneDoc 2.6.0.3 [129]. These alignments were then combined using the profile alignment function of ClustalX 1.83 [130]. The codeml program from the PAML 4.1 package is used for the Ka/Ks analyses [131].

2.3.3 Phylogenetic analysis

Both NJ and Maximum likelihood (ML) methods were used to perform the phylogenetic analyses. NJ trees were constructed using MEGA 4.0 with “pairwise deletion” option and “Poisson correction” model. Bootstrap test of 1000 replicates was

carried out to evaluate the reliability of internal branches. ML trees were generated using PHYML 2.4.4 [132] with 100 nonparametric bootstrap replicates. ProtTest 1.4 [133] was used to select the model and parameters for the ML analysis. In this study, WAG amino acid substitution model was used and both proportion of invariable sites and gamma distribution parameter were estimated from the data. In this study, we presented only the NJ trees with bootstrap values from both NJ and ML analyses.

2.4 RESULTS AND DISCUSSION

2.4.1 Distribution of AOD domain-containing proteins in major lineages

Since the AOD domain is the catalytic domain in the KDM1-type histone demethylases, we collected gene sequences for AOD domain proteins from selected animal, plant and fungal species following the procedure described in Methods. In total, 118 sequences were retrieved from 12 organisms (Table 2.1). The *AOD* genes are present in Eukaryotes and Eubacteria, but absent in Archaea. In this study, all *AOD* genes were named based on their domain structure. The genes which encode proteins with only the AOD domain were named as *AOD* genes, whereas the genes coding for proteins with both the SWIRM and the AOD domain were named as *KDM1*. *KDM1* genes only exist in Eukaryotes, and account for only a small fraction of the *AOD* genes (e.g., 2/8 in human and 4/14 in *Arabidopsis*). The *KDM1* genes have maintained a constant copy number of two in most animal species from the basal invertebrate sea anemone to human, except for insects and several nematodes, which contain one and three copies, respectively. A different trend was observed in plants. The number of *KDM1* genes increased from 2 in green algae to 4 in *Arabidopsis* and rice, with the highest number of 7 in poplar (*P. trichocarpa*), which is thought to have experienced a relatively recent genome-wide

duplication [134].

In fungi, *KDM1* was found in the fission yeast *Schizosaccharomyces pombe* but not the budding yeast *Saccharomyces cerevisiae* [100]. To investigate the distribution

Table 2.1 Number of AOD domain-containing genes and JmjC domain-containing genes included in this study

Organism	AOD domain-containing gene		JmjC domain-containing gene
	KDM1	*AOD*	
Human	2	6	26
Zebrafish[b]	2	5	
Drosophila	1	7	12
C. elegans	3	4	12
Mouse[a]	2		
Pufferfish[a]	2		
Sea squirt[a]	2		
Sea urchin[a]	2		
Mosquito[a]	1		
Honey bee[a]	1		
Beetle[a]	1		
Sea anemone[a]	2		
Arabidopsis	4	10	19
Poplar	7	17	25
Rice	4	10	15
Selaginella[b]	2	10	
Moss	3	10	14
Ostreococcus[b]	2	6	
Fission yeast	2	0	5
Budding yeast	0	1	4

a. Only *KDM1* genes are collected in these organisms.
b. Only *AOD* genes are collected in these organisms

of *KDM1* in fungi, we searched for *KDM1* genes in completely sequenced fungal genomes in the NCBI database. Our phylogenetic analysis of the fungal *KDM1* sequences (Appendix Fig. 2.1) indicates that one *KDM1* gene was present in the ancestor of Ascomycota and it was lost in the common ancestor of the budding yeast and *Candida albicans* after its divergence from *Y. lipolytica*. In fission yeast, the two *KDM1* genes were shown to have important functions in regulating heterochromatin [135, 136], which is marked by H3K9 methylation. In contrast, the budding yeast does not possess H3K9 methylation and employs a different set of proteins to fulfill the function of fission yeast *KDM1* genes in heterochromatin regulation [137]. Furthermore, in the absence of *KDM1* homolog, The H3K4 demethylation in the budding yeast is performed by a JmjC domain protein which will be discussed later.

2.4.2 Phylogenetic analyses of *AOD* genes

To investigate the evolutionary history of *AOD* genes, we carried out phylogenetic analyses with sequences from representative species using both NJ and ML methods, yielding very similar results. The phylogenetic tree (Fig. 2.1) indicates that all *KDM1* genes form a single clade with 90/83 bootstrap support. Within this clade, the animal *KDM1* genes form two highly supported (100/100) groups, each contains one *KDM1* gene from the two vertebrates, human and zebrafish. The only *Drosophila KDM1* gene is

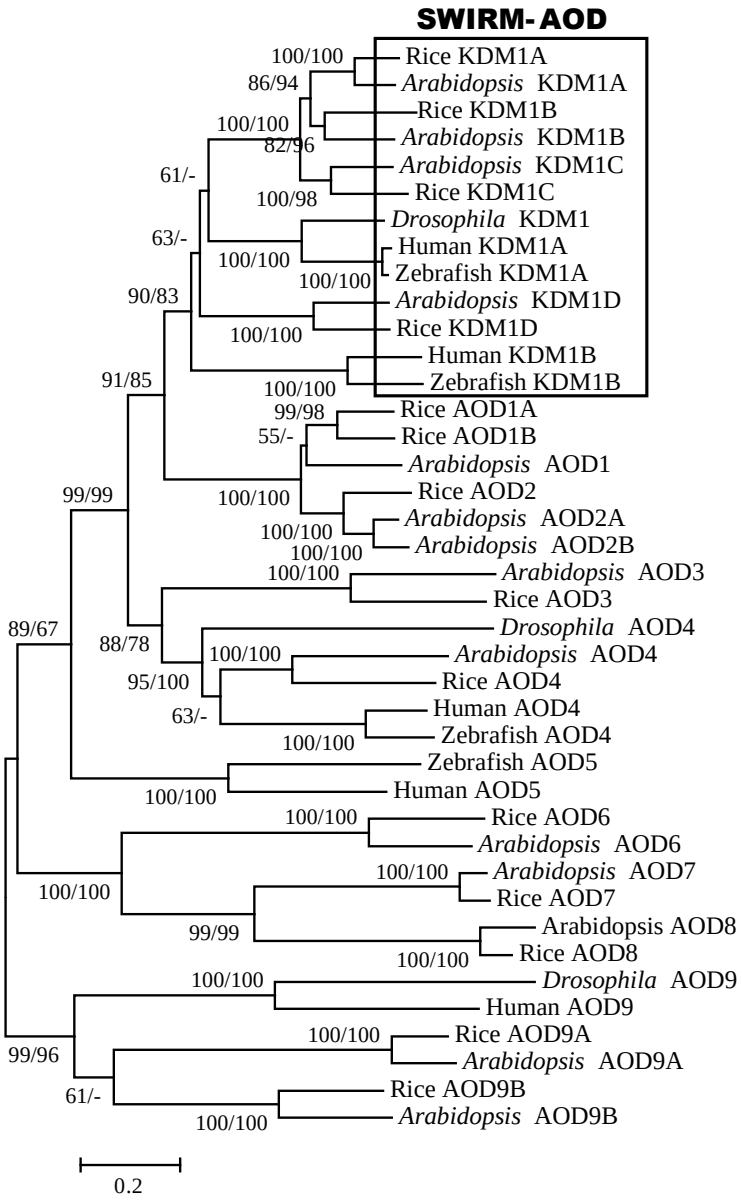


Figure 2.1 Phylogenetic tree of *AOD* genes from representative plant and animal species using the AOD domain region. Both NJ and ML methods were used to infer the evolutionary history, and only the NJ tree is shown. NJ/ML bootstrap values are presented for clades with support greater than 50%. All *KDM1* genes which code for proteins with both SWIRM and AOD domains form a single clade, which is highlighted in the box. The remaining *AOD* genes form six well supported clades.

in the same group as the human *KDM1A* gene. Similarly, the plant *KDM1* genes are also

divided into two separate groups, each with 100/100 support. The relationship between these animal and plant groups is unclear since the topology lacks strong bootstrap support. However, our results still suggest an early origin of *KDM1* genes prior to the divergence of animals and plants. Besides the *KDM1* clade, there are six major clades of *AOD* genes. One of these clades contains both animal and plant *AOD* genes, three are plant specific and one is animal specific. Based on these results, it could be estimated that, in the most recent common ancestor of animals and plants, there were at least one *KDM1* gene and six additional *AOD* genes.

Previous studies showed that the human KDM1A protein has an insertion in the AOD domain [100]. The insertion forms a coiled-coil protruding from the AOD domain and is required for the binding between human KDM1 and CoREST [101, 103, 138]. The alignment of AOD amino acid sequences showed that this insertion is conserved among animal KDM1A. The fungal KDM1 proteins also have an insertion at the same position, but the sequences are not similar to the animal insertions. Insertions of much shorter length can also be detected in several plant KDM1 proteins. By contrast, no insertion was found in other AOD proteins.

We used the COILS program to test whether the insertions in different KDM1 proteins are able to form a coiled-coil structure. Consistent with the crystal structure, the insertions in the human KDM1A protein is predicted to form a coiled-coil structure with high support. The same results were obtained for other animal KDM1A proteins,

suggesting that the interaction between human KDM1A and CoREST might be conserved in all animals. The lack of insertion in animal KDM1B suggests a functional divergence between these two proteins. Interestingly, although the fungal KDM1 proteins possess an insertion, no coiled-coil is predicted. Several studies showed that the two *S. pombe* KDM1 proteins form a complex with two PHD domain-containing proteins [139]. Hence, unlike their counterparts in animal, the insertions in fungal KDM1s might be involved in the interaction with these PHD proteins or have other functions. The absence of the insertion in other KDM1 proteins suggests that this insertion is not essential for the histone demethylase activity. Alternatively, the KDM1 proteins without the insertion might have different activities.

2.4.3 Plant and animal *KDM1* genes have different evolutionary patterns

To further understand the evolution of *KDM1* genes in different lineages, *KDM1* genes from more species were included in the phylogenetic analysis. A representative phylogenetic tree shown in Fig. 2.2 has high bootstrap supports for the two animal clades and two plant clades of *KDM1* genes. In this tree, the plant group I and animal *KDM1A* group cluster together to form a clade with 97/86 bootstrap support. The plant group II is placed outside this clade, and the animal *KDM1B* group occupies the basalmost position in the *KDM1* clade. However, while the position of plant group II is highly supported

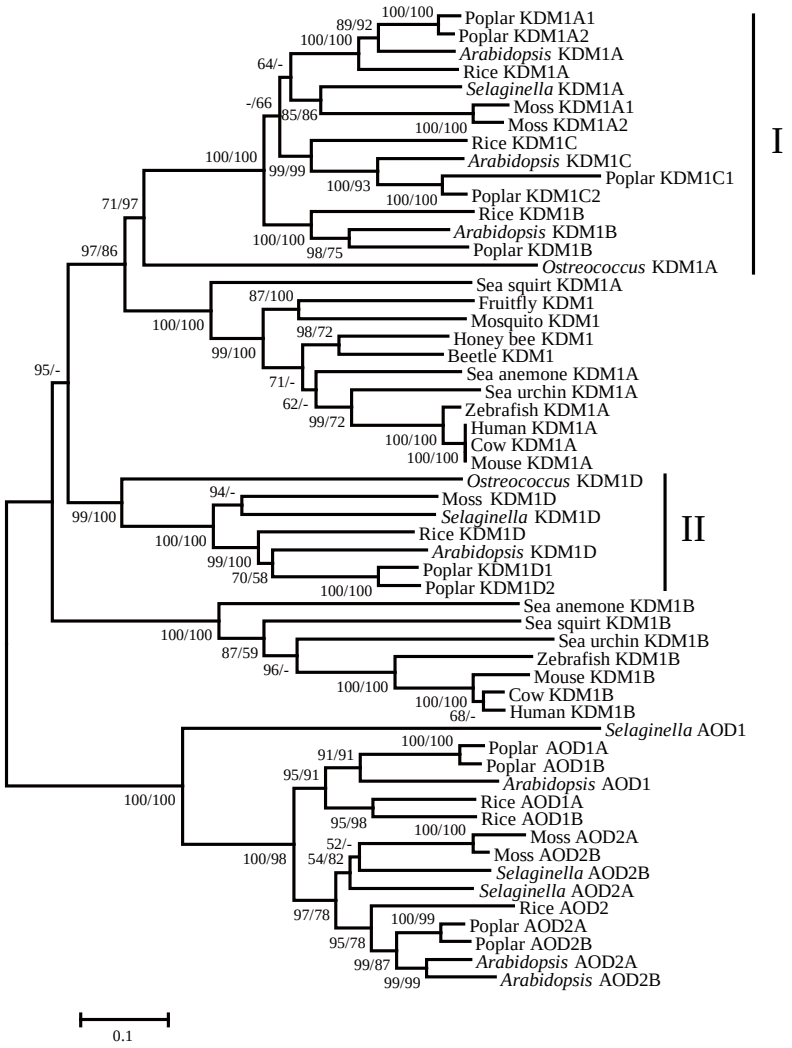


Figure 2.2 Phylogenetic tree for *KDM1* genes with *AOD1* and *AOD2* genes as outgroup. The tree was constructed using the AOD domain region. Plant and animal *KDM1* genes each form two separate clades; the plant specific group I and group II, and the animal specific KDM1A and KDM1B. The methods for tree construction and bootstrap values are given as in Fig. 2.1.

in the NJ tree, it has no support from the ML method. This discrepancy between the bootstrap values from two methods might be due to the long branches of the animal *KDM1B* genes. Therefore, according to these results, there were at least two copies of

KDM1 genes present in the most recent common ancestor of animals and plants. Furthermore, the inclusion of additional sequences revealed distinct evolutionary patterns of animal and plant *KDM1* genes. The animal *KDM1A* and *KDM1B* genes both maintain only one copy in most animals. However, *KDM1B* was not found in insects, implying that it was lost in the ancestor of insects. By contrast, the plant group I contains three subgroups and each subgroup consists of genes from monocots and eudicots, indicating the presence of three copies of group I *KDM1* genes in the most recent common ancestor of angiosperms. Due to the lack of complete genomic sequence and EST data, we did not detect sequences from gymnosperms. Hence it is unclear how many *KDM1* genes were present in the ancestor of seed plants. However, the basalmost position of the green algae *KDM1A* in this group suggests that all members in group I were derived from a single copy of *KDM1* in the ancestor of green plants. In addition, lineage specific duplication events were found in moss and poplar, as well as in group II.

Several genome-wide studies in fungi and *Drosophila* suggested that evolutionary patterns of gene families are correlated to their functions [140, 141]. The genes with low volatility in copy number during evolution are usually associated with essential functions. In fact, the *KDM1A* genes have been shown to be essential in mouse and *S. pombe* and are involved in important biological processes like meiotic progression and spermatogenesis [135, 142]. Although the function of animal *KDM1B* genes is not known, the similarity of their evolutionary pattern to that of *KDM1A* also implies functional

conservation and importance. Consistent with this idea, the residues critical for cofactor binding and catalytic activity are conserved in animal KDM1B proteins, suggesting that they have histone demethylase activity. Moreover, the expression of animal *KDM1B* gene is supported by considerable amount of EST data, although less abundant than *KDM1A*.

However, several potential substrate-binding residues are substituted in animal KDM1B, suggesting possible changes in substrate specificity of these proteins. Other lines of evidence also support the functional divergence between animal *KDM1A* and *KDM1B* genes. Besides the SWIRM and the AOD domain, the animal KDM1B proteins also contain a CW-type zinc finger near the N-terminus. The function of this zinc finger is not well characterized, but it is usually found in proteins which also have other domains involved in DNA binding or protein-protein interaction [143]. Interestingly, this domain is also found in a class of SET domain histone methyltransferases (HMTs), which have H3K36 methyltransferase activity [144]. Therefore, the zinc finger possibly facilitates the recognition of substrates other than methylated H3K4 and H3K9 by KDM1B. Furthermore, the tree in Fig. 2.2 also shows that the animal *KDM1B* genes have branches longer than those of *KDM1A*, indicating that the *KDM1B* genes have evolved at higher rates. To test this idea, we also conducted Ka/Ks analyses for several pairs of animal genes. The results indicate that: (1) both *KDM1A* and *KDM1B* genes were under purifying selection with Ka/Ks ratio lower than 0.1; (2) Ka/Ks values for *KDM1B* genes were significantly higher than those for *KDM1A* genes, indicating that the *KDM1B* genes

have evolved under less stringent selective pressure (Appendix Table 2.1). As all these results point to a functional divergence between animal *KDM1A* and *KDM1B*, it will be worth investigating the functions of KDM1B proteins in the future.

In plants, our results showed that the copy number of group I *KDM1* gene increased from one in the common ancestor of green plants to three in the common ancestor of flowering plants. The functional studies of *AtKDM1A*, *AtKDM1B* and *AtKDM1C* revealed that all three genes regulate the transition to reproductive development [112, 113]. It is possible that the expansion of plant group I might have contributed to the evolutionary success of flowering plants. The initiation of reproductive development is one of the most important developmental events in plants and is regulated by a complex regulatory network [145]. According to the duplication-degeneration-complementation (DDC) model [43], the duplicate group I *KDM1* genes would have undergone sub-functionalization or neo-functionalization, which might help to optimize the regulatory network controlling flowering. In fact, functional studies showed that these three genes have partially redundant functions in the repression of the expression of *FLC*, a major inhibitor of flowering [112, 113]. In addition, *AtKDM1B* and *AtKDM1C* can also affect the expression of *FWA*, a function independent of that of *AtKDM1A*.

In contrast, such duplication events were not observed for group II genes, suggesting a difference in the function between group I and group II *KDM1* genes. Since there is no reported study on *AtKDM1D*, it is unclear whether this gene also participates in the

regulation of flowering. The expression data from the GENEVESTIGATOR database and our previous microarray results [146, 147] showed that *AtKDM1D* is expressed at very low levels across all developmental stages. On the other hand, the sequence of *AtKDM1D* gene is well conserved. Hence it is possible that *AtKDM1D* has evolved a function in a specific group of cells or for a specific environmental situation.

2.4.4 The origin of SWIRM-AOD architecture

To investigate the origin of the *KDM1* genes, we performed additional phylogenetic analysis with eukaryotic *AOD* genes and the most similar *AOD* genes from Eubacteria. Our results (Fig. 2.3) showed that most major clades have one eubacterial *AOD* gene at or near the basal position. The *R. castenholzii AOD* gene is placed at the basal position outside all *KDM1* genes with 91/55 bootstrap support values. This topology suggests that all *KDM1* genes have a single origin from an *AOD* gene in the ancestor of Eukaryotes and Prokaryotes. However, it is still not clear whether the plant *AOD1* and *AOD2* genes have the same origin as *KDM1* since the position of *R. castenholzii AOD* gene was only weakly supported by the ML method.

As the KDM1 proteins also contain a SWIRM domain in addition to the AOD domain, how the SWIRM-AOD domain architecture originated is still a question. According to previously proposed mechanisms for the evolution of new gene structures

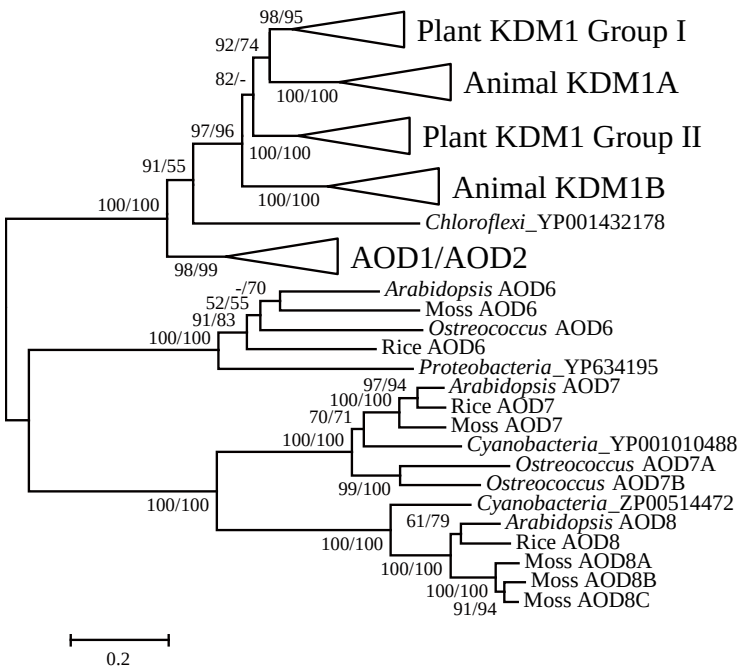


Figure 2.3 Phylogenetic tree showing the possible origin of selected eukaryotic *AOD* genes. The tree includes *KDM1*, *AOD6*, *AOD7* and *AOD8* genes and their most closely related eubacterial homologs. The tree was constructed using AOD domain region. The methods for tree construction and bootstrap values are given as in Fig. 2.1.

[148], there might be two possible origins for the first *KDM1* gene: (1) an exon shuffling/retrotransposition event that brought these two domain together; (2) *de novo* evolution of SWIRM domain coding region at the 5' of a preexisting *AOD* gene. Previous studies have shown that, in spite of its short length, the SWIRM domain is an evolutionarily conserved domain that occurred in proteins with different domain compositions [149]. Therefore the second possibility is unlikely.

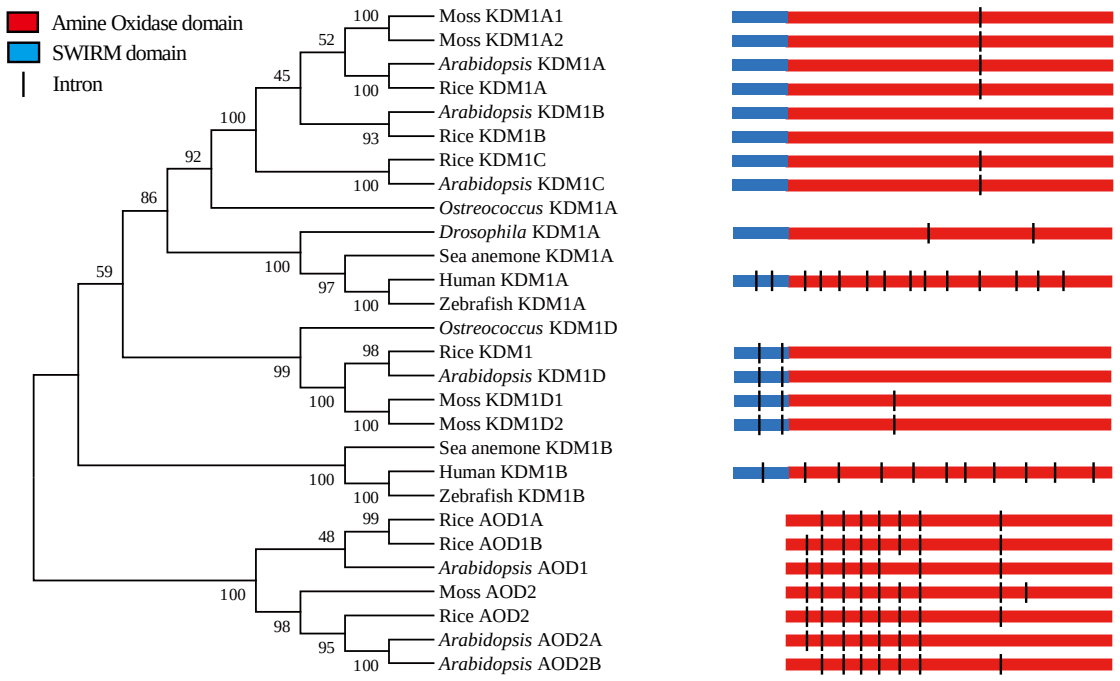


Figure 2.4 Schematic diagram of intron-exon structure of *KDM1*, *AOD1* and *AOD2* genes. Only the introns in SWIRM and AOD domain coding region are shown. The exons are drawn to scale.

To explore the first possibility, we analyzed the intron/exon structures of the *KDM1* genes and the closely related *AOD* genes (Fig. 2.4). Among the plant *AOD1* and *AOD2* genes, the number of introns ranged from 7 to 9 in the AOD domain-coding region. With the exception of only a few intron loss and gain events, the positions of all the introns are highly conserved. In contrast, the plant *KDM1* genes have many fewer introns. All plant *KDM1D* genes have two introns in the SWIRM domain-coding region and, except for the two *PpKDM1D* genes, have no intron in the AOD domain. Interestingly, the other plant *KDM1* genes have no intron in the SWIRM domain, but most of them possess an intron in the AOD domain at a different position from all the other introns mentioned above,

and the *AtKDM3B* and *OsKDM3B* are intronless for the entire gene. These intron/exon structures are also conserved in poplar and grape (*V. vinifera*) (not shown) [134, 150].

In comparison to the few introns observed in plant *KDM1* genes, the animal *KDM1* genes exhibit completely different patterns of intron positions. Most of the animal *KDM1* genes have one or two introns in the SWIRM domain and around 10 introns in the AOD domain, with the exception of insect *KDM1* genes, which have only 2 introns in the AOD domain only. Furthermore, although the positions of introns are conserved among animal *KDM1A* and *KDM1B* genes respectively, they are different from each other or that of the plant *KDM1* genes.

The most parsimonious explanation for the observed intron patterns is that the AOD domain of the ancestral *KDM1* gene in the most recent common ancestor of animals and plants was intronless, which supports the origin of *KDM1* gene through retrotransposition. After that, the plant *KDM1* genes have experienced limited or no intron gains, whereas the animal *KDM1* genes accumulated many introns during their evolution. It is still not clear what evolutionary pressure suppressed intron gain in plant *KDM1* genes. Nevertheless, these results again clearly support our conclusion that the animal and plant *KDM1* genes experienced very different evolutionary history.

2.4.5 Evidence for horizontal gene transfer (HGT) during the evolution of *AOD* genes

Another interesting result worth noting in Fig. 2.3 is that the most closely related eubacterial homologs of the plant *AOD7* and *AOD8* genes, respectively, are both from cyanobacteria. Previous studies revealed that both AtAOD7 and AtAOD8 proteins have chloroplast targeting signal and are localized to chloroplast [151]. It has been proposed that chloroplast originated from an eubacterium related to cyanobacteria through an endosymbiotic event, after which many genes have been transferred from the chloroplast to the nuclear genome [152]. Hence it is highly possible that the plant *AOD7* and *AOD8* genes are derived from the chloroplast. To examine this possibility, we performed further phylogenetic analysis that included the plant *AOD7* and *AOD8* genes and their eubacterial homologs. Besides plants, the only eukaryotic species in which we were able to find homologs of *AOD7* and *AOD8* was the brown alga *T. pseudonana,* which was proposed to have acquired a chloroplast through a secondary endosymbiotic event [153]. As shown in Fig. 2.5, the eukaryotic *AOD7* and *AOD8* genes cluster with their respective homologs from cyanobacteria with high bootstrap support (100/100). These results together suggest a cyanobacterium-like origin of the eukaryotic *AOD7* and *AOD8* genes.

AOD7 and AOD8 both are key enzymes important for the biosynthesis of

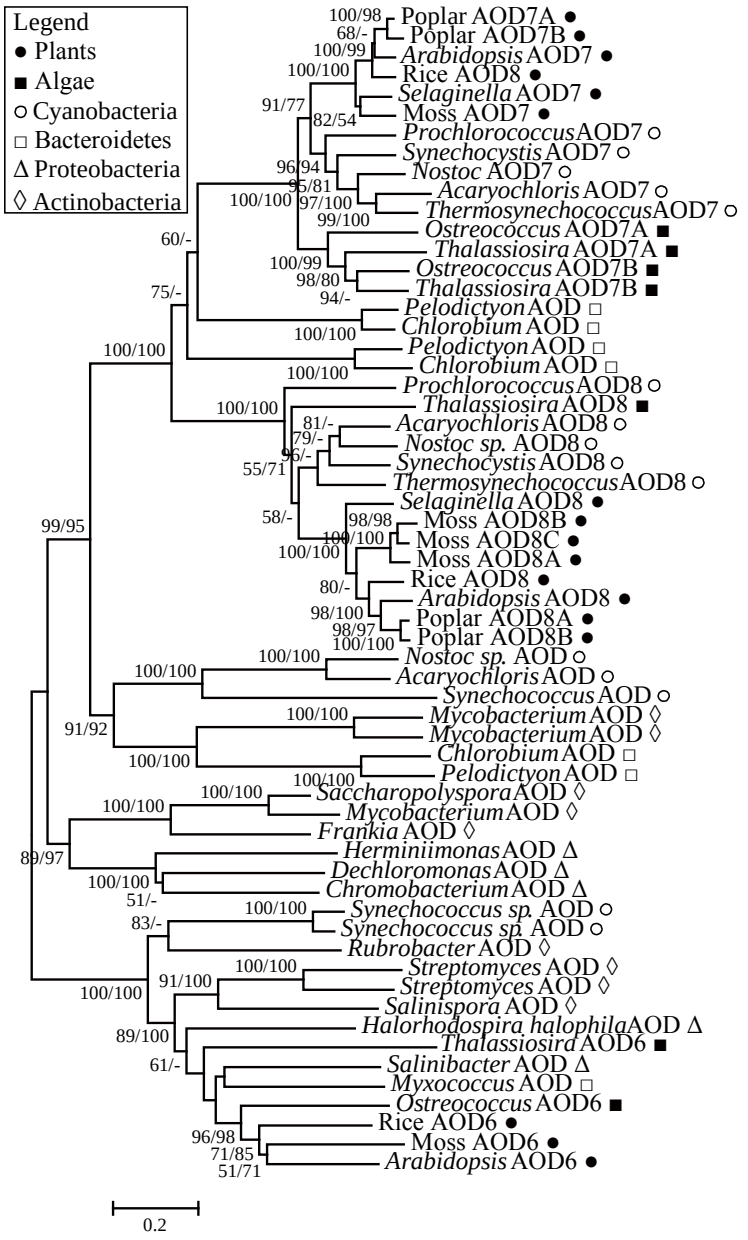


Figure 2.5 Phylogenetic tree of eukaryotic *AOD6*, *AOD7* and *AOD8* genes and their homologs from selected bacterial species, showing possible horizontal gene transfer between bacteria and plants. The tree was constructed using AOD domain region. The methods for tree construction and bootstrap values are given as in Fig. 2.1.

carotenoids in all photosynthetic organisms, including plants, algae and cyanobacteria [154]. In the carotenoid synthetic pathway, these two enzymes catalyze the two

dehydrogenation reactions that convert phytonen to *cis*-lycopene, which is then converted to all-*trans*-lycopene by an isomerase [151]. In many nonphotosynthetic organisms like fungi and eubacteria except cyanobacteria, these three steps are replaced by a single reaction that is catalyzed by a distinct enzyme [151]. It is possible that these two *AOD* genes were recruited to the carotenoid pathway in the common ancestor of cyanobacteria after its divergence from the other eubacteria, and then the photosynthetic eukaryotes acquired these two genes from cyanobacteria through HGT.

The plant-specific *AOD6* group, which is closely related to *AOD7* and *AOD8* groups, also shows a similar pattern. A homolog of plant *AOD6* genes can be found in the brown alga *T. pseudonana*, but not in animals and fungi. The AOD6 proteins are predicted to have a chloroplast-targeting signal, but the actual localization and function remains unknown. In addition, the region of the *AOD6* genes encoding the AOD domain is intronless. Our phylogenetic analysis (Fig. 2.5) showed that eukaryotic *AOD6* genes are most closely related to *AOD* genes from proteobacteria and bacteroidetes, suggesting that the eukaryotic *AOD6* genes also originated through HGT from a eubacterium. The results on *AOD6*, *AOD7*, and *AOD8* phylogeny together reveal an important role of HGT events in the evolution of *AOD* genes.

2.4.6 JmjC domain-containing proteins

Klose *et al.* have studied the evolutionary relationship between animal and fungal JmjC domain-containing proteins and they identified seven subfamilies based on both phylogenetic analysis of JmjC domain and domain architecture information [124]. JmjC proteins in six of the seven subfamilies have multiple domains and each family has a distinct domain structure. However, the evolutionary history of plant JmjC proteins is not clear. It has already been reported that two *Arabidopsis* JmjC proteins have an unusual domain architecture, which is not found in animals and fungi. Hence it will be of interest to elucidate the phylogeny of plant JmjC proteins, and compare between the evolutionary patterns of plant and animal JmjC proteins. To investigate the evolutionary history of plant JmjC domain histone demethylase genes, we retrieved sequences for JmjC domain-containing proteins from various plants and selected animals (Table 2.1). We also used the sequences of eukaryotic JmjC domains as queries to search for JmjC domain-containing proteins in prokaryotes. While proteins with limited similarity were found in Eubacteria, no homolog was detected in Archaea. Thus, our results indicate that neither AOD nor JmjC protein is present in Archaea. It is known that some archaea already possess the pseudonucleosomal tetrameric structures [155]. However, the absence of histone demethylase in Archaea is not surprising since archaeal histone proteins do not have N-terminal tails [155]. It is possible that, upon the acquisition of histone tails in

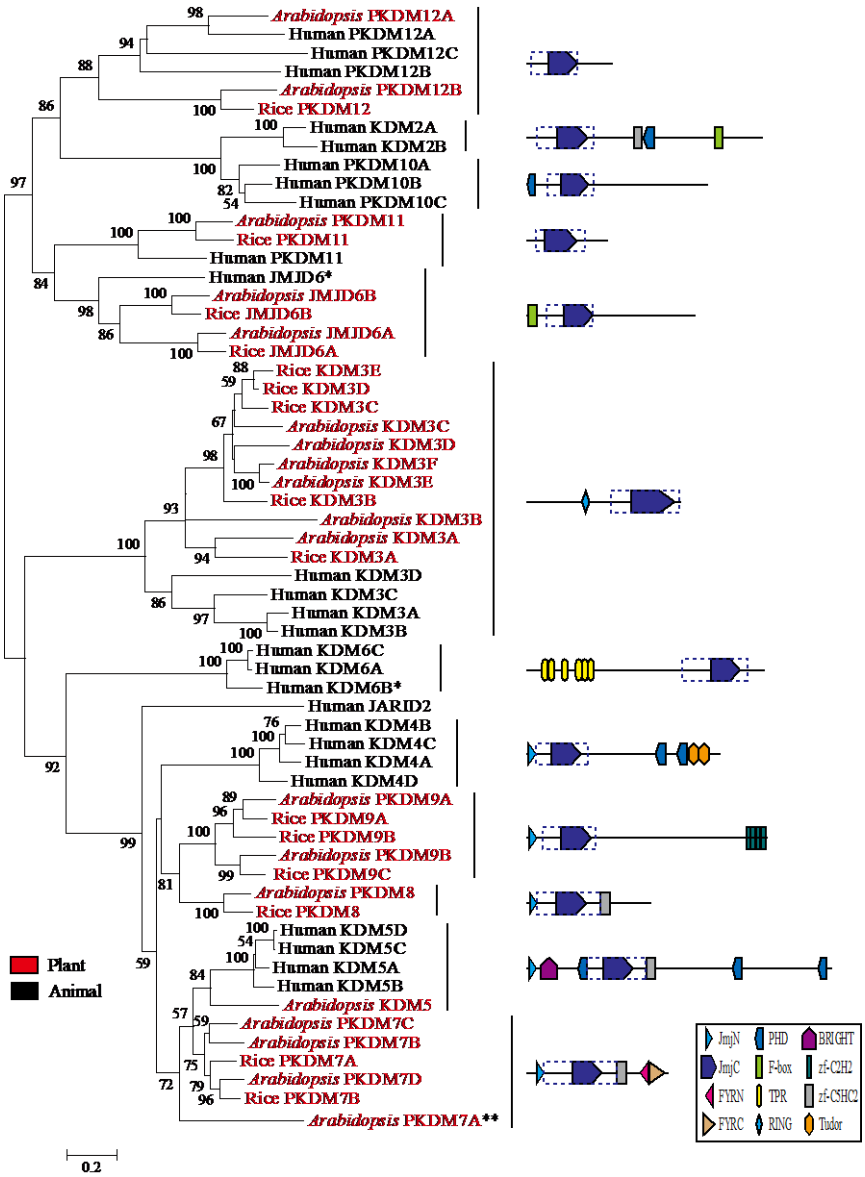


Figure 2.6 NJ tree showing the evolutionary relationship between JmjC genes from Human, Arabidopsis and Rice. The tree was constructed using JmjC domain region. Based on both phylogenetic information and domain architecture, 12 subfamilies can be defined. The representative domain architectures of each subfamily are shown next to the tree. JmjC proteins in the same subfamily have conserved regions extend from boundaries of JmjC domain, which are highlighted by dash line box.

*: Human JMJD6 and KDM6B do not have additional domain besides JmjC domain.

**: *Arabidopsis* PKDM7A has similar domain architecture to other PKDM7 proteins, but lacks the FYRN and FYRC domains. It is assigned to PKDM7 subfamily based on the phylogenetic tree shown in Fig. 2.7A.

early eukaryotes, AOD and JmjC proteins were recruited to serve as chromatin modifying enzymes.

The JmjC domains in most eubacterial JmjC proteins have low support from the SMART analysis and they are annotated as Cupin domain by Pfam with high e-value. Similarly, the JmjC domains in human MINA53, NO66, *Drosophila* CG2982 proteins retrieved in this study are also annotated as Cupin with strong support. The domain architecture analysis of the collected proteins shows that all the eubacterial proteins have only the JmjC domain. In contrast, most eukaryotic proteins contain other domain(s) besides the JmjC domain. Some domain architectures were observed only in plant members and others only in animals and/or fungi members. From the amino acid sequence alignment, the proteins with the same domain architecture have more similar JmjC domains and regions flanking the JmjC domains.

2.4.7 The birth-and-death evolution of genes encoding JmjC domain proteins

According to the NJ tree shown in Fig. 2.6, the *JmjC* family can be divided into 12 monophyletic subfamilies. These 12 subfamilies represent 11 different domain architectures, as two subfamilies contain only the JmjC domain. Previously, these two subfamilies were in a monophyletic group with low support and was defined as a single subfamily [124], but our study supports two separate subfamilies. On the other hand, the

other six subfamilies defined in the previous study [124] were confirmed by our result. Most of the 12 subfamilies are designated after the name of their animal members according to their chromatin modifying enzyme activities [156]. Among these subfamilies, the *KDM2*, *KDM4* and *KDM6* subfamilies are animal specific, while *KDM3*, *KDM5* and *JMJD6* have members from both plants and animals. Those subfamilies without a known histone demethylase function are named as PKDM (**P**utative-**KDM**). Among these, the *PKDM7*, *PKDM8* and *PKDM9* subfamilies are composed of only plant genes and *PKDM10* is animal specific; the remaining two subfamilies, *PKDM11* and *PKDM12*, contain both plant and animal genes. According to the tree topology, it could be estimated that there were at least nine JmjC genes in the most recent common ancestor of plants and animals. After the divergence of animals and plants, some copies were lost in plants, others in animals. When the human *MINA53*, *NO66*, *Drosophila CG2982* genes and their eubacterial homologs encoding Cupin proteins were included in the analysis, they form a separate clade that is sister to the *PKDM12* subfamily. Therefore it is possible that all *JmjC* genes originated from the ancestor of these *Cupin* genes.

Besides the above mentioned loss of specific subfamilies in plant or animal *JmjC* genes, three different patterns of birth-and-death evolution were also observed within the subfamilies. In the *PKDM9* and *PKDM11* subfamilies, a single copy has been stably maintained in both animals and plants, except for a recent duplication of *PKDM9* in poplar. In other subfamilies, *JmjC* genes experienced duplication in one of the animal and

plant lineages, but were stable or lost in the other lineage. For example, in the *JMJD6* subfamily, while one copy has been maintained in each animal, two gene duplication events can be detected in plants, one before the divergence of land plants and another in moss. In contrast, the *KDM5* subfamily has four members in humans, resulting from duplication events after the divergence of vertebrate animals from insects. This pattern is also found in the plant (e.g. *PKDM3*) or animal (e.g. *KDM6*) specific subfamilies. A third pattern is that *JmjC* genes were duplicated in both animals and plants, such as the *KDM3* subfamily. As shown in Fig. 2.7A, five well supported clades all include members from *Arabidopsis* and poplar, suggesting the presence of five *KDM3* genes in the most recent ancestor of *Arabidopsis* and poplar. These five clades were all derived from possibly one copy in the ancestor of plants and animals through gene duplication. In addition, lineage specific gene duplication events can be observed in plants. In animals, duplication events can also be inferred from the clade with 100/100 supports that is composed of one *Drosophila KDM3* and four human *KDM3* genes.

2.4.8 Potential histone demethylase activities of plant JmjC proteins

Among the twelve subfamilies identified in this study, six of them have at least one member with known histone demethylase activities [108]. However, as all functional studies so far are performed in animals and fungi, no plant JmjC protein in these six

subfamilies has been reported to have histone demethylase activity. In the absence of biochemical studies, our phylogenetic results can be valuable clues about possible functions of plant JmjC proteins. Here, we propose potential histone demethylase activities for the plant JmjC proteins based on the evolutionary relationships from this study, the conservation of enzymatic active sites and domain architectures.

As described above, three subfamilies have both members with known biochemical activities and members from plants. Two human proteins in *KDM3* subfamily, human KDM3A and KDM3B, have been shown to have H3K9me1/2 demethylase activity [157]. The plant KDM3 proteins have the same domain architecture as the animal members, with a zinc finger domain in addition to the JmjC domain. The predicted cofactor binding sites are also conserved in most plant KDM3 proteins, suggesting possible H3K9 demethylase function. Consistent with this idea, a recent study revealed an increased level of H3K9 methylation at the *BNS* locus in the *Arabidopsis kdm3c* mutant [123]. However, proteins in the two clades including *Arabidopsis* KDM3A and KDM3B have variant residues at the co-factor binding sites. Hence it is possible that they have evolved novel functions or become pseudogenes. To investigate these two possibilities, we examined the expression data of *Arabidopsis KDM3* genes from our previous microarray analysis [147] and the GENEVESTIGATOR database [146]. *AtKDM3A* has the highest expression level among these genes at all the developmental stages, and *AtKDM3B* is also expressed, suggesting they are functional.

Similar phenomenon can be observed in the *PKDM7* subfamily. All proteins in this subfamily are from plant and they contain a JmjN domain, a C5H2-zinc finger domain and C-terminal FYRN and FYRC domains. The cofactor binding sites are conserved in all members but the *Arabidopsis* and poplar PKDM7A proteins, which have evolved much faster than the other members. Nevertheless, the expression data shows that *AtPKDM7A* is expressed at a level comparable to *AtPKDM7B* and *AtPKDM7D* [146, 147]. Although the AtPKDM7C protein has intact cofactor binding sites, it has no detectable expression. The phylogeny in Fig. 2.7B shows that the *PKDM7* subfamily forms a clade with 99/97 bootstrap supports and is most closely related to the *KDM5* subfamily. Several animal KDM5 proteins have been shown to have H3K4me2/3 demethylase activities [119, 158-160]. Therefore, although the plant KDM5 and PKDM7 proteins have distinct domain architecture, they might have H3K4 demethylase activities.

Recently, the human JMJD6 protein was shown to have histone arginine demethylase activity [109]. Although most of the cofactor binding sites are conserved in plant JMJD6 proteins, the first KG binding site has been substituted by Ser and Ala in plant JMJD6A and JMJD6B proteins, respectively. It is unclear whether these substitutions will compromise the histone arginine demethylase activity of plant JMJD6 proteins. We noticed that *AtJMJD6A* and *AtJMJD6B* are expressed at a high level at specific developmental stages [146, 147] , suggesting that these proteins are functional.

In summary, we have used our phylogenetic results to propose histone demethylase

activities for plant JmjC proteins in four subfamilies. The other plant JmjC proteins are either in the plant specific subfamilies *PKDM8* and *PKDM9* or in the subfamilies *PKDM11* and *PKDM12* which do not have an animal member with known biochemical activities. Nevertheless, some of these plant JmjC proteins have already been implicated in chromatin modification. For example, an elevated histone H4 acetylation level is observed at the *FLC* locus in the *Arabidopsis pkdm9a* mutant, which is phenotypically similar to the *kdm1a* mutant [122]. Moreover, it is still not clear which proteins are responsible for the H3K9me3, H3K27 and H3K36 demethylation in plants, since the *KDM2*, *KDM4* and *KDM6* subfamilies do not have a plant member. One possibility is that these demethylase activities in plants are carried out by some of the other JmjC proteins without a known function.

2.4.9 Functional implications of differences in evolutionary patterns

Our phylogenetic analyses of these two histone demethylase families revealed a significant difference in evolutionary pattern between animal and plant proteins in both families. In the *AOD* family, the plant group I *KDM1* genes were duplicated several times before the diversification of flowering plants and further in specific lineages, whereas the animal *KDM1* genes have been maintained with a constant copy number in most species. The animal and plant JmjC domain-containing proteins show similar patterns of

evolution in some subfamilies but not in the others. Furthermore, certain types of histone demethylation might be conducted by plant JmjC proteins in subfamilies different from the animal JmjC proteins. These results indicate a divergence in the regulation of histone methylation between animals and plants, consistent with the proposed divergent roles of histone methylation in different organisms [161]. In animals, both H3K9me2 and H3K9me3 are enriched in heterochromatin. However, in *Arabidopsis*, while the H3K9me2 is considered as a hallmark of heterochromatin, H3K9me3 is mainly found in euchromatin [161]. In animals, H3K9me3 demethylation is catalyzed by members of the *KDM4* subfamily [162-165], which lacks plant members, suggesting that H3K9me3 demethylation in plants is catalyzed by proteins from another subfamily. Furthermore, previous phylogenetic analysis also revealed a similar evolutionary pattern in the HDAC families; one of the three major classes of SIR2 family of HDACs has members from animal but not plant, whereas the HD2 family is plant specific [166]. Thus, the functional and regulatory diversification might be a common feature of chromatin modification genes.

In addition, our study also showed distinct evolutionary patterns between the *AOD* and the *JmjC* families. Whereas the *KDM1* genes only experienced limited duplication events and maintained relatively constant domain architecture in their history, the *JmjC* gene have evolved several types of domain architectures before the divergence of major eukaryotic groups and underwent further duplication subsequently. As suggested by

genome-wide studies in *Drosophila* and fungi, such divergence in evolutionary patterns may indicate differences in functional essentiality [140, 141]. The KDM1 histone demethylases are reported to have a variety of functions. In animal, KDM1 is required for the ligand-dependent transcriptional activation by nuclear hormone receptors [105, 167]. It also plays important roles in cell differentiation, cell cycle control and spermatogenesis [142]. In addition, most of these functions are also shared by JmjC proteins. For example, members of the *KDM3* and *KDM4* subfamilies, which possess H3K9 demethylase activities, are also required for the steroid hormone induced gene expression [121, 157, 167] and KDM3A is crucial for spermatogenesis [168]. In addition, the KDM5A, an H3K4me2/me3 demethylase, has overlapping roles with KDM1 in the regulation of cell differentiation [158]. It is also possible that KDM1 has some distinct function from those of JmjC proteins. In fact, the work by Lan *et al.* showed that KDM1 is retained on the unmethylated H3K4 after its action and suggested a role of KDM1 in the prevention of H3K4 methylation [169].

Another explanation for the observed evolutionary patterns is that they reflect the difference in evolutionary potential of these two families of histone demethylase. Consistent with this idea, several lines of evidence support a greater functional potential of JmjC proteins than KDM1. First, the JmjC proteins have broader substrate specificity than KDM1 proteins. The requirement of a protonated nitrogen in KDM1-mediated demethylation limits the substrate specificity of KDM1 to mono- and dimethylated lysine

residues. By contrast, JmjC proteins are able to demethylate all the three states of lysine methylation. In addition, KDM1 proteins are only known to catalyze the demethylation on H3K4 and H3K9, whereas the substrates for JmjC proteins include H3K4, H3K9, H3K27, H3K36 and even H3R2. Studies on protein structures suggest that the interactions between KDM1 and the substrate are intricate and specific, leading to the exquisite substrate specificity of KDM1. Second, the JmjC domain is much smaller than the AOD domain in KDM1. The JmjC domain in most JmjC proteins are less than 200 amino acids, but the length of the AOD domain is usually more than 400 amino acids. Smaller domain might be combined with the other domains more easily, providing JmjC proteins greater evolutionary adaptability. This is supported by a recent study, which identified the protein domains with relatively high tendency to combine with different domains in eukaryotes [170]. In their list of highly versatile domains, most have 250 or fewer amino acids residues. Hence the short length of JmjC domain may allow JmjC proteins to evolve new functions quickly by combining with new domains, which can promote protein-protein interaction, DNA binding or recognition of chromatin modification.

2.4.10 Apparently convergent evolution of histone demethylases

The fact that the *KDM1* and *JmjC* genes belong to two phylogenetically distinct gene

families indicates that, during evolution, these two gene families were recruited to perform the histone demethylation activity independently, providing an example of convergent evolution. In fact, this phenomenon is prevalent among histone modifying enzymes. For instance, the enzymes that catalyze histone methylation belong to two different families, the widespread SET-domain family and the DOT1-related protein family [97]. Similarly, there are three distinct families of HDACs and four different families of HATs [166]. In addition, it is also common that families responsible for the same type of histone modification show distinct evolutionary patterns. While some families are widespread in eukaryotes (e.g. SET family HKMTs and SIR2 family HDACs), others are only present in specific lineages of eukaryotes (e.g. DOT1-related HKMTs in animals and fungi and HD2 family HDACs in plants) [97, 166]. The recruitment of more than one gene families to fulfill the same type of biochemical activities might have allowed these families to evolve specific roles under different circumstances (e.g. cell type, developmental stage, environmental cues) or toward different substrates. The multiple origins of histone modification enzymes have likely contributed to the complexity of epigenetic regulation.

2.5 CONCLUSIONS

In this paper, we present detailed phylogenetic analyses of the *KDM1* and *JmjC* families, whose members include the recently identified histone demethylases. Our results revealed a possible single origin of all *KDM1* histone demethylase genes through the acquisition of the region encoding the SWIRM domain by an *AOD* gene before the split of major eukaryotic lineages. The *KDM1* genes are conserved in both copy number and domain structure during evolution, although a few duplication events were observed in plants. We also identified the contribution of HGT events to the evolution of *AOD* genes. On the other hand, our analyses *JmjC* genes showed this family clearly experienced birth-and-death evolution and the subfamilies displayed lineage-specific duplication patterns. According to the evolutionary relationship revealed by our study, we proposed histone demethylase activities for several plant JmjC domain-containing proteins. Furthermore, we found distinct evolutionary patterns of histone demethylases in different lineages and between the *KDM1* and *JmjC* families. These results may imply functional divergence of certain types of histone methylation in different organisms and different classes of function associated with KDM1 and JmjC domain-containing histone demethylases. In summary, our study improves the understanding about the evolution and functions of histone demethylases and provides valuable information for future studies.

CHAPTER 3

PHYLOGENETIC DETECTION OF NUMEROUS GENE DUPLICATIONS SHARED BY ANIMALS, FUNGI AND PLANTS

The work described in this chapter has been published in Zhou *et. al, Geome Biol.,* 11:R38.

Note: In this publication, Dr. Zhenguo Lin has contributed to the analysis of protist sequences.

3.1 SYNOPSIS

Gene duplication is considered a driving force for evolution of genetic novelty by facilitating functional divergence and organismal diversity, including the process of speciation. Animals, fungi and plants are major eukaryotic kingdoms and the divergences between them are some of the most significant evolutionary events. Although gene duplications in each lineage have been studied extensively in various contexts, the extent of gene duplication prior to the split of plants and animals/fungi is not clear. Here, we have studied gene duplications in early eukaryotes by phylogenetic relative dating. We have reconstructed gene families (with one or more orthogroups) with members from both animals/fungi and plants by using two different clustering strategies. Extensive phylogenetic analyses of the gene families show that, among nearly 2600 orthogroups identified, at least 300 of them still retain duplication that occurred before the divergence of the three kingdoms. We further found evidence that such duplications were also detected in some highly divergent protists, suggesting that these duplication events occurred in the ancestors of most major extent eukaryotic groups. Our phylogenetic analyses show that numerous gene duplications happened at the early stage of eukaryotic evolution, probably before the separation of known major eukaryotic lineages. We discuss the implication of our results in the contexts of different models of eukaryotic

phylogeny. One possible model for the large number of gene duplication events is one or more large-scale duplications, possibly whole genome or segmental duplication(s), which might have provided a genomic basis for the successful radiation of early eukaryotes.

3.2 INTRODUCTION

The history of eukaryotic evolution is one with ever-increasing diversity and complexity at multiple levels. The increases in genotypic and phenotypic complexity are usually associated with expansion of gene families. These multigene families are subject to birth-and-death evolution and most new genes arise by gene duplication [13]. During the eukaryotic history, gene duplication has been a ubiquitous phenomenon and has contributed to evolutionary innovation by generating additional genetic materials for functional divergence and novelty [1]. Besides its important role in the evolution of new gene functions, gene duplication also has great contribution to the speciation process through the divergent resolution of duplicated genes in different populations [8]. In particular, large-scale gene duplication events have been documented in animals and fungi, and are particularly frequent in plants [22, 23, 30-34]. These large-scale duplications are believed to be associated with dramatic increases in species diversity, such as the radiation of vertebrates and the diversification of flowering plants [36, 37].

One of the most important evolutionary milestones is the early diversification of eukaryotes [78]. The traditional “crown-stem” model (Fig. 1.3A) of eukaryotic phylogeny based on small-subunit ribosomal RNA sequences suggests that plants, animals and fungi form a crown group in the eukaryotic tree and separated from each

other more recently than some early branching protists [68, 72, 81]. In contrast, the "six supergroups" model stems from recent phylogenomic studies and classifies the eukaryotic diversity into six supergroups (Fig. 1.3B) [69]. The diverging order of these supergroups is difficult to resolve, suggestive of a rapid diversification at the early stage of eukaryotic evolution [74-77]. Nonetheless, a number of studies suggest that the split between Archaeplastida (including plants) and Opisthokonta (including animals/fungi) is among the earliest known eukaryotic divergences, before the divergence of other major protist groups from either Archaeplastida or Opisthokonta [74-76]. Therefore, the separation of plants from animals/fungi would be much more ancient than what was suggested by the "crown-stem" model.

Previous phylogenetic studies of individual eukaryotic gene families for transcription regulators, kinesins, and recombinational proteins all indicate that there were duplication events before the split of animals and plants, suggestive of abundant gene duplication during the early eukaryotic evolution [50-52, 171-173]. This notion is also supported by a comparative genomic study, in which the established COG (prokaryotic clusters of orthologous groups) and KOG (eukaryotic clusters of orthologous groups) databases were used to reconstruct gene clusters and to analyze their phylogenies [174]. It was found that the inferred number of genes in the last eukaryotic common ancestor is 1.92 fold higher than in the first eukaryotic common ancestor, leading to the conclusion that early eukaryotes had significantly more gene duplication than prokaryotes during similar

periods [174]. However, a systematic investigation of the extent of gene duplication prior to the split of plants and animals/fungi is still lacking. Here, we present extensive phylogenetic analyses of gene families and our results supporting the hypothesis that many of these families had experienced at least one duplication event before the divergence of the three major eukaryotic kingdoms.

3.3 MATERIALS AND METHODS

3.3.1 Reconstruction of gene clusters

For Analysis I and II, the predicted protein sequences of the 14 representative species were retrieved from public databases (see Appendix Table 3.1 for the complete list of data sources). These protein sequences were compared using a all-to-all BLASTP search with a cut-off of $1e^{-10}$ [175]. Based on the BLASTP results, MCL clustering was performed with low stringency (Inflation value of 1.5) to produce gene clusters [176]. To check the clusters for common domains, the domain architectures of all cluster members were annotated using InterProScan v4.5 (InterPro release 22.0, including both integrated and un-integrated) [177].

For Analysis III, we started from the 1092 KOG-to-COG clusters identified in the study of Makarova *et al.* [174]. Since the original KOG database does not cover the genomes of *Physcomitrella*, *Chlamydomonas*, *Takifugu* and *Strongylocentrotus*, the predicted protein sequences from these four species were assigned to KOGs using BLASTP search. Then the sequences from the 14 representative prokaryotic and eukaryotic species were extracted from each KOG-to-COG cluster to form the dataset for the following phylogenetic analysis.

3.3.2 Phylogenetic analysis

For all the MCL gene clusters and KOG-to-COG clusters, highly similar sequences (more than 80% identity) from the same species were removed by using BLASTCLUST [175]. Multiple sequence alignments (MSAs) were generated by using MUSCLE 3.6 [127]. The MSAs were trimmed by removing poorly aligned regions using trimAl 1.2 with the automated1 option [178]. Neighbor-Joining (NJ) trees were constructed using PHYLIP 3.68 (JTT model) with 1000 bootstrap replicates [179, 180]. Maximum-likelihood (ML) trees were constructed using RAxML 7.2.0 (LG model plus Gamma correction) with 100 bootstrap replicates [181, 182]. The best-scoring ML trees were also evaluated with aLRT method by using Phyml 3.0 [183, 184]. For large clusters with more than 100 sequences, representative sequences were selected based on preliminary NJ tree. Phylogenetic trees were screened by custom scripts to identify orthogroups and duplication events. All scripts in this study, gene clusters and phylogenetic trees are available upon request.

3.3.3 Gene ontology analysis

Orthogroups with early eukaryotic duplication were compared with orthogroups that did not have such duplications for overrepresented GO terms [185]. Domains encoded by

the majority of orthogroup members were considered representatives for the orthogroup. Then GO annotations of representative InterPro domains were assigned to each orthogroup using InterPro2GO mapping [186]. Subsequently, all GO annotations were mapped to GO slims, a cut-down version of GO, using map2slim perl script and generic GO slim verstion 1.2 [186]. The overrepresentation of GO slims was examined using Ontologizer 2.0 [187] with Term-for-Term analysis and Bonferroni correction for multiple testing.

3.4 RESULTS

3.4.1 Reconstruction of gene clusters with MCL method

To identify gene duplication in early eukaryotic evolution, we reconstructed gene families from representative eukaryotic and prokaryotic species. The three multicellular eukaryotic kingdoms, plants, animals and fungi, belong to two of the six major eukaryotic supergroups (plants in Archaeplastida; animals and fungi both in Opisthokonta) [69]. According to the "six supergroups" model of eukaryotic phylogeny (Fig. 1.3B) and other recent phylogenies, the separation of plants and animals/fungi could be as early as the separation of any major groups of extent eukaryotes. Hence, the gene duplication prior to the split of plants and animals/fungi can be placed at an early stage of eukaryotic evolution.

In this study, we included three representatives of Archeaplastida (the flowering plant *Arabidopsis thaliana,* the moss *Physcomitrella patens* and the green alga *Chlamydomonas reinhardtii*), 3 animals (*Homo sapiens*, the pufferfish *Takifugu rubripes* and the sea urchin *Strongylocentrotus purpuratus*) and 2 fungi (the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*), which all have nearly complete genome sequences (Appendix Table 3.1). According to a widely

accepted model for the eukaryotic origin, the ancestral eukaryotic cell was derived from an Archaea-like organism, with additional genes originated from the endosymbiosis of a Proteobacterium-like cell, which evolved into the mitochondrion [152]. Therefore, we included genes from three bacteria (*Escherichia coli*, *Rickettsia prowazekii* and *Bacillus subtilis*) and three archaea (*Methanosarcina acetivorans*, *Sulfolobus solfataricus* and *Pyrobaculum aerophilum*) as outgroups (Appendix Table 3.1).

The predicted protein sequences from all these 14 species were clustered using the Markov Clustering Algorithm (MCL) (see Methods) which is among the most popular clustering methods and has been shown to be reliable [176]. By using a relatively low clustering stringency, 222436 annotated protein sequences from the 14 representative species were divided into 51396 gene clusters in total. Among these, 1394 clusters contained both prokaryotic and eukaryotic genes and 41444 clusters were eukaryote-specific. In addition, 794 out of the 1394 clusters and 2276 out of the 41444 clusters contained genes from both Archeaplastida and Opisthokonta. The numbers of clusters of other phyletic patterns are summarized in Appendix Table 3.2.

3.4.2 Analysis I – MCL clusters with both prokaryotic and eukaryotic genes

On the basis of the 794 clusters with genes from Archaeplastida, Opisthokonta, and prokaryotes, we retained only the clusters that had at least three eukaryotic genes, with at

least one from Archaeplastida and at least one from Opisthokonta, as this is the minimum requirement for the deduction of a possible early eukaryotic duplication prior to the divergence of these two lineages. Also, to ensure the quality of these clusters, we tested the clusters by searching for one or more common domains in all members and subsequently removed sequences, if any, that lacked the most common domain(s) from each cluster. As a result, we obtained 772 gene clusters that meet these criteria and used them for phylogenetic analyses. The phylogeny for each cluster was estimated by the Neighbor-Joining (NJ) method with bootstrap test (BS) and Maximum-Likelihood (ML) method with BS and approximate likelihood ratio test (aLRT) (see Methods). The resulting tree topologies were then examined. Most gene families known to have experienced duplication in early eukaryotes were successfully recovered by our analysis (Appendix Table 3.3). Since our clusters were established based on sequence similarity instead of strict orthology, the eukaryotic genes in one cluster might be derived from more than one prokaryotic ancestor. To best distinguishing the duplication in early eukaryotes from paralogy before the prokaryotes-eukaryotes separation, we identified orthogroups in each tree; each orthogroup consisted of eukaryotic genes that, most likely, originated from the same gene in the first eukaryotic common ancestor. According to the tree topology (see Fig. 3.1A for illustration), we defined an orthogroup as a eukaryotic clade that meet both of the following criteria: (1) having members from both plants and animals/fungi, and (2) having a prokaryotic outgroup (designated as Type I orthogroups;

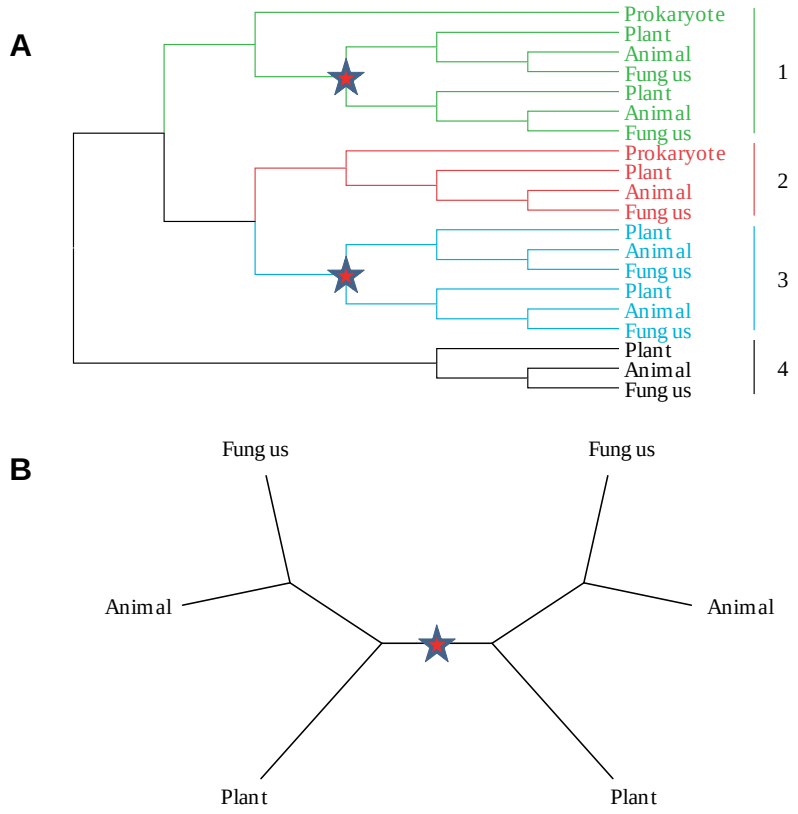


Figure 3.1 The design of phylogentic analysis. (A) Hypothetical phylogenetic tree showing the definition of orthogroups in Analysis I and III. Four possible orthogroup topologies are highlighted by colors: 1 (green), eukaryotic genes with prokaryotic outgroup and early eukaryotic duplication; 2 (red), eukaryotic genes with prokaryotic outgroup but no early eukaryotic duplication; 3 (blue), eukaryotic genes without prokaryotic outgroup but show early eukaryotic duplication; 4 (dark), eukaryotic genes without prokaryotic outgroup nor early eukaryotic duplication. (B) Hypothetical phylogenetic tree showing an example of eukaryote specific gene cluster with duplication. The stars indicate gene duplications**.**

e.g. the clades 1 and 2 in Fig. 3.1A) or being a sister to another orthogroup which has prokaryotic outgroup (designated as Type II orthogroups; e.g. the clades 3 and 4 in Fig. 3.1B). According to these criteria, we identified about 700 orthogroups. In each orthogroup, an ancient duplication event was inferred to be prior to the divergence of plants and animals/fungi if the tree topology of the orthogroup had two or more

eukaryotic clades with at least one clade consisted of members from both plants and animals/fungi. According to this definition, more than 35% [Bootstrap (BS) >= 50%] or 20% (BS support >= 70%) of the 700 orthogroups showed one or more ancient duplication events (Table 3.1). Furthermore, the aLRT test of ML phylogenies produced even higher percentages of orthogroups with an early eukaryotic gene duplication at both support levels of 50% and 70% (Table 3.1).

We reasoned that some of the gene duplications identified might be caused by long-branch attraction (LBA) artifacts in phylogenetic reconstruction. For example, in an orthogroup with the phyletic pattern of (plants, animals, fission yeast)(budding yeast), it was possible that fission yeast gene evolved rapidly and was placed at the basal position due to LBA. In this case, a duplication event would be inferred based on the incorrect topology. Therefore, to minimize the impact of LBA, we used a more stringent criterion for the identification of gene duplication before the divergence of plants and animals/fungi: at least one gene from at least one species must be present in each of two paralogous clades. Based on this conservative criterion, we still found about 25% (BS >= 50%) or 15% (BS >= 70%) of the orthogroups to have experienced an early eukaryotic duplication (Table 3.1, bold face). Also, the ML-aLRT test showed that more than 30% orthogroups (at both support levels of 50% and 70%) have experienced an early eukaryotic duplication (Table 3.1, bold face). The stringent criterion was also used in Analyses II and III (see below). Moreover, we arbitrarily selected a subset of the

Table 3.1 Number of orthogroups and early eukaryotic duplications identified in Analysis I

	NJ-BS*				ML-BS				ML-aLRT**			
	>= 50%		>= 70%		>= 50%		>= 70%		>= 50%		>= 70%	
Type I orthogroup with duplication	205	**136**	119	**88**	199	**135**	104	**82**	282	**188**	234	**166**
Type I orthogroup total	522		435		511		445		599		560	
Type II orthogroup with duplication	100	**63**	61	**43**	72	**46**	37	**29**	81	**60**	85	**66**
Type II orthogroup total	235***		260		229***		234		176***		196	
Total orthogroup with duplication	305	**199**	180	**131**	271	**181**	141	**111**	363	**248**	319	**232**
Orthogroup total	757		695		740		679		775		756	
Percentage	40.3%	**26.3%**	25.9%	**18.8%**	36.6%	**24.5%**	20.8%	**16.3%**	46.8%	**32.0%**	42.2%	**30.7%**

Type I orthogroup refers to orthogroups with a prokaryotic outgroup; Type II orthogroup refers to orthogroups without a prokaryotic outgroup. Bold face indicates that the duplications were inferred based on stringent criteria which required that at least one species was present in both paralogous clades.

*: BS = Bootstrap test
**: aLRT = Approximate Likelihood-Ratio test
***: These numbers of Type II orthogroups at support level of >= 70% are greater than that at support level of >= 50%, since some Type II orthogroups with >= 70% support were from Type I orthogroups with >= 50% support whose prokaryotic outgroup had support less than 70%.

orthogroups with topologies that were vulnerable to LBA, and added sequences from additional species to further test the impact of LBA. The results showed that phylogenies of the majority of orthogroups tested (15 out of 21) still supported early eukaryotic duplication (Appendix Table 3.4). Especially, all 6 orthogroups that initially showed

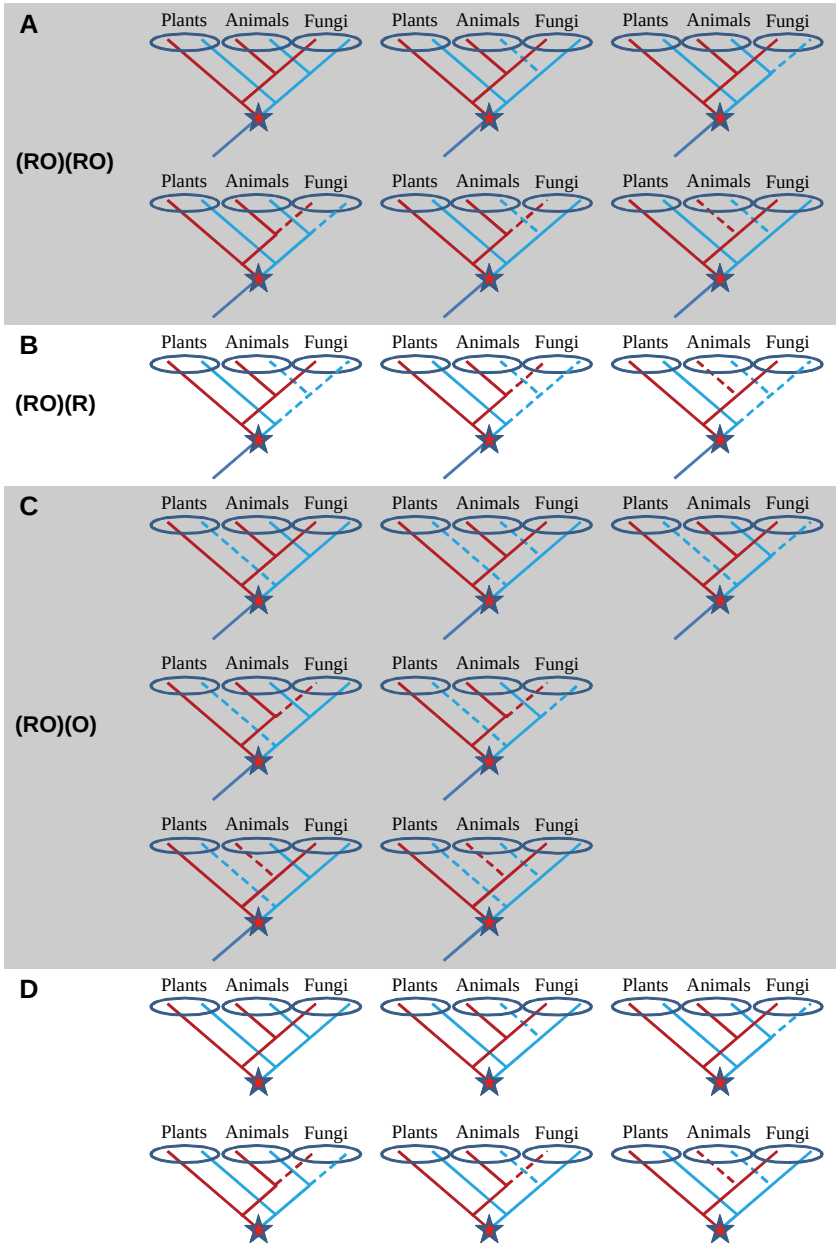


Figure 3.2 Hypothetical examples of phylogenetic tree showing the patterns of duplicates retention. (A) 6 phyletic patterns showing the (RO)(RO) pattern. (B) 3 phyletic patterns showing the (RO)(R) pattern. (C) 7 phyletic patterns showing the (RO)(O) pattern. (D) 6 phyletic patterns which supported an early eukaryotic duplication in eukaryote specific gene clusters.

duplication at support level of 70% still supported early eukaryotic duplication after

adding additional sequences. These results suggested that our phylogenetic topologies were quite reliable.

To learn about the fate of the ancient duplicates, we also examined whether specific duplicates were retained or lost, and found that different orthogroups varied in the patterns of retention of duplicates. One possible fate was that both of the duplicates were retained in plants and animals/fungi (Fig. 3.2A), abbreviated here as (RO)(RO) (R – Archaeplastida, O – Opisthokonta). Among all the orthogroups that showed early eukaryotic duplication, about 35% displayed this pattern (Table 3.2). Alternatively, one of the duplicates could be lost in either plants or animals/fungi, abbreviated here as (RO)(R) and (RO)(O), respectively (Fig. 3.2A and C). These two topologies were less frequent than (RO)(RO) (Table 3.2). Similar results were obtained with different phylogenetic methods and at different levels of support. A small number of remaining orthogroups had more complex patterns (Table 3.2, Other), possibly due to multiple rounds of duplication and gene loss. The detailed distribution of phyletic patterns is summarized in Appendix Table 3.5.

In the context of "six supergroups" model of eukaryotic evolution (Fig. 1.3B), the gene duplications we identified were very ancient events as they happened before the separation of Archaeplastida and Opisthokonta. This split possibly represents the most ancient eukaryotic divergence among extent groups. However, the "crown-stem" model (Fig. 1.3A) suggests that the plants-animals/fungi split is relatively recent in comparison

Table 3.2 Distribution of orthogroups with phyletic patterns supporting early eukaryotic duplication

Dataset	Method	Support	(RO)(RO)	(RO)(R)	(RO)(O)	Other***	Total
Analysis I	NJ-BS*	>= 50%	73 (36.7%)	56 (28.1%)	59 (29.6%)	11 (5.5%)	199
		>= 70%	52 (39.7%)	31 (23.7%)	34 (26.0%)	14 (10.7%)	131
	ML-BS	>= 50%	71 (39.2%)	55 (30.4%)	46 (25.4%)	9 (5.0%)	181
		>= 70%	46 (41.4%)	29 (26.1%)	21 (18.9%)	15 (13.5%)	111
	ML-aLRT**	>= 50%	102 (41.1%)	75 (30.2%)	64 (25.8%)	7 (2.8%)	248
		>= 70%	95 (40.9%)	63 (27.2%)	62 (26.7%)	12 (5.2%)	232
Analysis III	NJ-BS	>= 50%	90 (30.9%)	72 (24.7%)	94 (32.3%)	35 (12.0%)	291
		>= 70%	40 (26.3%)	41 (27.0%)	41 (27.0%)	30 (19.7%)	152
	ML-BS	>= 50%	92 (33.9%)	80 (29.5%)	62 (22.9%)	37 (13.7%)	271
		>= 70%	39 (30.2%)	33 (25.6%)	22 (17.1%)	35 (27.1%)	129
	ML-aLRT	>= 50%	299 (48.3%)	156 (25.2%)	156 (25.2%)	8 (1.3%)	619
		>= 70%	268 (46.4%)	136 (23.6%)	150 (26.0%)	23 (4.0%)	577

*: BS = Bootstrap test

**: aLRT = Approximate Likelihood-Ratio test

***: All the orthogroups for which the pattern of duplicates retention cannot be explicitly determined are assigned to the "Other" category.

to several "early branching" protists, such as members of Excavata and Chromalveolata.

To further place the duplications we identified, we added sequences from representative

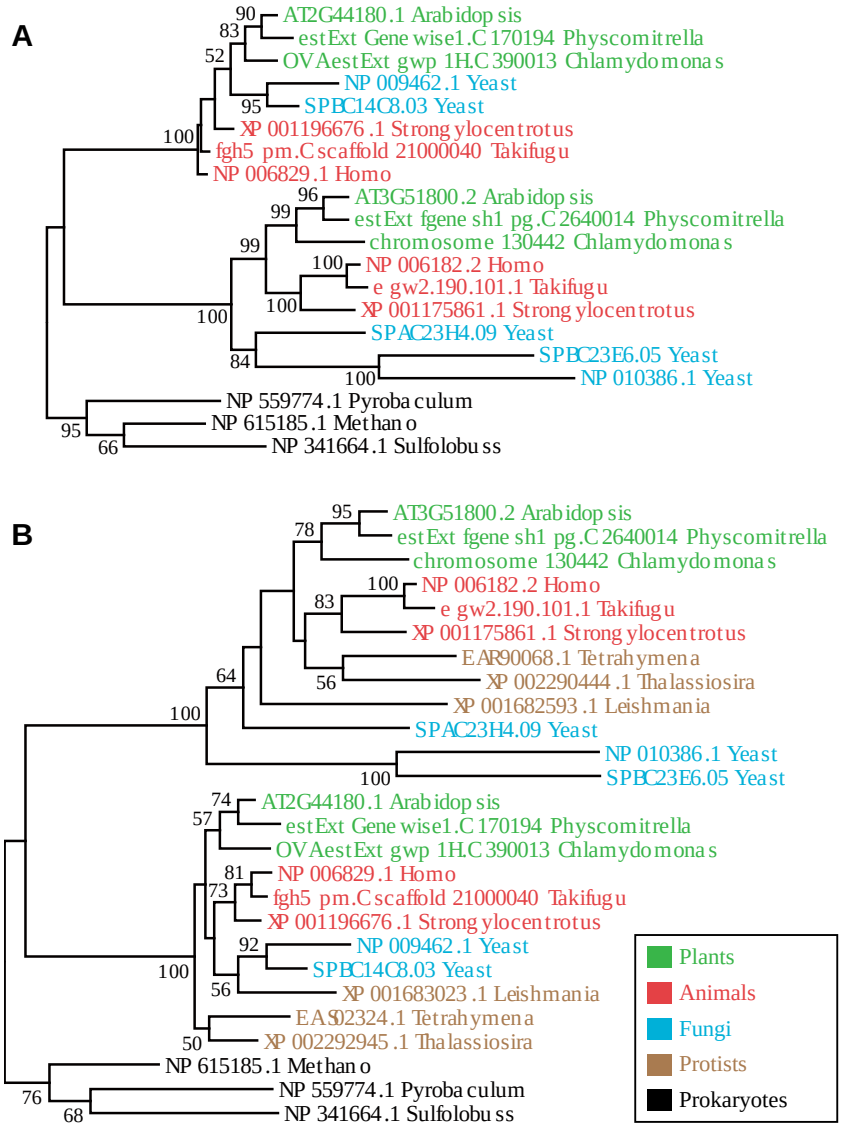


Figure 3.3 Examplar phylogenetic tree of an orthogroup (Cluster_212) with early eukaryotic duplication. (A) Topology of the ML tree, showing this orthogroup had experienced duplication before plants-animals/fungi split. (B) Topology of the ML tree with protist sequences, showing the duplication happened before the divergence of "early branching" protists.

"early branching" protists (Excavata: *Giardia lamblia*, *Trichomonas vaginalis*, *Trypanosoma brucei* and *Leishmania major*; Chromalveolata: *Plasmodium falciparum* and *Phaeodactylum tricornutum*; Amoebazoa: *Dictyostelium discoideum* and *Entamoeba*

histolytica) to orthogroups with duplication (identified by ML method at BS >= 70% support level). Additional protists (e.g. Chromalveolata: *Tetrahymena thermophila*, *Paramecium tetraurelia* and *Toxoplasma gondii*) were searched if no homolog can be found in the previous group of representative species. We found that the majority (84 out of 111) of the orthogroups had protist sequences in at least one of the paralogous clades (see Fig. 3.3 for example). Among the remaining 27 orthogroups, 19 orthogroups had no resolution, 2 orthogroups had no detectable protist homologs and only 6 orthogroups supported a different phylogeny that placed the duplication after the divergence of early protists from animals/plants. These results strongly suggested that most of these duplications were indeed very ancient events, regardless of which eukaryotic phylogenetic models ("crown-stem" or "six supergroups") were used.

3.4.3 Analysis II – MCL clusters with eukaryotic genes only

Because Analysis I required that each cluster contain some prokaryotic gene(s), the total number of the gene clusters was limited. To more widely represent the eukaryotic genomes in our study, we also examined gene clusters that contained only eukaryotic genes. Among the 41444 eukaryote-specific gene clusters (Appendix Table 3.2), there were 2276 clusters that contain members from both plants and animals/fungi, suggesting that they were likely descendants of ancestral genes in the early eukaryotes. Therefore,

Table 3.3 Number of orthogroups and early eukaryotic duplications identified in Analysis II

Method	Support	Orthogroup with duplication	Total	Percentage
NJ-BS*	>= 50%	275	1903	14.5%
	>= 70%	216		11.4%
ML-BS	>= 50%	248		13.0%
	>= 70%	194		10.2%
ML-aLRT**	>= 50%	304		16.0%
	>= 70%	283		14.9%

*: BS = Bootstrap test
**: aLRT = Approximate Likelihood-Ratio test

the phylogenies of these clusters could also provide evidence for early eukaryotic duplication. Due to the lack of prokaryotic outgroups, it was difficult to determine the root for the phylogeny of a eukaryote specific cluster. However, a duplication event could still be unambiguously inferred if a bipartition could be found in the tree in which both portions had sequences from plants and animals/fungi (see Fig. 3.1B for an illustration). It meant that the cluster should have at least two sequences from each of the plant and animal/fungal lineages.

After filtering out sequences that lack common domains, there were 1903 clusters that met this criterion and were further investigated by phylogenetic analysis. The results showed that, even at the support level of 70%, there were more than 10% of the clusters with evidence for duplication before the separation of plants and animals/fungi (Table 3.3).

3.4.4 Analysis III – reanalysis of the KOG-to-COG clusters

To further strengthen our investigation of ancient eukaryotic gene duplication, we wanted to test an independent dataset of gene clusters to evaluate the reliability of the results. We used an existing dataset of gene clusters with both eukaryotic and prokaryotic members that was established with a different methodology from that of our Analysis I [174] and performed our Analysis III. In their study, Makarova et al. used the established databases [188] of prokaryotic clusters of orthologous groups (COGs) and their eukaryotic counterparts (KOGs) to construct KOG-to-COG clusters. A COG was defined by best hits from BLAST analyses with members from at least three relatively distant prokaryotes among a total of 63 species included in the study [188]. Similarly, a KOG contains best hits from at least three eukaryotic species from a group of seven in the earlier study [188]; the total number of eukaryotes was increased to 11 subsequently [174]. The authors used RPS-BLAST search to find the best COG hit for each KOG and all the KOGs that have the same COG best-hit were assigned to one cluster [174]. In total, they identified 1092 KOG-to-COG clusters (each with one COG) which covered 2445 KOGs [174].

Since the KOG database does not include some of the representative species used in Analysis I, we first assigned the predicted protein sequences from *Physcomitrella*, *Chlamydomonas*, *Takifugu* and *Strongylocentrotus* to KOGs. Then, we extracted the

Table 3.4 Number of orthogroups and early eukaryotic duplications identified in Analysis III

	NJ-BS*		ML-BS		ML-aLRT**	
	>= 50%	>= 70%	>= 50%	>= 70%	>= 50%	>= 70%
Type I orthogroup with duplication	172	93	169	80	334	276
Type I orthogroup total	508	389	526	380	774	680
Type II orthogroup with duplication	119	59	102	49	285	301
Type II orthogroup total	724	597	605	504	581	659
Total orthogroup with duplication	291	152	271	129	619	577
Orthogroup total	1232	986	1131	884	1355	1339
Percentage	23.6%	15.4%	24.0%	14.6%	45.7%	43.1%

Type I orthogroup refers to orthogroups with a prokaryotic outgroup; Type II orthogroup refers to orthogroups without a prokaryotic outgroup.
*: BS = Bootstrap test
**: aLRT = Approximate Likelihood-Ratio test

sequences of the 14 representative species from each KOG-to-COG cluster, and retained only the clusters that had at least one prokaryotic gene and three eukaryotic genes, with at least one from plants and one from animals/fungi. As a result, 89 out of the 1092 KOG-to-COG clusters were excluded from further analysis due to their failure to meet the criteria. The phylogenies for the remaining 1003 clusters were estimated by using

both NJ and ML methods. The same criteria as used in Analysis I were followed to identify orthogroups and infer early eukaryotic gene duplication. As summarized in Table 3.4, while the total number of orthogroups (about 900 at the support level of BS >= 70%) was higher, the percentages of orthogroups with early eukaryotic duplication we observed were similar to those from Analysis I. Much higher percentages (more than 40%) of orthogroups with an early eukaryotic duplication were suggested by the ML-aLRT test at both support levels of 50% and 70% (Table 3.4). The distribution of orthogroups with different phyletic patterns was also similar to Analysis I (Table 3.2; Appendix Table 3.6).

3.4.5 Comparison of gene copy number between human and *Arabidopsis*

Many gene families have experienced duplication during the evolution of plants or animals, and gene copy can either remain similar or differ dramatically between organisms [50, 55, 171, 173, 189], possibly related to functional evolution. To further investigate the properties of families in our studies that showed detectable gene duplication before the animal-plant split, vs. the families that did not have such duplications, we plotted the gene copy number of each family in human vs. that in *Arabidopsis* and calculated the Spearman's correlation coefficients (Fig. 3.4). We found that among the families that had a prokaryotic outgroup, those exhibited the early eukaryotic duplication showed a positive correlation of gene copy number between

human and *Arabidopsis* (Fig. 3.4A), whereas the families that did not have detectable early duplication had a much less positive correlation between human and *Arabidopsis* (Fig. 3.4B). The difference between the two correlation coefficients was significant, according to the permutation test. Similarly, for the families that did not have a prokaryotic outgroup, the families with an early duplication showed a significantly stronger positive correlation than the families without the duplication (Fig. 3.4C and D).

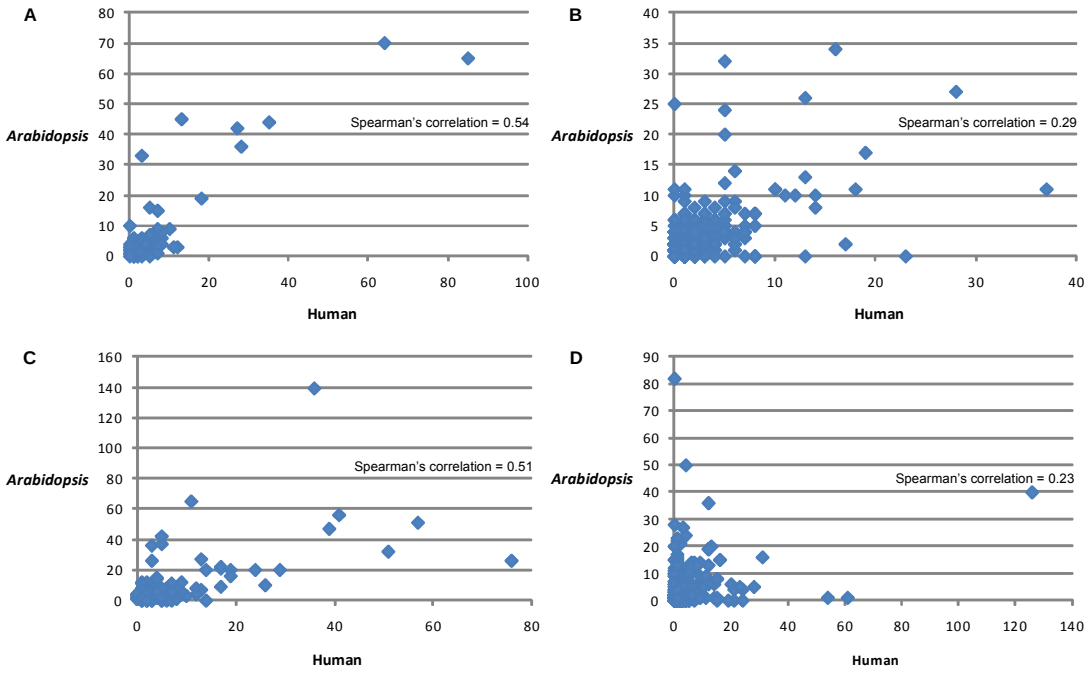


Figure 3.4 Comparison of gene copy number between human and *Arabidopsis*. The gene copy number of each family (ML approach, BS >= 70) in human vs. that in *Arabidopsis* was plotted. (A) Families with prokaryotic outgroups and early eukaryotic duplication. (B) Families with prokaryotic outgroups but no early eukaryotic duplication. (C) Families without prokaryotic outgroups but show early eukaryotic duplication. (D) Families without prokaryotic outgroups nor early eukaryotic duplication. The differences between Spearman correlation coefficients for both (A) vs. (B) and (C) vs. (D) were statistically significant (p-value < 0.01). The statistical significances were obtained through permutation test.

3.5 DISCUSSION

3.5.1 Detection of very ancient eukaryotic gene duplications

In this study, we investigated the extent of eukaryotic gene duplication before the divergence of plants and animals/fungi, by constructing gene clusters with members from representative prokaryotic and eukaryotic species and performing comprehensive phylogenetic analyses.

As we only sampled a small number of species from each lineage, additional cluster analyses were performed by adding genes from zebrafish (teleost fish), medaka (teleost fish), *Drosophila melanogaster* (insect) or the giant clam *Lottia gigantean* (mollusc), respectively. We found that adding genes from each of the additional species resulted in very slight changes of gene clusters numbers (Appendix Table 3.7). Therefore, we believe that our overall results would not be dramatically affected by inclusion of the additional animal species.

Our Analysis I was based on the gene clusters delineated by MCL method, and revealed that about 25% (BS >= 50%) or 15% (BS >= 70%) orthogroups had ancient gene duplication. Higher numbers and percentages of orthogroups that showed ancient gene duplication were reported by the ML-aLRT test (also in Analysis II and III),

possibly because the bootstrap test is consistently conservative [190]. It is known that, in comparative genomics studies like the ones we performed here, the accuracy of gene family clustering has great impact on the reliability of subsequent analyses such as phylogenetic reconstruction. Therefore, it is of interest to check whether alternative strategies of gene family clustering would lead to similar results as the MCL approach used in Analysis I. COG and its eukaryotic equivalent KOG are among the most widely used databases of orthologous gene clusters. In our Analysis III, we took the KOG-to-COG clusters identified by Makarova et al. and analyzed with the same procedures as in Analysis I. In comparison to Analysis I, in Analysis III we obtained very similar percentage of orthogroups showing early eukaryotic duplication, although the total number of orthogroups identified was higher. However, interestingly, we found that less than half of the orthogroups with duplication overlap between the two analyses. The differences were mainly caused by two reasons: first, the prokaryotic members in a particular MCL cluster were not in any COG or the corresponding COG were not in any KOG-to-COG cluster; second, a KOG-to-COG cluster may include sequences of very limited similarity, resulting in a phylogeny different from that of the corresponding MCL cluster. Nonetheless, the fact that different gene family clustering methods (MCL and COG/KOG) and phylogenetic approaches (NJ and ML) all revealed similar percentage of orthogroups that had experienced early eukaryotic duplication still supported the reliability of our results.

One possible bias in our Analysis I was that only the eukaryotic genes with detectable prokaryotic homologs were studied. This means that we focused on relatively conserved genes. In consideration of the antiquity of the gene duplication events we are interested in, some eukaryotic genes might lack detectable homologs in the prokaryotes in our study due to gene lost or sequence divergence and thus were not included in our Analysis I. For this reason, we also carried out Analysis II to analyze the eukaryote-specific MCL gene clusters and found that more than 10% of the 1903 gene clusters showed early eukaryotic duplication. It is possible that this figure is still an underestimation since some of the ancient duplicates might fail to be clustered together due to a high degree of divergence and would appear as separate gene clusters without early eukaryotic duplication.

Our phylogenetic analyses identified ~300 (BS support >= 70%) or ~500 (aLRT support >= 70%) gene duplications in the time window from the origin of eukaryotes to the plants-animals/fungi split. However, the estimation of the length of this time window varies depending on which eukaryotic phylogeny is adopted. According to the "crown-stem" model of eukaryotic phylogeny (Fig. 1.3A), plants and animals/fungi are members of a crown group and several groups of protists form deep branches in the tree [68, 81]. It was estimated that plants and animals/fungi separated at ~ 1600 million years ago (MYA), and Giardia, which was considered the deepest branch in the eukaryotic tree of life, diverged at ~ 2300 MYA [191]. Given the estimated origin of eukaryotes at ~

2700 MYA [192], the duplication events identified in our study could have taken place during the long time period before the separation of plants and animals/fungi, ~ 1100 million years. A contrasting picture is depicted by the more recent "six supergroups" classification of eukaryotes (Fig. 1.3B) [69, 193, 194]. In this model and other related models, both the "unikont-bikont" topology [74, 75] and the recent "photosynthetic-nonphotosynthetic" bipartition [76] suggest that the Archaeplastida-Opisthokonta separation might represent the first major split, or at least one of the early splits, in eukaryotic evolution (Fig. 1.3B). In this perspective, the duplication events we identified could be placed during a very early stage of eukaryotic evolution, prior to the divergence of most of the major extent protist groups.

Regardless of whether the "crown-stem" model, or "six supergroups" and other similar models were correct, we investigated gene duplications among a wider representation of eukaryotes, by phylogenetic analyses with additional sequences from examplars of divergent major protist groups, Excavata, Amoebazoa, and Chromalveolata (Fig.1.1B). For most of the gene families with 70% BS support, the duplication happened prior to the separation of these highly divergent protists from plants or animals/fungi. Even according to the "crown-stem" model of early eukaryotic history, these divergent protists separated from plants/animals/fungi at an earlier time. Therefore, irrespective of the models of early eukaryotic phylogeny, these duplications would be placed before any known major eukaryotic divergence. Therefore, our results support many gene

duplication events during very early eukaryotic evolution.

3.5.2 Functional implication for early eukaryotic evolution

The gene duplications we detected likely generated raw materials for functional evolution, as proposed before [1]. Indeed, the duplicates from the 300 or more gene duplications we identified would most likely be eliminated if they did not provide selective advantage. Therefore, these early eukaryotic gene duplications could be of great importance for the success and the radiation of early eukaryotes, and thus have been retained in the last common ancestor of extant major eukaryotic groups. If the duplicated gene families are involved in processes fundamental to early eukaryotes, which are likely to be also shared by extant eukaryotes, they might show similar evolutionary patterns in different eukaryotic kingdoms. Specifically, copy numbers for genes with highly conserved functions seem to be more than stable than the number of genes with more divergent functions (compare RAD51, MSH, SMC, vs JmjC and MADS-box genes) [50-52, 171, 173].

In fact, we observed a more positive correlation of gene family size between animals and plants in the families with early eukaryotic duplication than in the families without such duplication (Fig. 3.4). In other words, the families with the early eukaryotic duplication tend to have similar evolutionary patterns in both plants and animals/fungi

than those families without the early duplication, suggesting that these genes might have relatively conserved functions among the three major kingdoms. This idea of functional conservation is also supported by the finding that the pattern of [(RO)(RO)] retaining both duplicates in both the plants and animal/fungi lineages is the most frequent pattern among all possible patterns.

Also, it is of interest to know whether genes with specific biochemical or molecular functions or involved in specific processes are enriched among the families with duplication. Interestingly, our Gene Ontology (GO) analysis did not reveal any GO term being significantly enriched among the orthogroups with duplication (data not shown). This might suggest that the detected gene duplications, which we propose could have benefited the early eukaryotic ancestor and the ancestors of both the plant and animal/fungi lineages, affected many types of functions and processes, not just a few specialized classes of functions.

3.5.3 A hypothesis for early eukaryotic large-scale duplication

Gene duplication can be generated by several mechanisms, including tandem duplication, transposition and large-scale duplication [e.g. segmental/whole genome duplication (WGD)]. In principle, the 300 or more gene duplications we identified could be independent events through many times of tandem duplication and transposition

events. However, in the absence of supportive evidence, such a complex pattern of multiple independent events is not parsimonious. Alternatively, the duplications could be explained by one or a few large-scale duplications. Large-scale duplication like WGD is of special interest because it allows the generation of multiple new functional modules with many genes that are unrelated at the sequence level [27]. Also, segmental duplication(s) (SDs) are increasingly recognized as frequent phenomena, especially in primate genomes where ~5% of human genome consists of duplicated segments [195]. Therefore SDs with sufficiently large number of genes could also account for the gene duplications we detected. This kind of feat is not possible for other duplication mechanisms to accomplish. After WGD/SDs, the different fates of duplicated genes in different populations could generate the genetic diversity that then allows both reproductive isolation/speciation and environmental adaptation [28, 29].

The large number of ancient eukaryotic duplication events that we have detected here could have been the result of one or more early eukaryotic large-scale duplications. For relatively recent large-scale duplication events, it is possible to identify syntenic genomic regions [196]. For example, such syntenic regions were found for the most recent WGD in *Arabidopsis*, poplar and yeast, which likely occurred ~100 MYA or more recently [22, 23, 32, 134]. However, for older ones such as the WGDs in vertebrate (1R/2R; ~525-875 MYA [41], synteny is no longer detectable due to numerous genome rearrangements and gene loss [197]. If a large-scale duplication was the cause of the ancient gene duplication

events identified in this study, this event would have occurred at least 1600 MYA (possibly even earlier), making it exceedingly unlikely that any synteny can still be detected. Another approach to the detection of large-scale duplication is to analyze the rate of synonymous base substitutions (dS) between paralogous genes, as reported for many plant species [24, 198]. Unfortunately, this method is also not feasible for events older than ~150 million years because of the saturation of dS values.

An alternatively way to obtain evidence for large-scale duplication is to examine the phylogeny of a large number of gene families, as we have done here. Our results indicate that a significant fraction of the orthogroups in our dataset had experienced duplication before the divergence of the three major eukaryotic kingdoms. By combining the results of Analysis I and II, we estimated that the percentage of orthogroups showing duplication before the separation of plants and animals/fungi is over 15% (support level of BS >= 50%) and 10% (support level of BS >= 70%), or about 30% (aLRT support >= 50%) and 20% (aLRT support >= 70%). Similar large-scale phylogenetic analyses showed that, among the duplicate pairs resulting from more recent WGD in vertebrate (1R/2R; ~525-875 MYA) and yeast (~100 MYA), 26.6% and 20.1% of the pairs survived respectively [41, 199]. The early eukaryotic duplications we studied were much more ancient than the previously reported large-scale duplications in animals, plants and yeast. Thus, during the at least 1600 million years of evolution, the duplicate pairs arose in early eukaryotes might have higher chance to be lost or be too divergent to be recognized.

Therefore, it is reasonable to expect that a lower percentage of the duplicate pairs would survive, and our phylogenetic results could support the hypothesis that the duplication events identified here are the remnants of a large-scale duplication (e.g. WGD or SDs) in early eukaryotes. In other words, considering the antiquity of the early eukaryotic duplications, the 300 or more duplications we detected probably represent only a small fraction of the real number of duplications in early eukaryotes, which could be in the thousands. Our results could be most parsimoniously interpreted by one or more large-scale duplications, which were likely to be WGD/SDs, rather than thousands of independent duplications.

3.6 CONCLUSIONS

In this study, we conducted extensive phylogenetic analyses to investigate the extent of gene duplication in early eukaryotic evolution. We have found at least 300 orthogroups that had likely experienced an ancient eukaryotic duplication event, prior to the divergence of the major eukaryotic supergroups. Our results provide a better understanding of early eukaryotic evolution in several ways. The identification of numerous ancient eukaryotic gene duplication events suggests that gene duplication played important role in the evolution of early eukaryotes. The large number of duplicated genes might have allowed a large-scale evolution of new gene functions, increasing the chance of greater species diversity in changing environments. In particular, the shared duplications in plants and animals/fungi might have contributed to the three independent origins of multicellularity in these lineages. Furthermore, these ancient duplications could be most simply explained by a hypothesized early eukaryotic WGD/SDs. We further postulate that this/these WGD/SDs might have contributed to the early eukaryotic radiation. Therefore, like the cases of early vertebrate and angiosperm diversifications, the hypothesized WGD/SDs could provide an explanation at the level of genome evolution for the high rate of speciation near the origin of the three major eukaryotic lineages.

CHAPTER 4

PHYLOGENETIC RESOLUTION ACROSS WIDE EXPANSE OF EUKARYOTIC LANDSCAPE USING HIGHLY CONSERVED SINGLE-COPY GENES

4.1 SYNOPSIS

A comprehensive and reliable eukaryotic tree of life is important for many aspects of biological studies from comparative developmental and physiological analyses to translational medicine and agriculture. Both gene-rich and taxon-rich approaches have proven effective strategies to improve phylogenetic accuracy. For either approach, the understanding of the eukaryotic phylogeny will be greatly facilitated by identifying additional marker genes that are universally distributed, well conserved and orthologous across eukaryotes. However, a systematic search for marker genes suitable for the eukaryotic phylogeny is still lacking. In this study, we have identified ~1000 genes that are useful for phylogenomic studies and we show that these genes can robustly resolve the eukaryotic phylogeny, despite the challenges associated with resolving deep relationships in eukaryotes. In addition, we have demonstrated that smaller subset (~30) of these genes have strong performance in resolving controversial relationships in the eukaryotic tree of life among widely divergent taxa and at various depths, and can be obtained from organisms without genome sequences through PCR, thus are excellent for taxon-rich analyses. Strikingly, the moderate number of genes provides strong support both for the monophyly of eukaryotic supergroups and for the branching order among supergroups. Within supergroups, these marker genes result in fungal phylogenies that

are congruent with previous phylogenomic studies that used much more genes, and successfully resolve the placement of Microsporidia. In animals, we find high supports for both deep level branching patterns such as the monophyly of Protostomia and Ecdysozoa, and specific hypotheses for rooting the placental mammals. The marker genes reported here are powerful tools in both gene-rich and taxon-rich analyses and facilitate to provide a more complete picture of the eukaryotic tree of life.

4.2 INTRODUCTION

A eukaryotic tree of life provides the evolutionary framework for many facets of life sciences, including inferences of structural and functional relatedness, understanding of ecological interactions, translational medicine, and crop improvements. For example, the understanding of mammalian phylogeny can shed light on the origin and diversification of many characters, such as flight, and affect how findings from model systems can facilitate medical research. Recent methodological advances in sequence data collection and phylogenetic reconstruction have led to a substantial progress toward the goal of achieving the tree of life (e.g. [200]). Recent molecular phylogenies supported the classification of eukaryotes into six supergroups [69]: Opisthokonta (e.g. fungi and animals), Amoebozoa, Archaeplastida (e.g. green plants and red algae), Chromalveolata (e.g. brown algae and many diverse single-cell forms), Rhizaria (diverse unicellular eukaryotes, often amoeboids), and Excavata (*Euglena* and other single-cell free-living and parasitic eukaryotes). The separation of these supergroups probably represented deepest divergences in the extant eukaryotic tree of life. However, the current understanding of eukaryotic phylogeny is still rather incomplete; the eukaryotic diversity is only poorly represented in tree of life studies, and many uncertainties exist regarding relationships within and among major eukaryotic lineages [201].

Previous studies suggest that increasing both gene and taxon sampling densities are important to improve the accuracy of phylogenetic inference [85, 87, 95, 202, 203]: the sampling of more genes provides much greater resolving power than traditional gene-scale phylogenetics and thus overcomes the stochastic error; the sampling of more taxa reduces systematic errors, such as Long-Branch Attraction (LBA), which can lead to highly supported yet incorrect topologies in phylogenomic analyses. However, the number of eukaryotic species yet to be incorporated into the tree of life is vast, with new species being discovered ever more rapidly, making it impractical to acquire genome-scale data for all of them. Even with all of the genome-scale data available, the large number of species that need to be analyzed would also demand greater computational power and pose challenges in terms of computational time and phylogenetic accuracy [204].

In light of the difficulty of achieving both abundant genes and taxa, recent phylogenetic analyses prefer to only emphasize either of them. Many phylogenomic studies include more than one hundred genes and a relatively small number of taxa (e.g. [89]). In some other analyses, datasets consist of a small to moderate number of genes from a broad sampling of taxa (e.g. [94]). No matter what strategy is adopted, marker genes must be carefully chosen to avoid the violation of orthology assumption; such violation might be caused by gene duplication and/or horizontal gene transfer (HGT) and would result in incongruence between gene phylogeny and species phylogeny.

The marker genes used in recent studies of eukaryotic phylogeny mainly include previously identified universal markers (e.g. rDNA genes) and single-copy genes selected from targeted taxonomic groups, both with limitations. The number of known universal marker genes is small and a few of them (e.g. *eEF1α* [205] and *α-tubulin* [206]) have recently been shown to have a complex evolutionary history, rendering them non-orthologous. Recent phylogenomic studies included over one hundred genes [89, 207], but the gene selection usually focused on the organisms being studied. Hence, studies of different taxa usually have very different selections of marker genes (for example, the 146 genes used in a study of animal phylogeny [89] and the 246 genes used in a study of fungal phylogeny [207] only have 35 genes in common). This is not surprising because different genes are suitable for different phylogenetic questions [203, 208]. However, the lack of common thread among analyses of different organisms hinders the integration of different studies into a comprehensive eukaryotic tree of life.

Recent studies have suggested the importance of developing additional phylogenetic markers for eukaryotic phylogeny [209-211]. In addition, it is of great interest to identify genes that are suitable for analyzing organismal relationships in different parts of the eukaryotic phylogeny, which would more easily allow the assembly of a robust and complete eukaryotic tree of life. Here we report the identification of ~1000 low-copy genes that are widely distributed and well conserved across eukaryotes. We demonstrate that these genes can yield robust eukaryotic phylogeny. Furthermore, we demonstrate that

smaller subset (~30) of these genes can provide the power necessary to resolve difficult relationships across highly divergent eukaryotic lineages and at multiple levels of evolutionary depths. The marker genes we identified here can be used in both phylogenomic studies (the ~1000 genes) and taxon-rich analyses (the ~30 genes), and would greatly improve our understanding of the eukaryotic tree of life.

4.3 MATERIALS AND METHODS

4.3.1 Identification of marker genes

The complete list of orthogroups were downloaded from OrthoMCL-DB (version 4) [212]. Orthogroups in OrthoMCL-DB were delineated from 88 eukaryotic and 50 prokaryotic genomes. To obtain a more balanced representation of the eukaryotic diversity, 33 species were selected as representatives of five eukaryotic supergroups. Only the orthogroups that contain members from at least 75% of the 33 species were retained. For the 1291 selected orthogroup, phylogenetic analysis was performed as following: protein sequences from the 33 representative species were extracted and aligned using MUSCLE 3.8 [127]; conserved alignment blocks were selected using Gblocks [213]; then ML analysis was performed using RAxML 7.2.8 [182] with PROTGAMMALG option and 100 bootstrap replicates. The resulting phylogenetic trees were carefully examined (both manually and computationally using custom perl scripts) for gene duplication shared by multiple organisms. Bacterial sequences were added if the original gene tree suggested duplication(s) before the divergence of eukaryotic supergroups. Paralogous clades derived from duplication(s) in early eukaryotes were analyzed separately. An orthogroups was retained if: 1) it has no duplication, or only has

terminal duplication(s); or 2) it has a few duplications that are shared by closely related organisms.

4.3.2 Extended analysis of selected marker genes for taxon-rich analyses

For all selected species (see Appendix Table 4.1), both genomic sequence data and annotated protein sequence data of all relevant species were downloaded. The sequences of selected markers for well annotated genomes (e.g. human and Arabidopsis) were collected from the published literatures and used as queries in exhaustive homolog searches using an in-house developed program called "Phoenix" (see Supplemental Experimental Procedure). For each marker gene family, preliminary ML analysis was performed using RAxML v7.2.8 [182] to assign genes into subfamilies. The reliability of topologies was evaluated by bootstrap test with 100 replicates. Subfamilies were further analyzed using the ML approach to test for orthology. For paralogs derived from most recent duplications in individual species, only the most conserved copy was selected. For paralogs derived from duplications shared by multiple species, all paralogous copies were discarded. A special case is the well supported vertebrate-wide duplication of *SMC1*, in which *SMC1alpha* is apparently more conserved than *SMC1beta* [52]. Therefore, *SMC1alpha* was used for all analyses and both copies were used for mammal analysis. *eIF1A* has experienced multiple duplications within vertebrates; thus it was excluded

from animal and mammal analyses.

4.3.3 Phylogenetic analysis

For all the marker genes in this study, multiple sequence alignments were prepared using MUSCLE v3.8 [127]. The alignments of marker genes for taxon-rich analyses were manually inspected in GeneDoc v2.6 [214]. Conserved alignment blocks were selected using Gblocks [213] and concatenated using custom Perl script. ML analyses were performed using RAxML v7.2.8 with PROTGAMMALGF option for protein sequences and GTRGAMMA option for DNA sequences. Support values for topologies were estimated from bootstrap test with 100 replicates. Phylobayes v3.3b [215] was used for the Bayesian analyses of protein sequences under CAT model [216]. Each analysis consisted of two independent chains of 15,000 cycles.

In mammalian analysis, we generated dataset2 and dataset3 to reduce potential compositional heterogeneity caused by synonymous substitution. Dataset2 was constructed by using custom Perl script and dataset3 was constructed by using the "LeuArg1_v1_3.pl" Perl script [217]. MrBayes v3.1.2 [218] was used for the Bayesian analyses of DNA sequences under GTR+I+G model. Each analysis consisted of two independent runs, each with 1 cold chain and 3 heated chains, and each chain was run for 2 million generations. Trees were sampled every 100^{th} generations.

In animal and eukaryotic analyses, sub-datasets were constructed by removing fast evolving sites in order to test the potential impact of LBA. The fast-evolving sites were identified using PAML v4.4 [131]; sites were classified into 8 discrete gamma categories given the best scoring ML tree topology in each analysis and the top three categories were eliminated to create the sub-dataset. In animal analysis, the amino acids in the alignment were recoded into functional categories according to six Dayhoff groups. For the estimation of parallel amino acid changes, ancestral character states were estimated using PAML v4.4 [131].

Alternative topologies of previously controversial relationships in mammals were evaluated by the approximately unbiased (AU) test, the Kishono-Hasegawa (KH) test, the Shimodaira-Hasegawa (SH) test, and the weighted version of the latter two tests (wKH test and wSH test) as implemented in Consel v0.1k [219]. The site-likelihood values were calculated using RAxML under GTRGAMMA model.

4.4 RESULTS

4.4.1 Genome-scale Identification of phylogenetic marker genes

By analyzing orthogroups from OrthoMCL-DB [212], we identified 945 low-copy genes that have wide phylogenetic distribution and good orthology in eukaryotes. These genes are present in more than 75% of 33 species representing major eukaryotic lineages, and most of them (898/945) are present in all the five eukaryotic supergroups that are covered by OrthoMCL-DB. We divided the 945 genes into two classes according to their evolutionary patterns: the Class I included 480 genes which are single-copy or only show terminal duplication(s); the Class II included 465 genes which have experienced a limited number of duplication(s) shared by closely related species (e.g. duplication in early vertebrates, possibly due to the 1/2R whole genome duplication).

We then compared individual trees of the 945 genes with a reference eukaryotic phylogeny that emerges from recent studies [73, 80, 94], which is a common strategy to verify the orthology of candidate phylogenetic marker genes [220]. In this analysis, only the most conserved gene copy was kept if multiple copies exist in a species. We found that only 8.5% of the supported bipartitions (bootstrap [BS] $\geq$ 70%) in single gene trees were incongruent with the reference phylogeny. Many of the incongruencies are

regarding relationships that are known to be difficult (e.g. the monophyly of Ecdysozoa and Excavata). Well accepted clades such as animals, fungi, green plants, Stramenopiles, Apicomplexa and Euglenozoa were recovered by most supported single gene trees (Fig. 4.1). These results suggest that the genes we identified are suitable for the analysis of eukaryotic phylogeny.

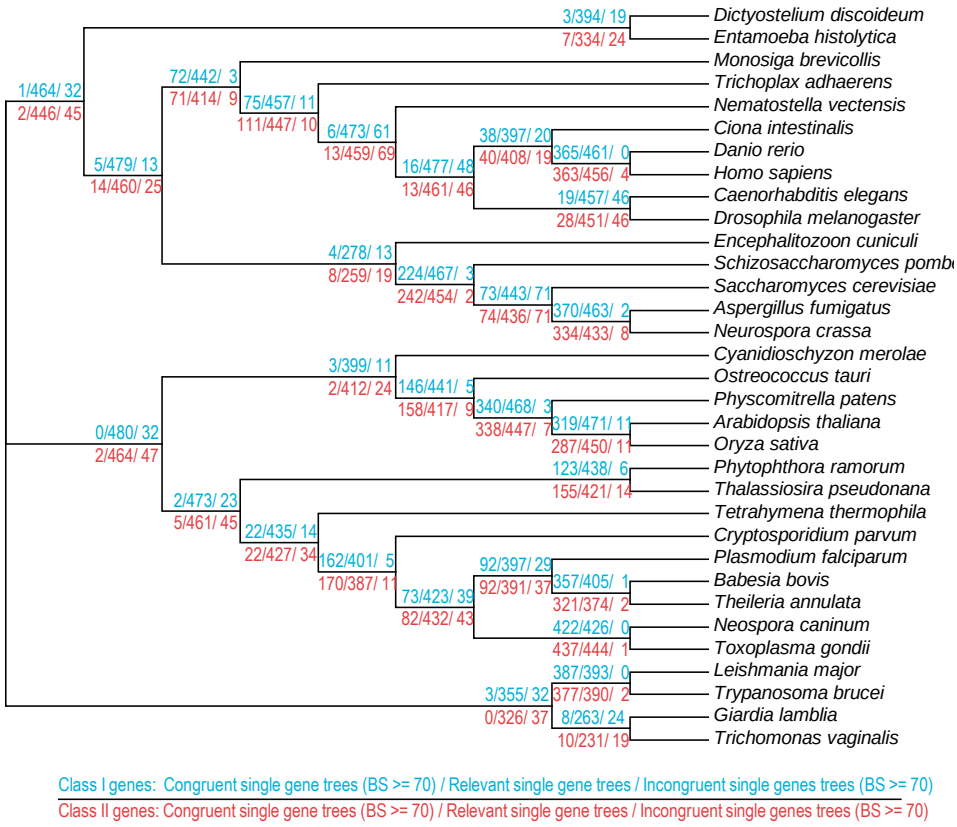


Figure 4.1 Cladogram of eukaryotes showing the supports for well-established relationships from single gene phylogenies. The tree topology is adapted from previous studies [73, 80, 94].

To evaluate the performance of the identified genes in resolving eukaryotic phylogeny, we assembled four supermatrix data sets from the 33 representative

eukaryotic species: 1) 70 Class I genes (alignable region length ≥ 400aa, average identity ≥ 40%); 2) 140 Class I genes (alignable region length ≥ 300aa, average identity ≥ 40%); 3) 78 Class II genes (alignable region length ≥ 400aa, average identity ≥ 40%); and 4) 139 Class II genes (alignable region length ≥ 300aa, average identity ≥ 40%). The concatenated analysis of all Class I (or II) genes was not feasible due to computational limitation. Bayesian analyses of the four data sets revealed the same topology, therefore only the tree from data set 1was shown (Fig. 4.2). Strikingly, the resulting tree was in

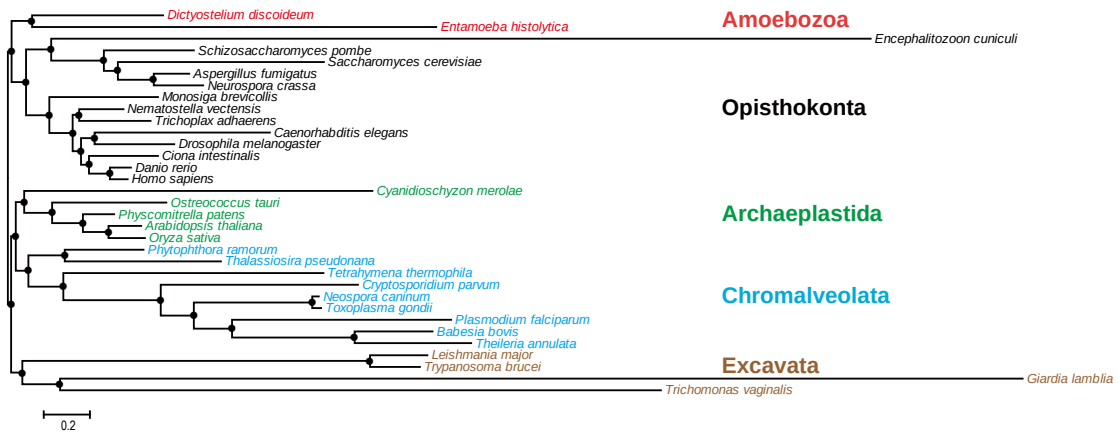


Figure 4.2 An unrooted Bayesian tree of eukaryotes using 70 Class I marker genes. The topology was estimated by Phylobayes under CAT model. The five eukaryotic supergroups are coloured as following; red, Amoebozoa; black, Opisthokonta; green, Archaeplastida; blue, Chromalveolata; and brown, Excavata. The branch leading to Giardia is shown as a quarter of the original length. Black dots indicate 100% support from Posterior Probability (PP).

complete agreement with the reference eukaryotic phylogeny, with the only exception of the grouping of *Trichoplax adhaerens* and *Nematostella vectensis*. All the relationships, including deeper ones such as the monophyly of the five supergroups and the branching

order among supergroups, received maximum support. Our results strongly support the utility of the genes we identified in studying the eukaryotic tree of life.

4.4.2 Selection of marker genes for taxon-rich analyses

Among the 945 genes, there are commonly used universal markers (e.g. elongation factors and ATPase subunits) and recently reported markers for taxon-rich analyses of the eukaryotic phylogeny [210]. In order to identify additional marker genes for taxon-rich analyses, we selected a subset of the 945 genes and performed thorough analyses in different eukaryotic lineages. Here, we focused on five gene families, including *recA/RAD51* [50], *MutS* [51], *MutL* [51], *SMC* [52] and *MCM* [53], mainly because previous phylogenetic studies showed that genes in these families are broadly distributed, highly conserved and orthologous in representative eukaryotes, making them excellent candidate markers.

These families each comprises several paralogs that resulted from ancient duplications before the diversification of major eukaryotic lineages and have maintained unambiguous orthologous relationships for major eukaryotic groups [50-53] except for a vertebrate-wide duplication of *SMC1* [52] and a few recent lineage (family)-specific duplications. Genes in these five families have essential functions in DNA replication,

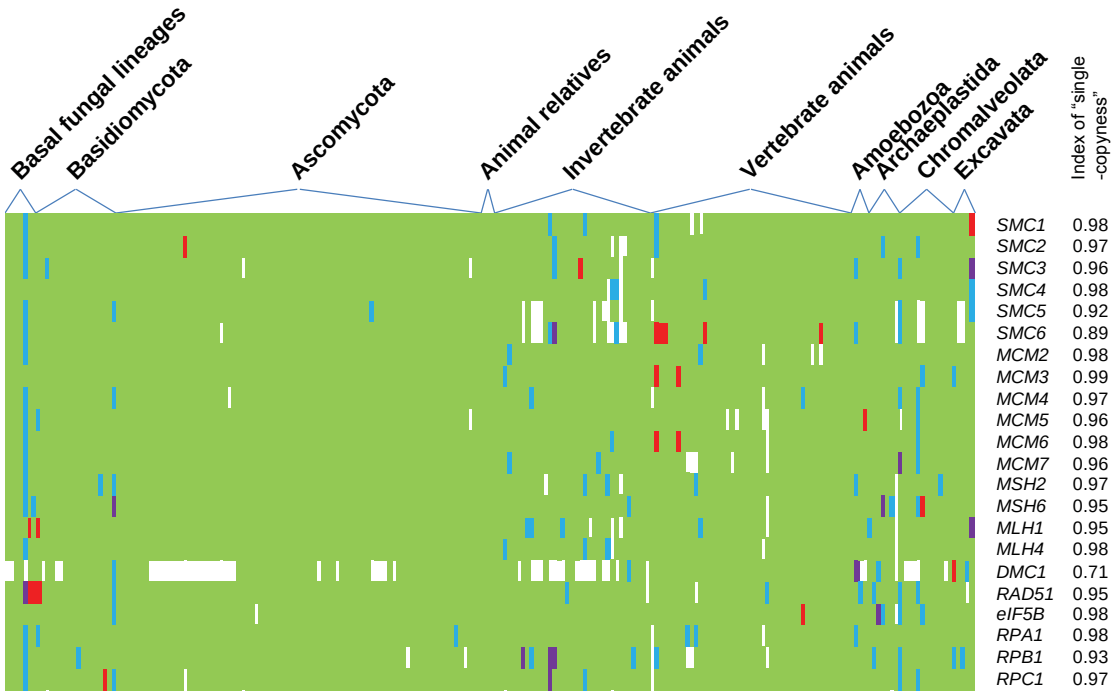


Figure 4.3 A matrix showing the distribution of identified marker genes in eukaryotes. The presence/absence of genes is highlighted by colour: blank, absence; green, single copy; blue, two copies due to intra-specific duplication; purple, more than two copies due to intra-specific duplications; red, more than one copy due to duplications shared by more than one species. The Index of Single Copyness (ISC) is defined as $(\sum_{i=1}^{n} \mathbf{1}/\boldsymbol{m_i} - \boldsymbol{k})/\boldsymbol{n}$, where n is the total number of species, mi is equal to the gene copy number for species with a single copy of the gene or more than one copies of terminal paralogs (mi is greater than 0; for species that do not have the gene, $1/m_i = 0$), k is equal to the total number of species with paralogs shared by two or more species. This matrix only includes the 22 genes that were used in all four major analyses in this study (i.e. phylogenetic analyses of fungi, animals, mammals, and eukaryotes).

repair and recombination. To reduce possible bias of only sampling from genes that function in DNA replication, repair, and recombination, we also included a few widely used phylogenetic markers related to transcription (*RPA1*, *RPB1* and *RPC1*) or translation (*eIF1A* and *eIF5B*). Our exhaustive homolog searches of these genes in a large number of sequenced eukaryotic genomes confirmed the single-copy status of these genes in most

eukaryotes (Fig. 4.3). Although in a few cases two or more copies of a gene were detected in one species, our phylogenetic analyses of individual genes indicated that most such paralogs were derived from intra-specific duplications.

We then further examined the individual gene phylogenies to investigate if these genes followed the orthology assumption. Since single genes usually have very limited phylogenetic signal, it is expected that single gene phylogenies would display variations in topology and have many unresolved nodes. Therefore, instead of a stringent comparison between our single gene phylogenies and the topology based on previous multi-gene dataset, we verified the monophyly of previously well-supported clades of closely related species. For example, major fungal clades, including basal lineages such as Microsporidia and Mucoromycotina as well as lineages within Basidiomycota and Ascomycota, all received moderate to high supports from most of the single gene phylogenies (Appendix Fig. 4.1). Major animal clades such as nematodes, arthropods, vertebrates and mammals were also well supported. These results suggest that the genes we identified have most likely maintained orthologous relationship in eukaryotes.

In total, we identified approximately 30 genes that are widely present and have essential functions such as replication, transcription and translation, suggesting that they are less prone to horizontal gene transfer (HGT) [221]. Furthermore, these genes are well conserved and most of them are fairly large (coding for proteins with >700aa, except for

recA/RAD51 genes and *eIF1A*). Therefore, it is relatively easy to identify gene regions that are well conserved within lineages (or even supergroups) for primer design. As an experimental validation, we have successfully amplified six genes (*DMC1, RAD51* and *SMC1-4*) from eight *Aspergillus* species (diverged ~200 MYA) using degenerate primers. Moreover, some of these genes have regions that are constant across wide eukaryotic diversity and universal primers can be designed.

These desirable features make these genes promising molecular markers for eukaryotic phylogeny. If shown to have good resolving power at various taxonomic levels, the same selected markers can be used to study different branches on the eukaryotic tree of life and will allow easy integration of these separate studies. Therefore, we assessed the performance of the selected markers in resolving relationships within or among major eukaryotic lineages. As a test of principle, we examined here eukaryotic phylogenies using the species with complete genome sequences (Appendix Table 4.1).

4.4.3 Deep relationships within and between eukaryotic supergroups

To investigate the general applicability of the highly conserved genes to wide eukaryotic diversity, we analyzed the most ancient relationships in the eukaryotic tree of life. The estimation of relationships among supergroups is challenging and the monophyly of some of the supergroups (e.g. Chromaveolata) has not been not

consistently supported [201]. A recent analysis with broad taxon sampling and moderate number of genes yielded promising results yet several major nodes received only low support [94], such as the monophyly of Amoebozoa and the grouping of Amoebozoa and Opisthokonta. Therefore, our analysis of the eukaryotic phylogeny has two major objectives; first, to test the utility of the selected markers at a broad taxonomic level; and second, to probe the deepest relationships in the eukaryotic tree of life.

Eukaryotic diversity has been very unevenly sampled by genome sequencing projects, with the overwhelming majority of the sequenced genomes belonging to Opisthokonta (animals and fungi), while no Rhizaria genome is currently available. To maximize the representation of eukaryotic diversity in our study, we included nearly all of the sequenced genomes in Amoebozoa, Chromalveolata and Excavata, and selected species from fungi (including Microsporidia), animals, plants and their closely related protists, with a total of 39 species, including representatives from all five supergroups that have whole-genome sequences (Appendix Table 4.1). The Maximum Likelihood (ML) analysis provided strong support for many previously reported relationships, including the monophyly of Opisthokonta, but also had obvious misplacement of several species (e.g. the grouping of the *Entamoeba histolytica* with Excavata; see Appendix Fig. 4.2), likely due to LBA. Therefore, as others have suggested [222-224], we think that the topology from Bayesian analysis with CAT model is more reliable.

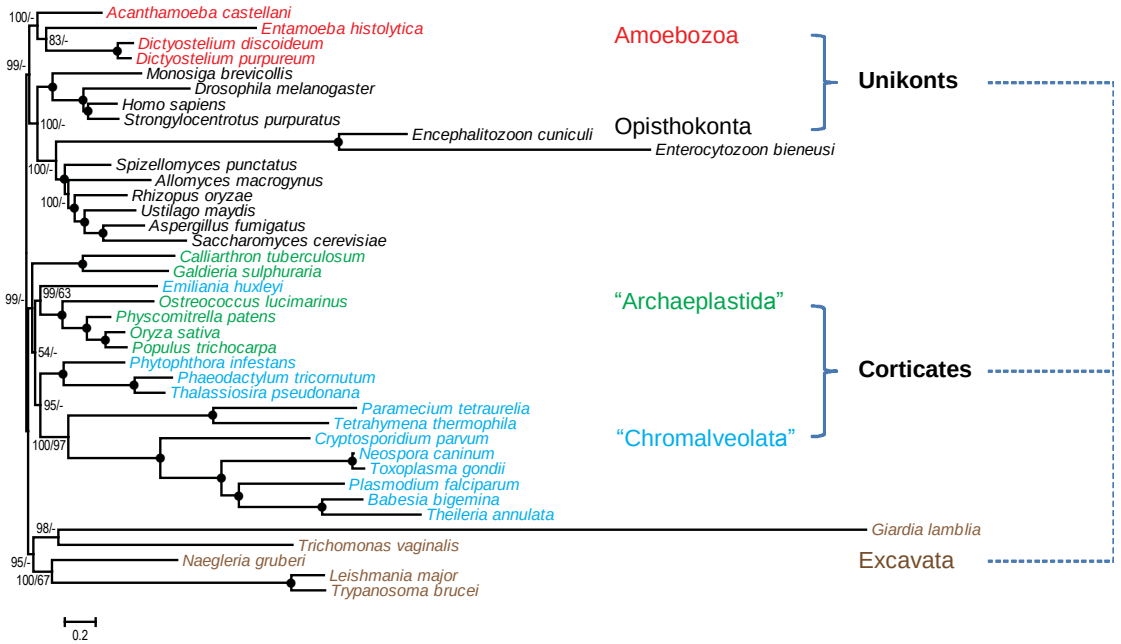


Figure 4.4 An unrooted Bayesian tree of eukaryotes using selected markers for taxon-rich analyses. The topology was estimated by Phylobayes under CAT model. The five eukaryotic supergroups are coloured as following; red, Amoebozoa; black, Opisthokonta; green, Archaeplastida; blue, Chromalveolata; and brown, Excavata. The branch leading to Giardia is shown as a quarter of the original length. Black dots indicate 100% support from both Posterior Probability (PP) and bootstrap (BS). Support values from PP/BS are shown for nodes that do not receive 100% support. Dashes indicate inconsistent topology between Bayesian and ML methods.

Our Bayesian analysis was able to recover a robust phylogeny with strong support for overwhelming majority of nodes in the eukaryotic tree. In the Bayesian inference (Fig. 4.4), there was strong support for the monophyly of Opisthokonta, Amoebozoa, and Excavata, much stronger than previous single-gene analyses, where the monophyly of these supergroups has often been uncertain [201]. Our result indicated that, not only their monophyly, but also their internal nodes were robustly resolved. Within the supergroup Opisthokonta, our results strongly supported a close relationship of Microsporidia with

fungi. In addition, the supergroup Amoebozoa was strongly supported even with the highly divergent *E. histolytica*. The supergroup Excavata encompasses protists with diverse characteristics and its phylogeny has been difficult to resolve previously, resulting in the paraphyly of Excavata from both gene-scale and genome-scale analyses [201, 209, 225]. Strikingly, with only a moderate number of genes here, we successfully recovered the monophyletic Excavata clade and even its internal topology is in accordance with other analyses [73, 80, 94].

The supergroup Chromalveolata is a large group of diverse organisms and is further divided into several major groups, including Apicomplexans, Ciliates, Dinoflagellates, Cryptomonads, Haptophytes and Heterokonts. All but the ciliates have evidence for secondary endosymbiosis of red algae-like primary photosynthetic organisms. Our analysis included 12 species representing four of the major groups, Apicomplexans, Ciliates, Haptophytes, and Heterokonts, which were all robustly resolved in our tree. In addition, our analysis strongly supported a sister relationship between Apicomplexans and Ciliates, with Heterokonts forming a more basal group. However, the monophyly of the entire supergroup was not recovered here. Specifically, Haptophytes, represented by *Emiliania huxleyi*, was placed sister to green plants with strong support. Similarly, the supergroup Archaeplastida was not recovered in our analysis; the evolutionary affiliation of red algae remained unresolved. The Bayesian analysis provided almost equal support to 1) the sister relationship between red algae and green plants plus Haptophytes, to 2) the

grouping of red algae and the remaining Chromalveolata, and to 3) the basal position of red algae in a clade uniting Archaeplastida and Chromalveolata.

Recent analyses have also hypothesized several different scenarios for the monophyly of and relationships between the eukaryotic supergroups [201]. In this regard, our eukaryotic tree strongly supported the union of Opisthokonta and Amoebozoa ("Unikonta"), which was proposed to be related on the basis of shared internal duplication of phosphofructokinase [226]. In addition, members of Archaeplastida and Chromalveolata formed a well-supported clade ("Corticates"; Fig. 4.4), even though photosynthesis in these groups has different origins. The markers genes used here are not related to photosynthesis or other plastid functions, suggesting that these diverse groups share a nuclear ancestry separate from their similarity in energy production. Therefore, the dominant multicellular heterotrophs (animals and fungi) and autotrophs (plants and algae) occupy respective positions that have been separated very early in eukaryotic history, with many groups of unicellular protists being more related to either animals or plants than these latter ones are to each other. Excavata was placed between the Unikonta and the Corticates (Fig. 4.4), indicating that they are very distant from most eukaryotes, including all major groups of multicellular organisms. This topology was consistent with previous phylogenomic results [73, 80, 94], with stronger support for the sisterhood of Archaeplastida and Haptophytes. Our results were also compatible with a new scenario of eukaryotic origin proposed by Cavalier-Smith [227], where the root of the eukaryotic tree

is placed within the current Excavata supergroup, while Unikonta and Corticates (Archaeplastida, Rhizaria and Chromalveolata) form a clade called Neozoa [227]. Till now, the root of eukaryotic tree still remains highly uncertain. The fact that the selected markers here all have ancient paralogs might provide a great opportunity to probe the root of eukaryote tree via the reciprocal rooting strategy.

4.4.4 Well resolved fungal phylogenies

To further assess the utility of the selected markers within eukaryotic lineages, we studied the phylogeny of fungi, the lineage that encompasses a largest fraction of completely sequenced eukaryotic genomes. The relationships within fungi have been extensively studied [228-230], allowing a comparison of our results. Our fungal dataset consisted of 29 genes (Appendix Table 4.2) from 108 fungal species (Appendix Table 4.1), which covers almost all of the sequenced fungal genomes, except for a few closely related to species already included here. We first analyzed two subsets of species (42 and 70, respectively) that matched, respectively, those studied previously [228, 229]. Our results using 29 genes (Appendix Fig. 4.3 and Appendix Fig. 4.4) were in excellent agreement with the previously reported phylogenies based on 153 genes or whole genome data, except for minor differences regarding a few controversial relationships (e.g. relationships within the CTG clade), indicating that the selected markers have strong

resolving power, as much as those of studies using many more genes.

We then analyzed the large dataset of 98 fungal species with both Bayesian and ML approaches (only representatives from *Fusarium*, and *Saccharomyces*, *Trichoderma* and *Verticillium* were included in the 98 fungal species dataset, due to computational limitation). The two approaches resulted in phylogenies that are largely congruent and provided maximum supports for more than 80% of all nodes including both higher level and recent relationships (Fig. 4.5; Appendix Fig. 4.5 and Appendix Fig. 4.6). Microsporidia are rapidly evolving parasites and their phylogenetic placement has been controversial and difficult to resolve. Earlier phylogenies based on small subunit ribosomal RNA showed an early divergence of Microsporidia in the eukaryotic tree of life [67, 231], which was likely affected by the extreme long branches of Microsporidia [232]. Recent analyses of protein coding genes with improved phylogenetic methods have indicated a fungal origin of Microsporidia [230, 233-236], yet the placement of Microsporidia with fungi remains largely unresolved; different possibilities have been suggested, including a position within Zygomycota [234] and a sister relationship to almost all fungi [230], but these hypotheses were not well supported. In particular, we found maximum support from both Bayesian and ML analyses for the basal-most position of Microsporidia in fungi, again supporting the usefulness of the genes used here.

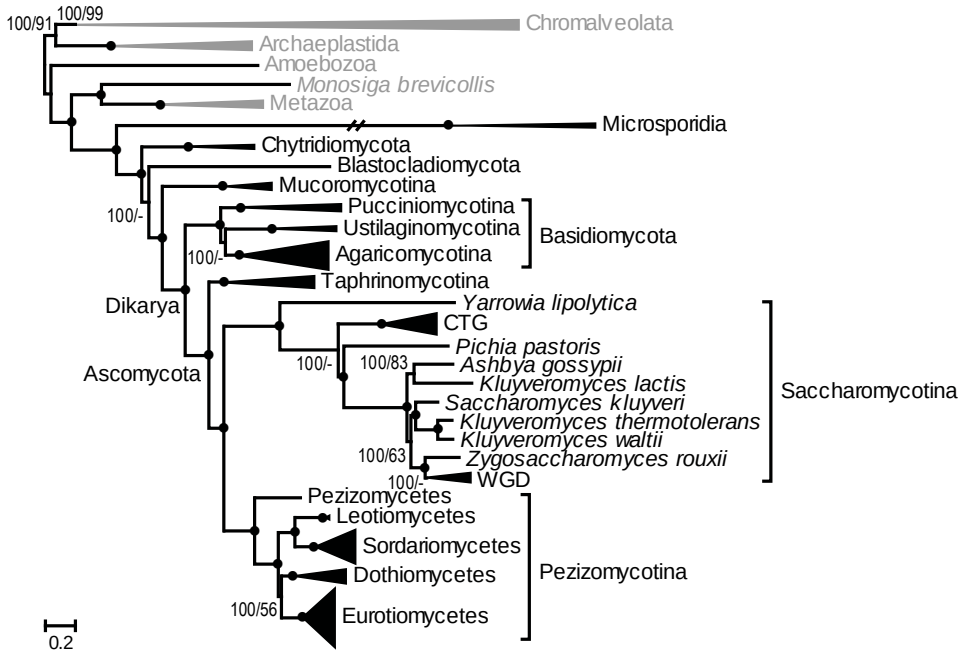


Figure 4.5 Cladogram of 98 fungal species with three animals as outgroups. The topology was estimated by the Phylobayes using the CAT model. Black dots indicate 100% support from both Posterior Probability (PP) and bootstrap (BS) support from ML analysis (based on 100 replicates). Support values are only shown for nodes that do not receive 100% support. Dashes indicate <50 % support from PP or BS, or inconsistent topology between Bayesian and ML methods.

Furthermore, the Bayesian analysis resulted in maximal support for progressively basal positions for Blastocladiomycota and Chytridiomycota (Fig. 4.5; Appendix Fig. 4.5), although ML analysis did not support these relationships (Appendix Fig. 4.6). Regardless of their relative positions, Chytridiomycota and Blastocladiomycota were placed at the base of fungi by both analyses. Chytridiomycota species are morphologically similar to oldest fungal fossils, consistent with their phylogenetically very basal position. Both Chytridiomycota and Blastocladiomycota species are aquatic and have flagellated gametes, suggesting that these are ancestral characteristics of fungi

but are lost in more derived lineages. Recently, a tour de force phylogeny of about 200 fungi was reported using six genes [230]; however, this phylogeny had low support for the placements of basal lineages such as Microsporidia, Chytridiomycota and Blastocladiomycota, probably because six genes did not have sufficient phylogenetic signal for these relationships. In this study, we have demonstrated the power of the ~30 marker genes in resolving basal fungal phylogeny by using sequenced genomes. The moderate number of genes can be obtained from taxa that lack genome sequences; much greater taxon sampling is also easily feasible in these basal lineages and holds the promise of determining their relationships.

Regarding the other fungal lineages, there was strong support for the sister relationship of Mucoromycotina (Zygomycota) with Dikarya, the subkingdom that includes the phyla Basidiomycota and Ascomycota. Within the phylum Ascomycota, we found strong support for the sister relationship of two major subphyla, Saccharomycotina and Pezizomycotina, and the basal position for Taphrinomycotina. In addition, we also recovered the monophyly of two clades with unusual genetic or genomic characteristics: the CTG clade including species that translate CTG as serine instead of leucine and the WGD clade including species that have experienced a whole-genome duplication. Our analyses also resolved the positions of several lineages that were not included in previous phylogenomic studies [228, 229]. For example, Pezizomycetes (represented by *Tuber melanosporum*) was placed at the base of Pezizomycotina with high confidence. Also, the

monophyly of Taphrinomycotina (represented by *Pneumocystis carinii* and Schizosaccharomyces) received maximum support from both Bayesian and ML analyses.

Within the CTG clade, our results were different from the two reported phylogenies [228, 229] (Appendix Fig. 4.5 and Appendix Fig. 4.6), which include a monophyletic clade formed by haploids *Candida lusitaniae*, *Candida guilliermondii*, *Debaryomyces hansenii* and *Pichia stipits* [228, 229], and a closer relationship between the latter two [229]. In contrast, our results placed *C. lusitaniae* at the base of the CTG clade and *P. stipites* sister to a clade of diploid species including *Candida albicans*. The topology here suggests that haploid represents the ancestral status of the CTG clade. Interestingly, our analysis of 42-species revealed the same result as the published phylogenies with similar taxon sampling [228], while our analyses of more species (70 and 108 species) supported a different topology, suggesting the possible impact of increased taxon sampling.

Similarly, improved taxon sampling here also led to a change in the position of Dothideomycetes, a class in Pezizomycotina with nearly 20 thousand morphologically diverse species. Our fungal tree showed strong support for the monophyly of the subphylum Pezizomycotina and its four main classes: Dothideomycetes, Eurotiomycetes, Sordariomycetes and Leotiomycetes. However, the relationships among these four classes have been controversial [228]; this has limited the understanding of the origin and evolution of traits such as lichenization and ascoma shape in Pezizomycotina. For example, both Dothideomycetes and Eurotiomycetes have members that are lichenized

and can produce fissitunicate asci, whereas Sordariomycetes and Leotiomycetes are nonlichenized and they produce unitunicate asci [237]. Therefore, different phylogenies of these four clades would lead to very different reconstruction of the ancestral status. Although the closer relationship between Leotiomycetes and Sordariomycetes was strongly supported in our analyses and a few others [228, 229, 238, 239], previous studies suggested three alternative placements of Dothideomycetes; 1) sister to Eurotiomycetes [228, 229]; 2) sister to Leotiomycetes-Sordariomycetes [228, 238]; or 3) basal to the other three clades [239]. When only one species from Dothideomycetes was included, our analysis of 42 species and the 42-species supermatrix phylogeny of Fitzpatrick *et al* [228] both supported the second placement. However, with more species sampled in the four clades, especially in Dothideomycetes, the first topology became supported by our analysis of 108-species datasets as well as the 82-species phylogeny of Wang *et al* [229], suggesting that lichenization and the production of fissitunicate asci might have a recent origin in the ancestor of Dothideomycetes and Eurotiomycetes.

Moreover, inclusion of more Basidiomycota species in our study also resulted in a better understanding of the previously uncertain relationship among Pucciniomycotina (e.g., *Puccinia graminis*), Ustilaginomycotina (e.g., *Ustilago maydis*) and Agaricomycotina (e.g., *Agaricus bisporus*) [230]. Our analysis of 70-species favored the sister relationship between Pucciniomycotina and Agaricomycotina (Appendix Fig. 4.4), but our Bayesian analysis of 108-species placed Pucciniomycotina basal to

Ustilaginomycotina-Agaricomycotina with maximum support (Appendix Fig. 4.5), again highlighting the importance of the taxon-rich approach. The union of Ustilaginomycotina and Agaricomycotina in our results is consistent with previous analysis of the structure of septal pores which is an important ultrastructural character in fungi [240]. Septal pore swellings can be found in both Ustilaginomycotina and Agaricomycotina, while Pucciniomycotina only possesses simple septa. Well resolved relationships among the major clades are important for understanding the evolution of such biological traits in Basidiomycota. Sampling with even greater number of taxa, as facilitated by using the marker genes described here, will likely yield better understanding of the relationships among these ecologically and economically important groups.

To test whether a well resolved fungal phylogeny can be recovered with even fewer genes, we successively removed genes that were lost in some taxa (*MSH4-5*, *MLH2-3* and *DMC1*) and genes related to transcription and translation (*RPA1*, *RPB1*, *RPC1*, *eIF1A* and *eIF5B*), resulting in two datasets with 24 and 19 genes, respectively. Both datasets revealed almost the same topology as the 29-gene dataset (Appendix Fig. 4.7 and Appendix Fig. 4.8), with maximum support for about 75% of all nodes. In conclusion, our analyses of fungal phylogeny indicate that the moderate number of markers we identified can achieve the same level of accuracy and resolution as genome-scale data.

4.4.5 Strongly supported relationships between metazoan clades

Animals phylogeny is of wide interest and have been greatly modified in recent years yet many relationships remains unsettled [241]. The deep relationships at the base of Metazoa are unclear as even very recent phylogenomic studies do not agree on the positions of basal lineages, including Porifera (e.g. sponges), Placozoa (e.g. *Trichoplax*) and Cnidaria (e.g. corals) [90, 242-244], relative to Bilateria, which encompasses the vast majority of animal diversity. Competing hypotheses include the Epitheliozoa hypothesis that places Placozoa as sister to Cnidaria-Bilateria with the exclusion of Porifera, the Urmetazoon hypothesis that groups Porifera, Placozoa and Cnidaria into a monophyletic clade [243], and other possibilities such as Placozoa being the closest relative of Bilateria [244].

Within Bilateria, the relationships among vertebrates and invertebrate clades such as nematodes and arthropods have also been hotly debated [245]. The traditional "Coelomata" hypothesis proposed that Coelomata includes vertebrates, arthropods and others, and the earlier divergence of Pseudocoelomata (e.g. nematodes) and Acoelomata (e.g. Platyhelminthes). Alternatively, a more recent animal phylogeny proposed the division of Bilateria into Deuterostomia and Protostomia, the latter of which is further divided into Ecdysozoa and Lophotrochozoa.

We analyzed animal phylogeny with a dataset including 29 genes from 8

representative vertebrates and 35 invertebrates with three closely related protists and two fungi as outgroups (Appendix Table 4.1 and Appendix Table 4.2). The ML analysis yielded a topology that is obviously affected by LBA (Appendix Fig. 4.9); therefore, we analyzed the dataset with the Bayesian approach and the more realistic CAT model

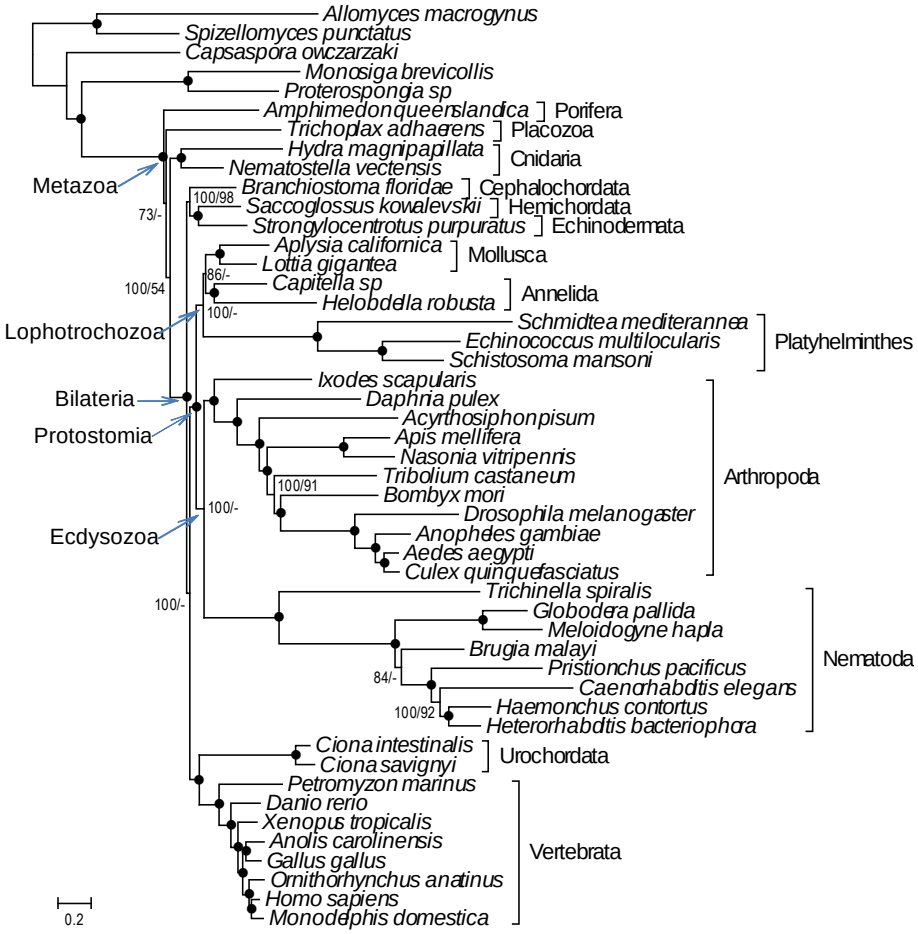


Figure 4.6 Phylogenetic relationships of animals. A Bayesian tree of 43 animal and five outgroup species. The tree was constructed by Phylobayes using the CAT model. Black dots indicate 100% support from both Posterior Probability (PP) and bootstrap (BS). Support values from PP/BS are shown for nodes that do not receive 100% support. Dashes indicate <50 % support from PP or BS, or inconsistent topology between Bayesian and ML methods.

[246]. The Bayesian inference (Fig. 4.6) had maximum support for the majority of nodes and most inferred relationships were consistent with those reported in recent phylogenomic studies [90, 242]. Several controversial relationships were resolved in our study. At the base of animal phylogeny, our result supported the Epitheliozoa hypothesis despite limited taxon sampling. The well-supported sister relationship of Cnidaria and Bilateria supports a single origin of the neural system, which is found in both groups, but not in either Porifera or Placozoa. In addition, the monophyly of Protostomia, Ecdysozoa and Lophotrochozoa were all highly supported and the detailed relationships inside these clades were also largely congruent with those revealed by other studies [89, 247].

However, some noticeable unusual aspects of our result were the polyphyly of Chordata and the paraphyly of Deuterostomia. These topologies have been observed previously [247, 248], and the impacts of compositional heterogeneity and unrealistic phylogenetic model have been suggested [248, 249]. Therefore, we performed additional analyses aiming to overcome various phylogenetic artefacts and further investigated these unexpected relationships. As a results, the Chordata monophyly was recovered when distant outgroups were removed (Appendix Fig. 4.10 and Appendix Fig. 4.11), or if the protein data set was recoded into functionally similar groups (Appendix Fig. 4.12), suggesting that the position of Cephalochordata revealed in Fig. 4.6 is likely artefactual. On the other hand, Deuterostomia remained paraphyletic in all of our experiments,

including the removal of distant outgroups (Appendix Fig. 4.10 and Appendix Fig. 4.11), the recoding of protein data set (Appendix Fig. 4.12), and the removal of fast-evolving sites (Appendix Fig. 4.13). It is possible that more efficient strategy to reduce phylogenetic artefacts and improved taxon-sampling near the base of Deuterostomia might provide a solution for the Deuterostomia paraphyly.

Deuterostomia is a major bilaterian group; it is proposed based on embryonic developmental characters (i.e. the mouth develops from a second opening at the opposite end from the blastopore, which becomes the anus) and receives consistent high supports from previous molecular phylogenies [250, 251]. However, the monophyly of Deuterostomia was questioned or only poorly supported in recent phylogenomic studies [247, 252, 253]. In addition, two groups of organisms, Lophophorata and Chaetognatha, were previously considered members of Deuterostomia, because the mouth of these animals develop at the opposite end of the blastopore [241], but have been placed in Protostomia by recent molecular studies [250, 254].

Here our Bayesian analysis of a moderate number of carefully selected, highly conserved single-copy genes also supported the paraphyly of Deuterostomia, thus further challenging the long-held idea of Deuterostomia monophyly. If the basal position of Hemichordata and Echinodermata in Bilateria is accepted, there would be important revisions of our understanding of the bilateral evolution. Firstly, it would suggest that several Deuterostomia characters, such as the secondary formation of the mouth during

gastrulation and the radial cleavage, were already present in the ancestor of Bilateria. This notion is further supported by the fact that two lineages (Lophophorata and Chaetognatha) with the defining characteristics of Deuterostomia (development of mouth from the opposite end from the blastopore and radial cleavage) have been placed in Protostomia by molecular phylogenetic analyses [250, 254], suggesting a deuterostome-like ancestor of Protostomia and Deuterostomia.

In addition, this topology would also shed light on the understanding of the evolution of the nervous system centralization. It has been proposed that the central nervous systems (CNS) in Chordata and Protostomia originated from an ancestral CNS in early Bilateria [255]. This would imply that Echinodermata, Hemichordata and Xenoturbellida, which are the closest relatives of Chordata according to the conventional Deuterostomia phylogeny, should also have CNS homologous to the dorsal Chordata CNS and ventral Protostomia CNS. However, the highly divergent Echinodermata CNS and diffuse nervous system in Xenoturbellida were instead considered as support for the independent origin of the Chordata CNS [256]. Our result suggests a closer relationship among Chordata and Protostomia, thus favouring the common origin of CNS in these two clades. The recently characterized CNS in adult Hemichordata possesses both dorsal and ventral neural strands [257], which might represent a status ancestral to the dorsal Chordata CNS and ventral Protostomia CNS.

4.4.6 Shallow Divergences in mammals

Besides the relatively early animal phylogeny, recent relationships in mammals have also received great attention [258]. The knowledge of mammalian phylogeny is fundamental to many aspects of mammalian evolutionary studies which include, but not limited to: establishing the evolutionary history of morphological and physiological characters, determining the divergence time among lineages, and guiding the comparative genomic studies among mammals. For example, the traditional mammalian phylogeny based on morphology has supported the grouping of Xenarthrans (armadillos) with pangolins, both of which have poorly-developed molars, and the sister relationship between bats and flying lemurs, which have common features such as the flight membrane [258]. Later molecular phylogenies have refuted these relationships, suggesting that the morphological similarities among these species are due to convergent evolution [258]. However, despite the availability of genome-scale data and continuous endeavour, there are still several controversial relationships in mammals [259], including: 1) the order of divergence among Afrotheria (e.g. elephant), Xenarthra (e.g. armadillo and sloth) and Boreoplacentalia (most other placental mammals); 2) the position of bats within Laurasiatheria (e.g. horse, dog, and cow); 3) the relationships within Rodentia regarding squirrel, guinea pig and mouse-rat; and 4) the relative orders of tree shrew, Glires and Primates.

Previous studies using rare genomic changes (e.g. retroposons) [260-264], small number of house-keeping genes [265] and large genomic dataset (e.g. combined analysis of ~ 3000 genes) [259, 266-269] were unable to confidently resolve these relationships. For instance, the analyses of retroposon insertions, which are assumed homoplasy-free and more reliable than sequence substitution, found similar supports for each of the three alternative branching patterns among Afrotheria, Xenarthra and Boreoplacentalia [260, 261]. Phylogenomic studies with different sets of species and genes also yielded inconsistent topologies. It is possible that, due to mechanisms such as incomplete lineage-sorting, introgression and hybrid species, the divergences of these controversial lineages might not follow a strict bifurcation thus the relationships are difficult to resolve [259, 266]. However, the uncertainties in mammal phylogeny are also possibly due to the inadequacy of current phylogenetic approaches; rare genomic changes being also subject to homoplasy [270], and genome-scale data not necessarily generating the correct phylogeny [85].

The moderate number of highly-conserved single-copy genes we selected might provide better resolution for some of the problematic relationships in mammals. To test the effectiveness of selected markers for resolving mammal phylogeny, we constructed a dataset of 35 genes from genomes of 35 mammals and 3 vertebrate outgroups (Appendix Table 4.1 and Appendix Table 4.2). Our analysis of protein sequences was able to provide strong support for the root of placental mammals; however, the other

controversial relationships could not be confidently resolved (Appendix Fig. 4.14), probably due the very high degrees of sequence conservation. Therefore, in addition to the analysis using protein sequences, we also used the DNA sequences of the protein coding regions for mammal phylogeny because they might provide more phylogenetic signals. Two additional datasets were generated by sequentially removing the 3rd codon position and the 1st position in the codons coding for L and R to reduce potential compositional heterogeneity caused by synonymous substitution [217]. Our analyses of different datasets with two phylogenetic methods (ML and Bayesian) yielded the same topology (Fig. 4.7) and resolved some of the controversial relationships with high support.

Regarding the root of placental mammals, our analyses of both protein and DNA datasets uniformly and strongly supported the position of Afrotheria basal to Xenarthra and Boreoplacentalia (Fig. 4.7; Appendix Fig. 4.14), and alternative topologies could be confidently rejected by most datasets and statistical methods (Appendix Table 4.3). In previous phylogenomic studies, the relationships among these three clades were unstable and sensitive to taxon sampling; for example, the analyses of 22 species [259] and 37 species [266] supported the grouping of Afrotheria and Xenarthra and the basal position of Afrotheria, respectively. We also analyzed a smallerdataset with the 22 species as in a previous study [259] and the earlier divergence of Afrotheria was still supported (Appendix Fig. 4.15), suggesting that our results were reliable. Early analyses of

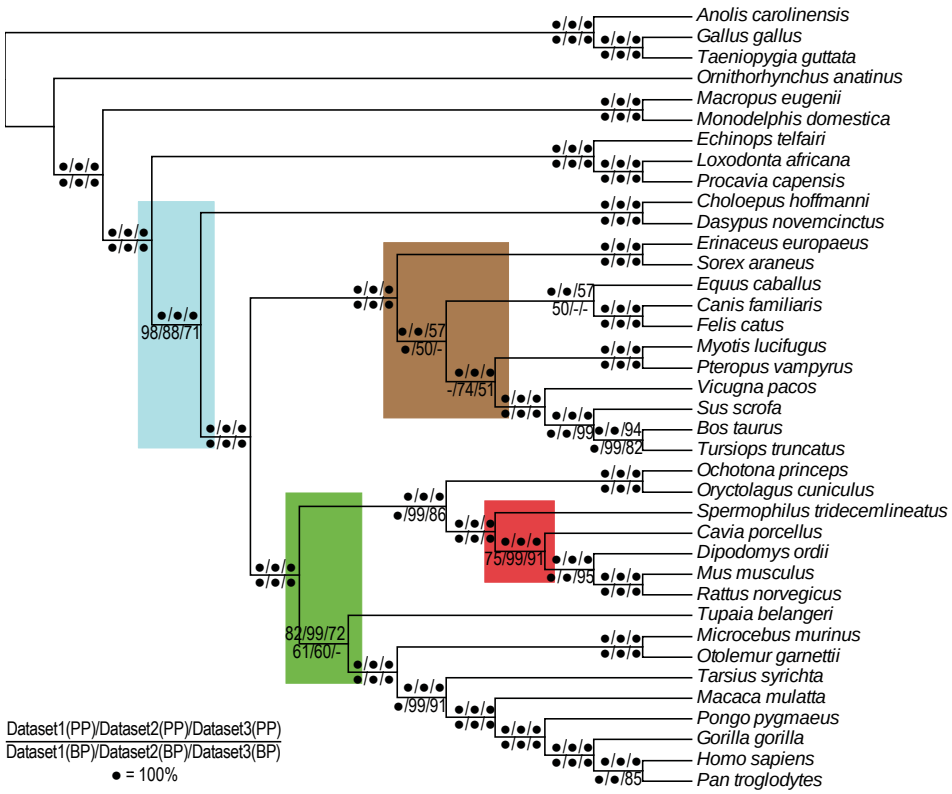


Figure 4.7 Phylogenetic relationships of mammals. Cladogram of 35 mammal species with green anole and birds as outgroup. Previously controversial relationships are highlighted by colour: blue, the relationships among Afrotheria, Xenarthra and Boreoplacentalia; brown, position of bats within Laurasiatheria; red, the relationships among squirrel, guinea pig and mouse-rat; green, position of tree shrew. Both Bayesian and ML approaches were employed to analyze each of the three mammal datasets: dataset1 included all three codon positions; dataset2 included the first two positions of each codon; and dataset3 excluded the first position in the codons coding for L and R from dataset2. The topology shown was consistently recovered by different datasets and phylogenetic methods. Dashes indicate <50 % support from Posterior Probability (PP) or bootstrap (BS), or inconsistent topology between Bayesian and ML methods.

morphological characters led to the “Epitheria” hypothesis which unites Afrotheria and Boreoplacentalia [271]. The basal position of Afrotheria that was revealed in our study strongly suggests that the defining features of “Epitheria” (e.g. the “stirrup-shaped stapes in the middle ear”) are likely ancestral to all placental mammals.

Besides the earliest relationships among placental mammals, our results also supported the sister relationship between mouse-rat and guinea pig instead of squirrel. In addition, the grouping of bats with Artiodactyla (e.g. cow), the grouping of Perissodactyla (horse) with Carnivora (e.g. dog), and the grouping of tree shrew with Primates, received strong support from some of the analyses, indicating the resolving power of these marker genes. We further performed statistical tests for alternative topologies of these problematic branches (Appendix Table 4.3). The results showed that positions of squirrel and bats revealed in our analyses are strongly favored and alternative topologies could be confidently rejected by most datasets and statistical methods. For positions of tree shrew and horse, alternatives were less favored but could not be rejected. In summary, our phylogenetic analyses and topology tests lend strong supports to the basal position of Afrotheria and squirrel, but resolution of remaining controversies might require further expansion in taxon sampling.

4.5 DISCUSSION

4.5.1 The discovery of a wealth of eukaryotic markers

In this study, we performed a systematic search and identified 954 eukaryotic marker genes. The development of additional phylogenetic markers is important for the study of eukaryotic phylogeny and there have been continuous efforts in recent years. For example, Tekle et al. analyzed the KOG database and identified 17 maker genes that are promising for eukaryotic phylogeny [210]. In addition, a list of 146 genes curated by Phillipe et al. [89] has been widely used in recent phylogenomic analyses and the performance of these genes has been demonstrated in resolving organismal relationships within and among eukaryotic lineages [73, 91, 92, 242, 272]. Almost all the genes identified in [210] (15/17) and [89] (135/146) were recovered in this study, suggesting that our approach is able to capture genuine marker genes for eukaryotic phylogeny. On the other hand, there is much less overlap between the genes we identified and two other gene sets used in phylogenomic studies focusing on more restricted groups of taxa [90, 207]. Only 153 of the 246 single-copy genes identified in fungi [207] and 70 of the 150 single-copy genes identified in animals [90] were recovered by us. Among the 93 fungal single-copy genes and the 80 animal single-copy genes that were rejected in our study,

most genes (65 and 64, respectively) failed to meet our criterion on phylogenetic distribution.

In total, 652 out of the 954 eukaryotic marker genes are newly identified in this study; they are not included in any of the forementioned gene sets [89, 90, 207, 210]. The implication of our results is two-fold. First, the 652 new marker genes we present here is a valuable supplement to currently available eukaryotic markers. Sequence features of these new marker genes, such as alignable region length and average identity, are similar to those of the 146 genes [89] (Appendix Fig. 4.16). In addition, the phylogenetic informativeness profiles of the new marker genes indicate that they carry phylogenetic signals even for ancient relationships in eukaryotes. At the deepest nodes, most genes have informativeness per site values greater than 0.2, which is as good as the markers identified in [210]. More importantly, we have demonstrated the great resolving power of these genes; a highly resolved eukaryotic phylogeny can be obtained using data sets which mainly (more than 50%) consist of the new marker genes (Fig. 4.2). Therefore, the 652 newly identified marker genes have characteristics and performance that are at least comparable to other eukaryotic markers, suggesting that they are promising tools for eukaryotic phylogeny.

Second, our phylogenetic results provide additional support for the eukaryotic phylogeny shown in Fig. 4.1. Previous molecular phylogenetic evidence for the relationships within and among eukaryotic supergroups mainly came from similar gene

sets (e.g. a few universal markers, see [201]); especially, most recent phylogenomic studies were based on the 146 genes [73, 80, 273-275]). It is of great interest to compare the results from independent sets of marker genes. Our data sets, which was denser in gene sampling and included many new marker genes, strongly supported the results of previous phylogenomic studies (Fig. 4.2). Together with other current available markers, the marker genes we report here may provide important insights into the understanding of the evolutionary history of eukaryotes.

4.5.2 Implications for eukaryotic phylogeny

Our phylogenetic analyses demonstrated the excellent performance of the marker genes we identified. At the same time, the well resolved phylogenies have enabled the test of several important hypotheses, including some that have been hotly debated in recent years, such as the origin of Microsporidia, relationships within bilateria, and the monophyly and relationships of deep eukaryotic supergroups.

Microsporidia: Microsporidia are rapidly evolving parasites and their phylogenetic placement has been controversial and difficult to resolve. Earlier phylogenies based on small subunit ribosomal RNA showed an early divergence of Microsporidia in the eukaryotic tree of life [67, 231], which was likely affected by the extreme long branches of Microsporidia [232]. Recent analyses of protein coding genes with improved

phylogenetic methods have indicated a fungal origin of Microsporidia [230, 233-236], yet the placement of Microsporidia with fungi remains largely unresolved; different possibilities have been suggested, including a position within Zygomycota [234] and a sister relationship to almost all fungi [230], but these hypotheses were not well supported. Our analyses consistently placed Mircrosporidia basal to other fungi (Fig. 4.2, Fig. 4.4 and Fig. 4.5), suggesting a fungal origin of Microsporidia.

Protostomia: Relationships between invertebrate and vertebrates animals within Bilateria has been controversial. The Coelomata hypothesis is originally based on morphological data and it is still supported by recent genome-scale analyses [276, 277]. However, these analyses included limited number of species and their results were likely affected by phylogenetic artefacts such as LBA. On the other hand, the new animal phylogeny is supported with broader taxon sampling and strategies to overcome LBA [89, 90, 278]. Our analysis of metazoan phylogeny using the ~30 marker genes strongly supported the monophyly of Protostomia and its inner clades. These results are remarkable given the relatively small number of taxa and genes.

Eukaryotic supergroups: Both our analyses of the larger data set (Fig. 4.2) and the ~30 selected markers (Fig. 4.4) provided strong support for the monophyly of major eukaryotic lineages and the relationships among them. For example, the monophyly of the highly divergent Excavata was recovered with high confidence values.

However, the ~30 selected markers did not support the supergroup Chromaveolata

(Fig. 4.4); Haptophytes was placed sister to green plants. Other recent studies also suggested the grouping of Haptophytes and Cryptomonads with the primary photosynthetic Archaeplastida [73, 80, 94]. These intriguing relationships seem robust to the removal of fast-evolving sites [73, 80], thus are unlikely to be caused by LBA. A possible explanation for the close affinity of primary (green plants) and secondary (e.g., Haptophytes) photosynthetic lineages might be due to phylogenetic signals from the secondarily endosymbiotic genomes via HGT. However, the marker genes we selected all have core functions (not related to the chloroplast) and are less prone to HGT. In fact, the nucleomorph (secondary endosymbiont) genome of the Cryptomonad *Hemiselmis andersenii* still retained homologs of some of the marker genes used here, instead of having them been transferred into the new host nuclear genome, arguing against the possibility of HGT for the markers used here. Regardless whether some of the lineages had experienced HGT, the topology of Chromalveolata strongly suggest that Haptophytes (and perhaps Cryptomonads) had a different history from the others, such as Ciliates and Apicomplexans. Therefore, our phylogenetic results support the paraphyly of Chromalveolata and raise the possibility of two or more separate secondary endosymbiosis events in different lineages of Chromalveolata, in addition to the likely independent secondary endosymbiosis event in *Euglena* (an Excavate) and its close relatives [279].

In addition, the position of red algae was not resolved in our analysis of the ~30

selected markers (Fig. 4.4). One possible reason is the limited taxon sampling; in our analysis, Archaeplastida was only represented by red algae and green plants. Another major lineage in Archaeplastida, Glaucophyta, was not included, because no sequenced genome is available. The monophyly of Archaeplastida has also been uncertain in recent phylogenetic analyses [94, 201, 209, 280]. As suggested by other studies [94, 209], improved taxon sampling might be critical for understanding the evolutionary history of Archaeplastida. Interestingly, the grouping of red algae and green plants received maximum support from the larger data set (Fig. 4.2). It is possible that, given the limited taxon sampling in Archaeplastida, much more genes are needed to resolve the relationships among these photosynthetic lineages.

4.5.3 The marker genes are useful for both gene-rich and taxon-rich approaches

Both gene-rich and taxon-rich approaches have greatly contributed to the assembly of the eukaryotic tree of life. Although the relative importance of more gene or more taxa has been hotly debated for a long time, this controversy is largely attributed to our limited ability of sequence acquisition. It is unlikely that, genome-scale data from a wide selection of organisms will soon become available, which is desirable for phylogenetic analyses of eukaryotic phylogeny. However, recent gene-rich and taxon-rich analyses have both emphasized more balanced sampling of genes and taxa. On the one hand, in

gene-rich analyses, broader taxon sampling has been achieved through EST projects in a few targeted organisms, which help to alleviate systematic errors in phylogenomic studies [89, 90, 244]. One the other hand, taxon-rich analyses have expended the selection of marker genes [94]; a moderate number of genes carry much stronger phylogenetic signals than previous analyses using single, or small number of genes. In the near future, at least, both approaches will continue to play important roles in the study of eukaryotic phylogeny.

In this study, we identified 945 promising eukaryotic marker genes. These genes can be sampled through transcriptomic/genomic projects, and are valuable tools for phylogenomic studies. We also provided additional information of these genes, such as alignable region length, average identity, and phylogenetic informativeness profile. Based on such information, researchers can freely decide which genes to include in their analyses. In addition, we demonstrated that smaller subset of these genes can provide sufficient resolving power for both deep divergences in eukaryotes and relatively recent relationships in various eukaryotic lineages. We also showed that the ~30 selected genes can be obtained from organisms without genome sequences through PCR. Therefore, these selected genes are excellent for taxon-rich analyses combined with moderate number of markers. It should also be noted that, the ~30 selected genes we present here are not necessarily the only or the best ones for taxon-rich analyses; there might be many more in the set of 945 genes. Again, the large number of marker genes we identified are

highly useful for both gene-rich and taxon-rich analyses; they will greatly facilitate our understanding of the evolutionary history of eukaryotes.

APPENDIX

Appendix Table 2.1 Ka/Ks analysis of animal *KDM1A* and *KDM1B* genes

A. Ka/Ks ratios between animal *KDM1A* genes

	Cow	Mouse	Chicken	Frog
Human	0.0010	0.0010	0.0010	0.0036
	Pufferfish	Sea urchin	Sea anemone	Sea squirt
Human	0.0034	0.0052	0.0022	0.0080

B. Ka/Ks ratios between animal *KDM1B* genes

	Cow	Mouse	Chicken	Frog
Human	0.0945	0.0547	0.0613	0.0407
	Pufferfish	Sea urchin	Sea anemone	Sea squirt
Human	0.0256	0.0348	0.0289	0.0148

C. LRT for difference in Ka/Ks ratio between animal KDM1A and KDM1B clades

Models	$Ka/Ks_{(KDM1A)}$	$Ka/Ks_{(KDM1B)}$	Log-likelihood	LRT
One-ratio Model	0.0234	= $Ka/Ks_{(KDM1A)}$	-14180.1864	NA
Two-ratio Model	0.0036	0.0708	-14093.6031	$P < 0.0001$

Appendix Table 3.1 Summary of representative species included in this study

	Species	Data source	Predicted proteins	Analyzed proteins*
Plants	*Arabidopsis thaliana* (Flowering plant)	TAIR Version 8 (www.arabidopsis.org)	32825	13299
	Physcomitrella patens (Moss)	JGI Version 1.1 (ftp.jgi-psf.org/pub/JGI_data/)	35938	9517
	Chlamydomonas reinhardtii (Green algae)	JGI Version 4.0 (ftp.jgi-psf.org/pub/JGI_data/)	16709	4697
Animals	*Homo sapiens* (Human)	Build 36.3 (ftp.ncbi.nih.gov/genomes/)	37742	11830
	Takifugu rubripes (Pufferfish)	JGI Version 4.0 (ftp.jgi-psf.org/pub/JGI_data/)	26721	9205
	Strongylocentrotus purpuratus (Sea urchin)	Build 2.1 (ftp.ncbi.nih.gov/genomes/)	42420	18495
Fungi	*Saccharomyces cerevisiae* (Budding yeast)	Build 2.1 (ftp.ncbi.nih.gov/genomes/)	5861	2872
	Schizosaccharomyces pombe (Fission yeast)	06Aug2008 (ftp.sanger.ac.uk/pub/yeast/)	5026	2844
Eubacteria	*Escherichia coli str. K-12 substr. MG1655* (Gram-negtive bacteria)	(ftp.ncbi.nih.gov/genomes/Bacteria/)	4132	953
	Rickettsia prowazekii str. Madrid E (Gram-negtive bacteria)	(ftp.ncbi.nih.gov/genomes/Bacteria/)	835	277
	Bacillus subtilis subsp. subtilis str. 168 (Bacilli)	(ftp.ncbi.nih.gov/genomes/Bacteria/)	4105	894
Archaea	*Methanosarcina acetivorans C2A*	(ftp.ncbi.nih.gov/genomes/Bacteria/)	4540	720
	Sulfolobus solfataricus P2	(ftp.ncbi.nih.gov/genomes/Bacteria/)	2977	707
	Pyrobaculum aerophilum str. IM2	(ftp.ncbi.nih.gov/genomes/Bacteria/)	2605	516

*: proteins included in MCL clusters analyzed in Analysis I and II.

Appendix Table 3.2 Summary of MCL gene clustering results

Eukaryotes		Prokaryotes	Number of clusters
Archaeplastida	Opishtokonts		
+	+	+	794
+	-	+	443
-	+	+	157
+	+	-	2276
+	-	-	21874
-	+	-	17294
-	-	+	8558
Total			51396

Appendix Table 3.3 Summary of gene families known to have experienced early eukaryotic gene duplication

Gene family	Reference	Gene cluster	Pattern
SMC	[52]	Analysis I: OG_101	(RO)(RO)
recA	[50]	Analysis I: OG_102	(RO)(RO)
MutS	[51]	Analysis I: OG_67	(RO)(RO) by ML-aLRT / (RO)(R) by NJ
MutL	[51]	Analysis I: OG_100	(RO)(RO)
Spo11	[281]	Analysis I: OG_426	(RO)(R) by ML-BS and NJ / (RO)(RO) by ML-aLRT
Kinesin	[172]	Analysis II: OG_10	(RO)(RO)
MADS	[171]	The paralogous clades divided into separate gene clusters (Type-I - Analysis II: OG_52)	No duplication
KDM1	[173]	Analysis II: OG_143	(RO)(RO)
JmjC	[173]	Analysis II: OG_426 (PKDM11 and JMJD6)	(RO)(RO)
RDRP	[282]	The paralogous clades do not cover the representative species used in this study	
RNA Pol II	[283]	Analysis I: OG_64	(RO)(R) by ML-aLRT
DNA Pol	[284]	Analysis I: OG_142	(RO)(RO)
MCM	[285]	Analysis I: OG_53	(RO)(RO)
TCP1	[286]	Analysis I: OG_26	(RO)(O) by ML-BS / (RO)(RO) by ML-aLRT
Proteasome subunits	[287]	Analysis I: OG_29	(RO)(RO) by ML-aLRT

R – Archaeplastida; O – Opishtokonta.

References:

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci* **97**: 5328-5333.

Archambault, J. and Friesen, J.D. 1993. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol Rev* **57**: 703-724.

Filee, J., Forterre, P., Sen-Lin, T., and Laurent, J. 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* **54**: 763-773.

Gupta, R.S. 1995. Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. *Mol Microbiol* **15**: 1-11.

Hughes, A.L. 1997. Evolution of the proteasome components. *Immunogenetics* **46**: 82-92.

Kearsey, S.E. and Labib, K. 1998. MCM proteins: evolution, properties, and role in DNA replication. *Biochim Biophys Acta* **1398**: 113-136.

Lin, Z., Kong, H., Nei, M., and Ma, H. 2006. Origins and evolution of the *recA/RAD51* gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci* **103**: 10328-10333.

Lin, Z., Nei, M., and Ma, H. 2007. The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution. *Nucleic Acids Res* **35**: 7591-7603.

Malik, S.B., Ramesh, M.A., Hulstrand, A.M., and Logsdon, J.M., Jr. 2007. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Mol Biol Evol* **24**: 2827-2841.

Miki, H., Okada, Y., and Hirokawa, N. 2005. Analysis of the kinesin superfamily: insights into structure and function. *Trends Cell Biol* **15**: 467-476.

Surcel, A., Zhou, X., Quan, L., and Ma, H. 2008. Long-term maintenance of stable copy number in the eukaryotic *SMC* family: origin of a vertebrate meiotic *SMC1* and fate of recent segmental duplicates. *J Syst Evol* **46**: 19.

Zhou, X. and Ma, H. 2008. Evolutionary history of histone demethylase families: distinct evolutionary patterns suggest functional divergence. *BMC Evol Biol* **8**: 294.

Zong, J., Yao, X., Yin, J., Zhang, D., and Ma, H. 2009. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447**: 29-39.

Appendix Table 3.4 Test of the impact of long-branch attraction on orthogroups with vulnerable topologies

	Cluster ID	Original support	Support after adding sequences
(PAF)(P)	371	>= 70	>= 50
	287	>= 50	Undetermined*
	478	>= 50	>= 50
(PA)(P)	87	>= 50	>= 50
	146	>= 70	>= 70
	696	>= 50	>= 50
(PF)(P)	119	>= 70	>= 70
	177	>= 50	>= 50
	716	>= 70	>= 50
(PAF)(A)	187	>= 70	>= 70
	288	>= 50	Undetermined
	369	>= 50	>= 50
(PA)(A)	63	>= 50	>= 70
	365	>= 50	>= 70
	44	>= 70	>= 70
(PAF)(F)	232	>= 50	Undetermined
	396	>= 70	>= 70
	485	>= 50	Undetermined
(PF)(F)	414	>= 70	>= 70
	711	>= 50	No duplication
	725	>= 50	Undetermined

P – Plants; A – Animals; F – Fungi.

*: Undetermined means the topology could not be resolved at bootstrap support level of 50.

In this analysis, a subset of the orthogroups that have topologies potentially vulnerable to

long-branch attraction (LBA) was arbitrarily selected to test the impact of LBA. Additional sequences from the following species were added to each orthogroup:

(PAF)(P), (PA)(P), (PF)(P): *Selaginella moellendorfii* (spikemoss), *Oryza sativa* (rice) and *Vitis vinifera* (winegrape);

(PAF)(A), (PA)(A): *Nematostella vectensis* (sea anemone), *Ciona intestinails* (sea squirt) and *Xenopus tropicalis* (frog);

(PAF)(F), (PF)(F): *Ustilago maydis*, *Aspergillus* and *Neurospora crassa*.

Appendix Table 3.5 Distribution of orthogroups with phyletic patterns supporting early eukaryotic duplication – Analysis I

		NJ-BS		ML-BS		ML-aLRT	
		>= 50%	>= 70%	>= 50%	>= 70%	>= 50%	>= 70%
(RO)(RO)	(PAF)(PAF)	46	38	49	33	65	63
	(PAF)(PA)	12	6	8	6	12	12
	(PAF)(PF)	2	2	3	2	6	5
	(PA)(PA)	8	5	7	3	11	9
	(PF)(PF)	1	0	1	0	1	1
	(PA)(PF)	4	1	3	2	7	5
	Total	73	52	71	46	102	95
(RO)(R)	(PAF)(P)	25	14	23	13	36	29
	(PA)(P)	25	15	20	10	26	22
	(PF)(P)	6	2	12	6	13	12
	Total	56	31	55	29	75	63
(RO)(O)	(PAF)(AF)	1	1	2	1	4	3
	(PAF)(A)	15	12	16	4	22	21
	(PAF)(F)	11	6	8	3	9	9
	(PA)(AF)	0	0	1	1	2	2
	(PA)(A)	25	11	12	7	20	21
	(PF)(AF)	2	2	2	2	2	2
	(PF)(F)	5	2	5	3	5	4
	Total	59	34	46	21	64	62

R – Archaeplastida; O – Opishtokonta; P – Plants; A – Animals; F – Fungi.

Appendix Table 3.6 Distribution of orthogroups with phyletic patterns supporting early eukaryotic duplication – Analysis III

		NJ-BS		ML-BS		ML-aLRT	
		>= 50%	>= 70%	>= 50%	>= 70%	>= 50%	>= 70%
(RO)(RO)	(PAF)(PAF)	48	23	52	20	158	145
	(PAF)(PA)	18	5	18	7	69	59
	(PAF)(PF)	4	2	4	2	10	7
	(PA)(PA)	12	7	13	7	43	40
	(PF)(PF)	1	0	1	1	1	2
	(PA)(PF)	7	3	4	2	18	15
	Total	90	40	92	39	299	268
(RO)(R)	(PAF)(P)	26	15	30	13	67	53
	(PA)(P)	34	16	30	11	58	51
	(PF)(P)	12	10	20	9	31	32
	Total	72	41	80	33	156	136
(RO)(O)	(PAF)(AF)	3	1	4	1	8	8
	(PAF)(A)	38	19	20	7	52	53
	(PAF)(F)	20	6	10	3	24	28
	(PA)(AF)	1	0	0	0	7	6
	(PA)(A)	23	11	16	5	54	45
	(PF)(AF)	0	0	2	1	2	3
	(PF)(F)	9	4	10	5	9	7
	Total	94	41	62	22	156	150

R – Archaeplastida; O – Opishtokonta; P – Plants; A – Animals; F – Fungi.

Appendix Table 3.7 Results of MCL clustering analyses with genes from additional animal species

	Number of clusters with genes from Archeaplastida, Opisthokonta and prokaryotes	Number of eukaryote-specific clusters with genes from both Archeaplastida and Opisthokonta
Original dataset	794	2276
Adding genes from zebrafish	804	2273
Adding genes from medaka	807	2261
Adding genes from *Drosophila*	807	2271
Adding genes from *Lottia*	814	2336

Appendix Table 4.1 List of 88 eukaryotic species included in OrthoMCL-DB

Supergroup	Species	33 representative species
Opisthokonta	*Encephalitozoon cuniculi GB-M1*	+
	Aspergillus fumigatus Af293	+
	Aspergillus oryzae RIB40	
	Candida glabrata CBS 138	
	Coccidioides immitis RS	
	Coccidioides posadasii RMSCC 3488	
	Cryptococcus neoformans var. neoformans JEC21	
	Debaryomyces hansenii CBS767	
	Eremothecium gossypii	
	Gibberella zeae PH-1	
	Kluyveromyces lactis NRRL Y-1140	
	Laccaria bicolor S238N-H82	
	Magnaporthe grisea 70-15	
	Neurospora crassa OR74A	+
	Phanerochaete chrysosporium RP-78	
	Pichia stipitis CBS 6054	
	Saccharomyces cerevisiae S288c	+
	Schizosaccharomyces pombe	+
	Yarrowia lipolytica CLIB122	
	Monosiga brevicollis MX1	+
	Trichoplax adhaerens	+
	Nematostella vectensis	+
	Acyrthosiphon pisum	
	Aedes aegypti	
	Anopheles gambiae str. PEST	
	Apis mellifera	
	Bombyx mori	
	Brugia malayi	
	Culex pipiens	
	Drosophila melanogaster	+
	Pediculus humanus	
	Caenorhabditis briggsae AF16	
	Caenorhabditis elegans	+

	Schistosoma mansoni	
	Ciona intestinalis	+
	Danio rerio	+
	Takifugu rubripes	
	Tetraodon nigroviridis	
	Gallus gallus	
	Ornithorhynchus anatinus	
	Monodelphis domestica	
	Canis lupus familiaris	
	Mus musculus	
	Rattus norvegicus	
	Homo sapiens	+
	Pan troglodytes	
Amoebozoa	*Dictyostelium discoideum AX4*	+
	Entamoeba dispar SAW760	
	Entamoeba histolytica HM-1:IMSS	+
	Entamoeba invadens IP1	
Archaeplastida	*Cyanidioschyzon merolae strain 10D*	+
	Guillardia theta	
	Chlamydomonas reinhardtii	
	Ostreococcus tauri	+
	Physcomitrella patens subsp. patens	+
	Oryza sativa Japonica Group	+
	Ricinus communis	
	Arabidopsis thaliana	+
Chromaveolata	*Babesia bovis T2Bo*	+
	Cryptosporidium hominis TU502	
	Cryptosporidium muris RN66	
	Cryptosporidium parvum Iowa II	+
	Neospora caninum	+
	Plasmodium berghei	
	Plasmodium chabaudi	
	Plasmodium falciparum 3D7	+
	Plasmodium knowlesi strain H	
	Plasmodium vivax SaI-1	
	Plasmodium yoelii yoelii str. 17XNL	
	Phytophthora ramorum	+
	Tetrahymena thermophila SB210	+

	Thalassiosira pseudonana CCMP1335	+
	Theileria annulata strain Ankara	+
	Theileria parva strain Muguga	
	Toxoplasma gondii	+
Excavata	*Giardia intestinalis ATCC 50581*	
	Giardia lamblia P15	
	Giardia lamblia ATCC 50803	+
	Leishmania braziliensis MHOM/BR/75/M2904	
	Leishmania infantum JPCM5	
	Leishmania major strain Friedlin	+
	Leishmania mexicana	
	Trichomonas vaginalis G3	+
	Trypanosoma brucei gambiense	
	Trypanosoma brucei TREU927	+
	Trypanosoma congolense	
	Trypanosoma cruzi strain CL Brener	

Appendix Table 4.2 Marker genes used in different analyses

Analysis	Marker genes	Number of positions	Percentage of gaps and undetermined positions
Fungi[a]	*MCM2-7, MLH1-4, MSH1-6, SMC1-6, DMC1, RAD51, RPA1, RPB1, RPC1, eIF1A, eIF5B*	21,681 (AA)	11.37%
Animals[b]	*MCM2-9, MLH1-4, MSH2-6, SMC1-6, DMC1, RAD51, RPA1, RPB1, RPC1, eIF5B*	21,535 (AA)	15.50%
Mammals[b,c]	*MCM2-9, MLH1-4, MSH2-6, SMC1-6, SMC1beta, DMC1, RAD51, RAD51B-D, XRCC2-3, RPA1, RPB1, RPC1, eIF5B*	95,556 (CDS)	16.43%
Eukaryotes	*MCM2-7, MLH1,4, MSH2,6, SMC1-6, DMC1, RAD51, RPA1, RPB1, RPC1, eIF1A, eIF5B*	13,886 (AA)	10.12%

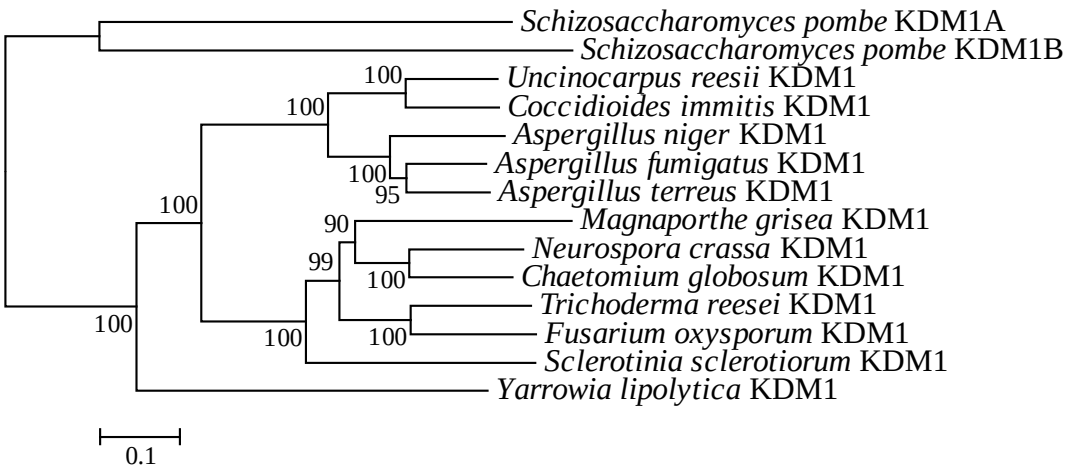
[a]The fungal dataset included one additional gene from the *MSH* gene family – *MSH1*.

[b]The *eIF1A* gene was excluded from animal and mammal analyses due to multiple duplications within vertebrates.

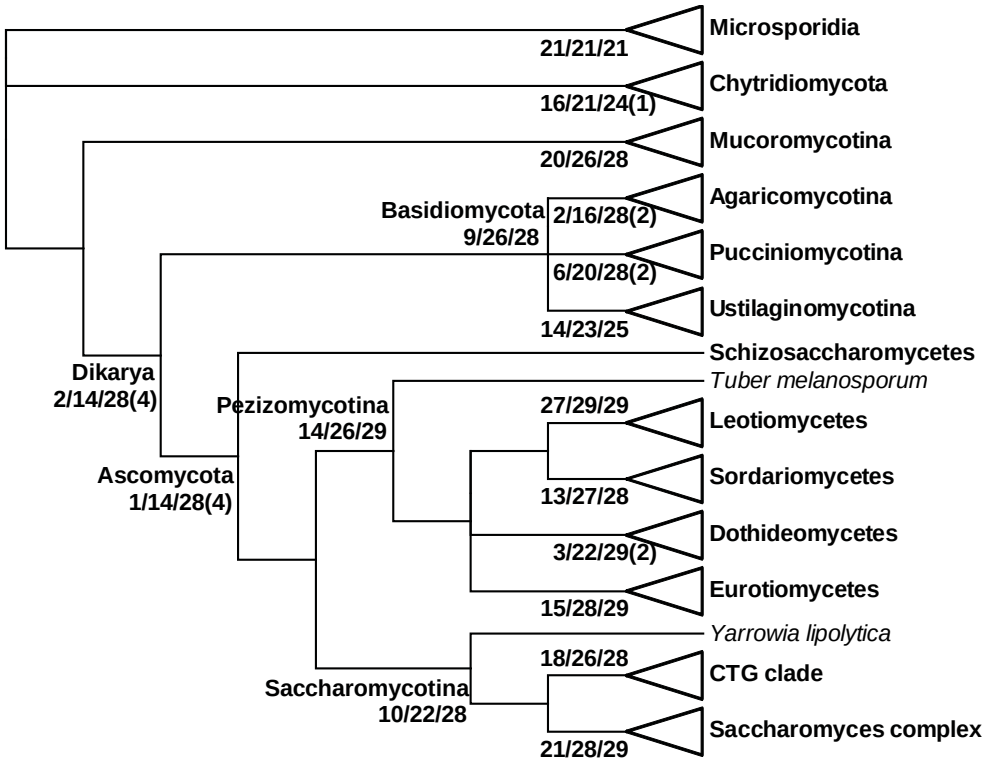
[c]The mammalian dataset included one additional genes from the *SMC* gene family – *SMC1beta*, and five additional genes from the *recA/RAD51* gene family – *RAD51B*, *RAD51C*, *RAD51D*, *XRCC2*, *XRCC3*.

Appendix Table 4.3 Statistical tests of alternative topologies of previous controversial relationships in mammals

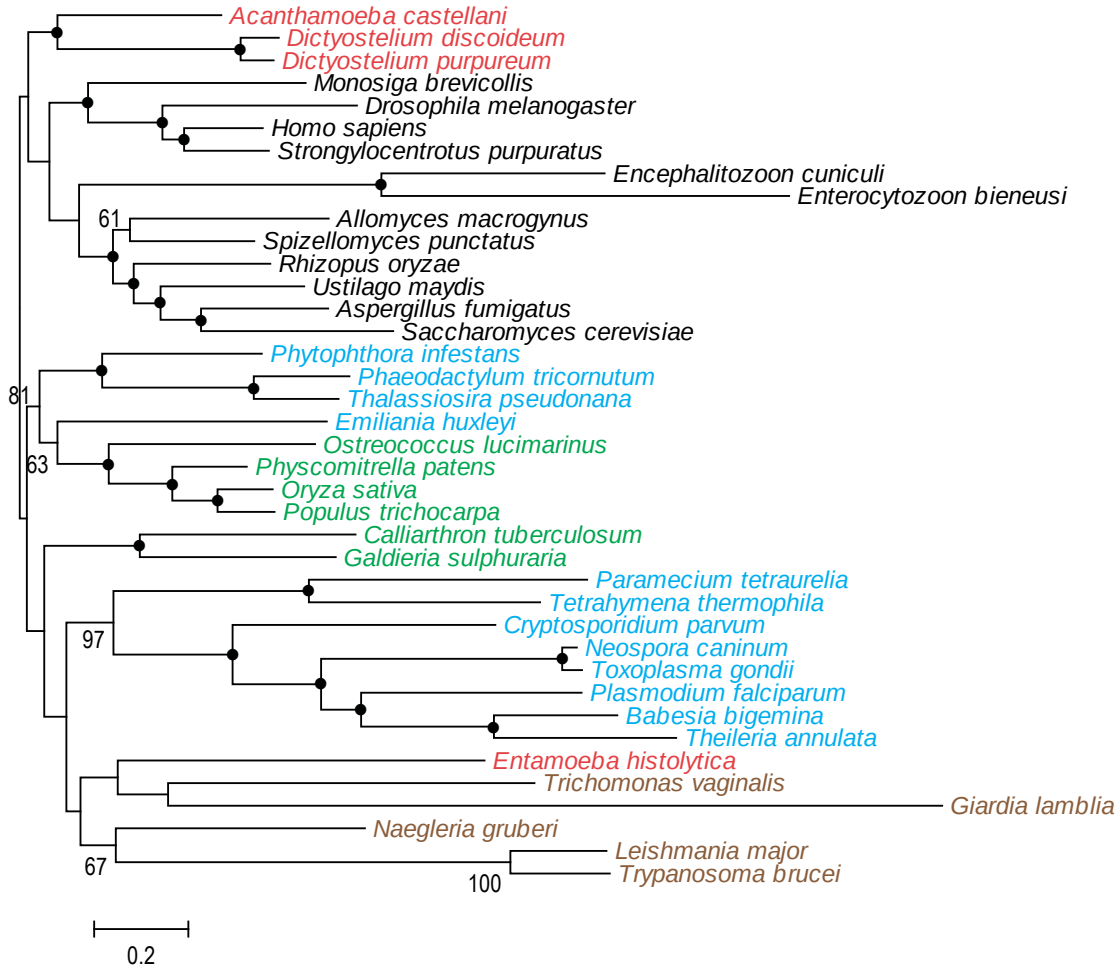
	AU			KH			SH			wKH			wSH		
	dataset1	dataset2	dataset3	dataset1	dataset2	dataset3	dataset1	dataset2	dataset3	dataset1	dataset2	dataset3	dataset1	dataset2	dataset3
	0.985	0.861	0.823	0.98	0.827	0.775	0.994	0.913	0.877	0.98	0.827	0.775	0.991	0.903	0.869
	0.019	0.202	0.268	0.02	0.173	0.225	0.021	0.232	0.317	0.02	0.173	0.225	0.037	0.278	0.353
	0.001	0.043	0.095	0.002	0.043	0.089	0.002	0.061	0.138	0.002	0.043	0.089	0.003	0.081	0.161
	0.716	0.789	0.743	0.711	0.741	0.712	0.835	0.846	0.828	0.711	0.741	0.712	0.835	0.84	0.825
	0.333	0.312	0.348	0.289	0.259	0.288	0.429	0.365	0.42	0.289	0.259	0.288	0.441	0.402	0.442
	0.329	0.174	0.251	0.286	0.161	0.224	0.424	0.243	0.345	0.286	0.161	0.224	0.436	0.278	0.366
	0.794	0.982	0.95	0.791	0.972	0.933	0.89	0.985	0.96	0.791	0.972	0.933	0.867	0.983	0.954
	6.00E-05	0.003	0.009	0.002	0.006	0.019	0.002	0.006	0.019	0.002	0.006	0.019	0.004	0.011	0.036
	0.207	0.025	0.064	0.209	0.028	0.067	0.249	0.028	0.067	0.209	0.028	0.067	0.316	0.05	0.114
	0.839	0.984	0.99	0.821	0.961	0.961	0.939	0.984	0.98	0.821	0.961	0.961	0.957	0.992	0.993
	0.205	0.029	0.042	0.168	0.027	0.039	0.49	0.097	0.171	0.168	0.027	0.039	0.442	0.078	0.095
	0.221	0.044	0.019	0.179	0.039	0.035	0.465	0.147	0.166	0.179	0.039	0.035	0.391	0.105	0.108
	8.00E-13	8.00E-06	0.006	7.00E-05	0.001	0.007	1.00E-04	0.001	0.008	4.00E-05	0.001	0.007	7.00E-05	0.005	0.02
	4.00E-05	0.025	2.20E-02	5.00E-04	0.028	0.023	0.001	0.032	0.026	2.00E-04	0.028	0.023	5.00E-04	0.06	0.052
	0.714	0.797	0.687	0.66	0.664	0.604	0.933	0.947	0.908	0.66	0.664	0.604	0.924	0.926	0.893
	0.147	0.207	0.388	0.145	0.101	0.278	0.507	0.474	0.741	0.145	0.101	0.278	0.504	0.385	0.743
	0.097	0.09	0.036	0.136	0.061	0.105	0.347	0.193	0.27	0.136	0.061	0.084	0.382	0.2	0.26
	0.428	0.146	0.326	0.34	0.092	0.274	0.639	0.248	0.511	0.34	0.092	0.274	0.668	0.266	0.575
	2.00E-04	0.407	0.47	2.00E-04	0.336	0.396	0.002	0.567	0.651	2.00E-04	0.336	0.396	0.001	0.601	0.666
	2.00E-73	0.001	0.003	0	0.011	0.025	9.00E-06	0.043	0.101	0	0.005	0.018	0	0.019	0.065
	2.00E-05	0.173	0.104	0.001	0.117	0.108	0.072	0.518	0.51	0.001	0.117	0.108	0.005	0.423	0.398



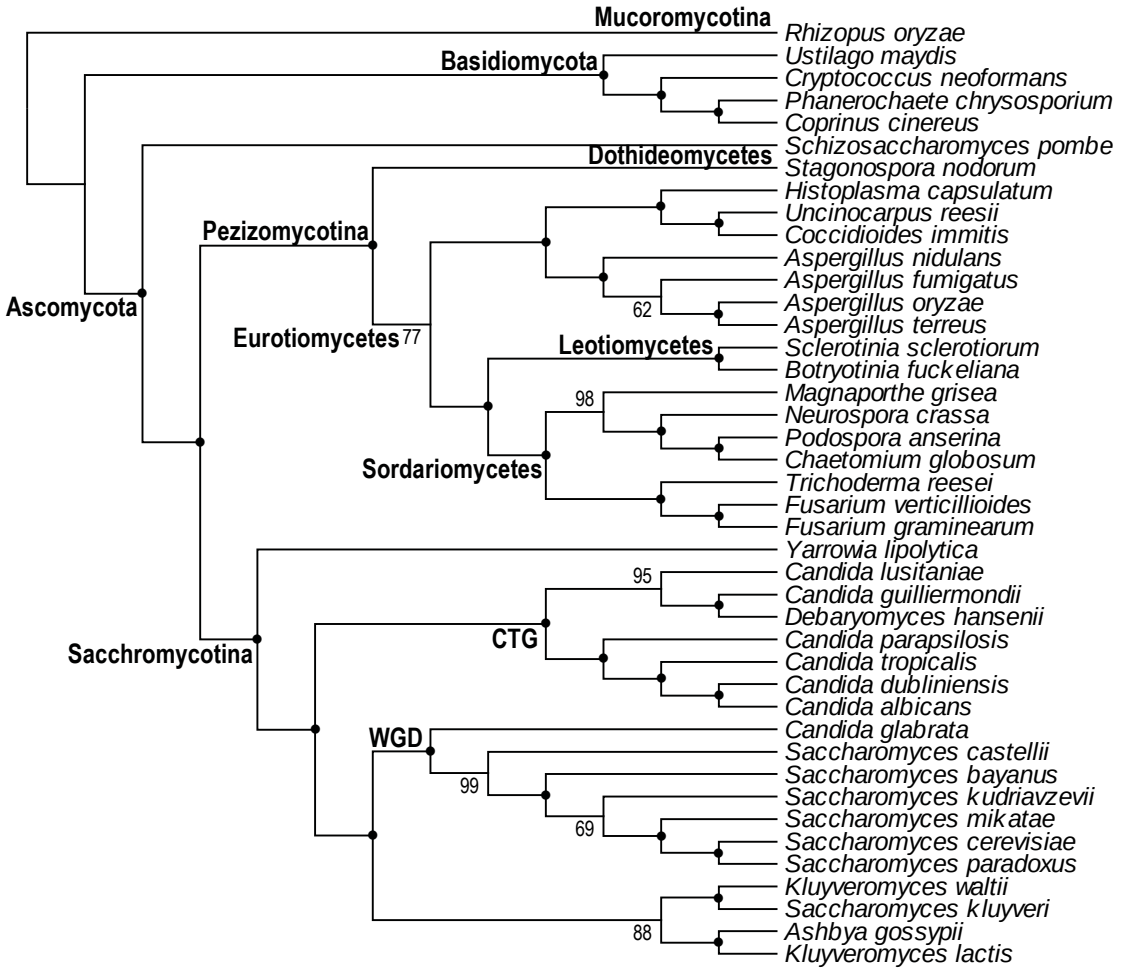
Appendix Figure 2.1 A NJ tree for fungal *KDM1* genes.



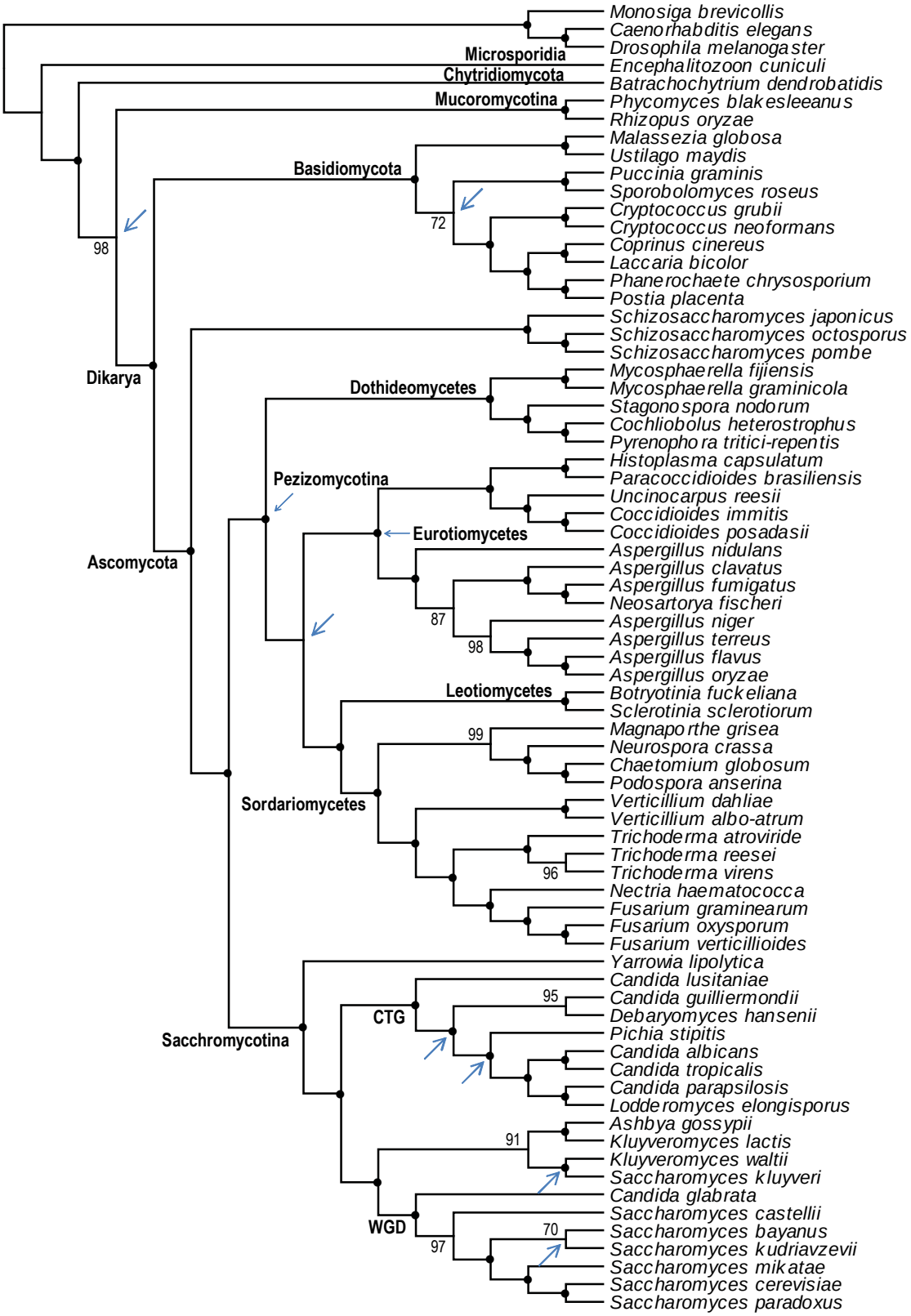
Appendix Figure 4.1 Cladogram of fungi showing the supports for major fungal clades from single gene phylogenies. The tree topology is adapted from previous studies [230, 288]. Uncertain relationships among basal linages and with Basidiomycota and Pezizomycotina are shown as unresolved. Numbers shown indicate number of single gene phylogenies supporting the corresponding nodes. The first number indicate number of single gene phylogenies with 100% support for the node; the second number indicate number of single gene phylogenies with at least 50% support for the node; the third number indicate total number of single gene phylogenies that are relevant to the node; the numbers in parentheses represent those gene families that have at least 70% support for alternative topology.



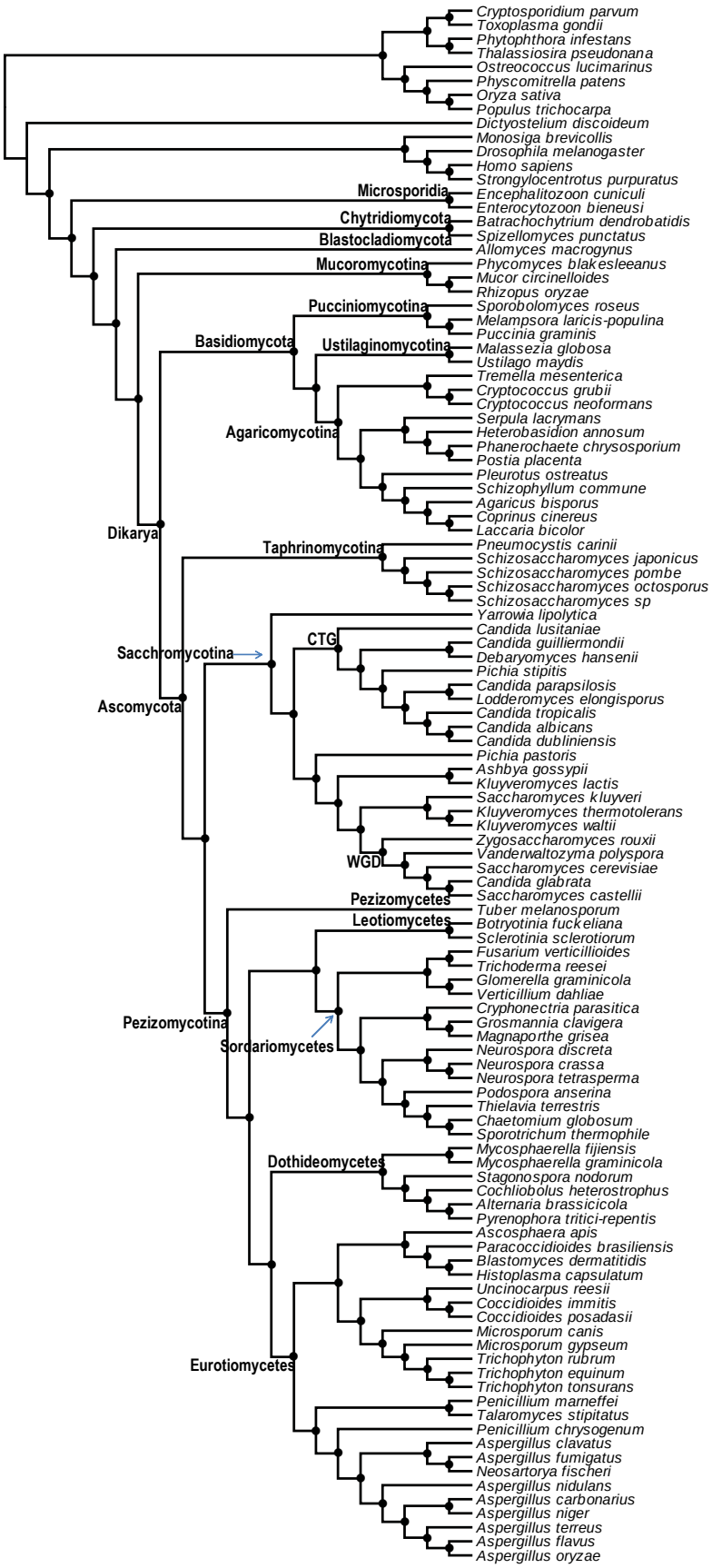
Appendix Figure 4.2 ML analysis of eukaryotic phylogeny. Support values were calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support. Only support values greater than 50% were shown. The five eukaryotic supergroups are coloured as described in Fig. 4.7.



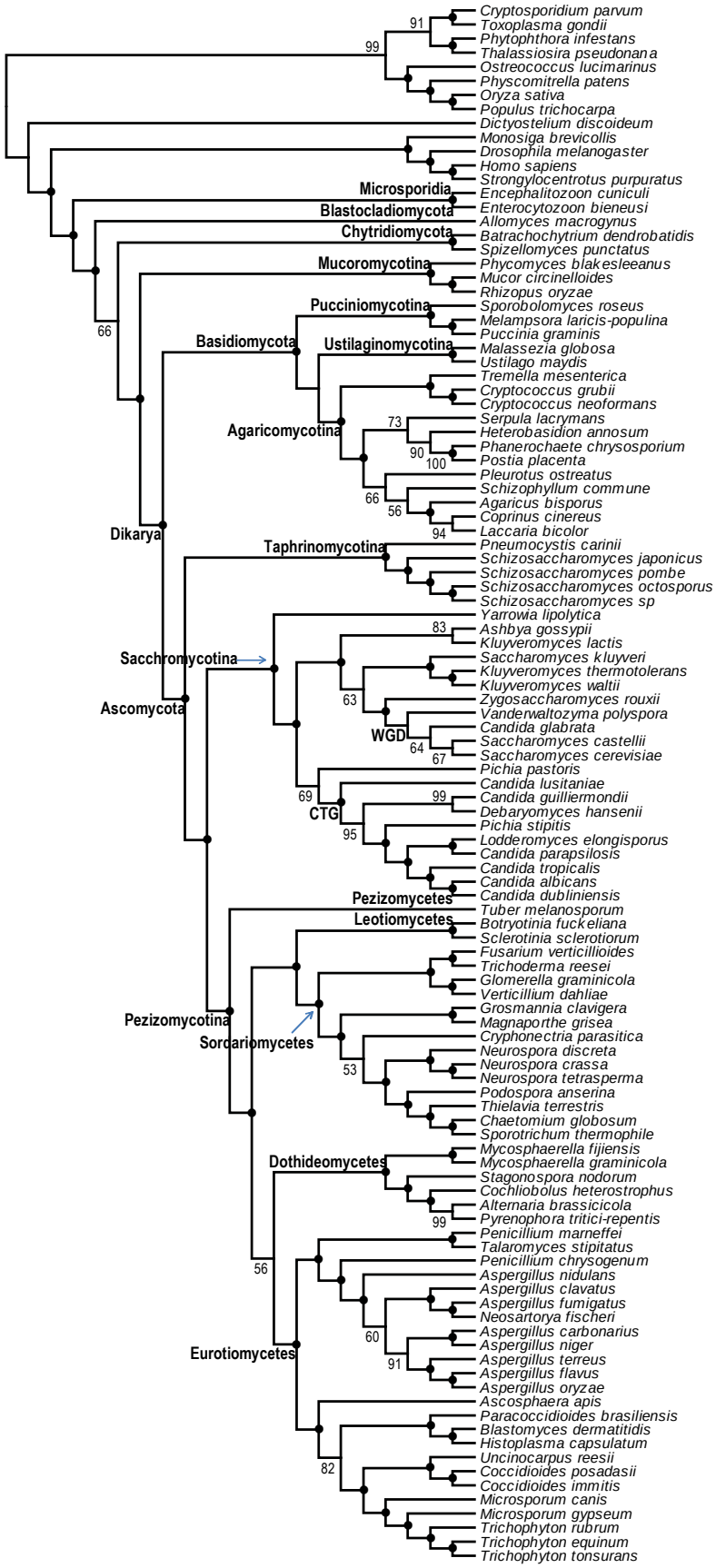
Appendix Figure 4.3 Cladogram of 42 fungal species that correspond to the 42 taxa analyzed by Fitzpatrick *et al* [228]. The topology was estimated by the ML approach with support values calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support.


Monosiga brevicollis
Caenorhabditis elegans
Drosophila melanogaster
Microsporidia
Encephalitozoon cuniculi
Chytridiomycota
Batrachochytrium dendrobatidis
Mucoromycotina
Phycomyces blakesleeanus
Rhizopus oryzae
Malassezia globosa
Ustilago maydis
Basidiomycota
Puccinia graminis
Sporobolomyces roseus
72
Cryptococcus grubii
Cryptococcus neoformans
98
Coprinus cinereus
Laccaria bicolor
Phanerochaete chrysosporium
Postia placenta
Schizosaccharomyces japonicus
Schizosaccharomyces octosporus
Schizosaccharomyces pombe
Dikarya
Mycosphaerella fijiensis
Dothideomycetes
Mycosphaerella graminicola
Stagonospora nodorum
Cochliobolus heterostrophus
Pyrenophora tritici-repentis
Histoplasma capsulatum
Pezizomycotina
Paracoccidioides brasiliensis
Uncinocarpus reesii
Coccidioides immitis
Eurotiomycetes
Coccidioides posadasii
Ascomycota
Aspergillus nidulans
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
87
Aspergillus niger
98
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Leotiomycetes
Botryotinia fuckeliana
Sclerotinia sclerotiorum
99
Magnaporthe grisea
Neurospora crassa
Chaetomium globosum
Podospora anserina
Sordariomycetes
Verticillium dahliae
Verticillium albo-atrum
Trichoderma atroviride
Trichoderma reesei
96
Trichoderma virens
Nectria haematococca
Fusarium graminearum
Fusarium oxysporum
Fusarium verticillioides
Yarrowia lipolytica
Candida lusitaniae
95
Candida guilliermondii
CTG
Debaryomyces hansenii
Pichia stipitis
Sacchromycotina
Candida albicans
Candida tropicalis
Candida parapsilosis
Lodderomyces elongisporus
Ashbya gossypii
91
Kluyveromyces lactis
Kluyveromyces waltii
Saccharomyces kluyveri
Candida glabrata
Saccharomyces castellii
WGD
70
Saccharomyces bayanus
97
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus

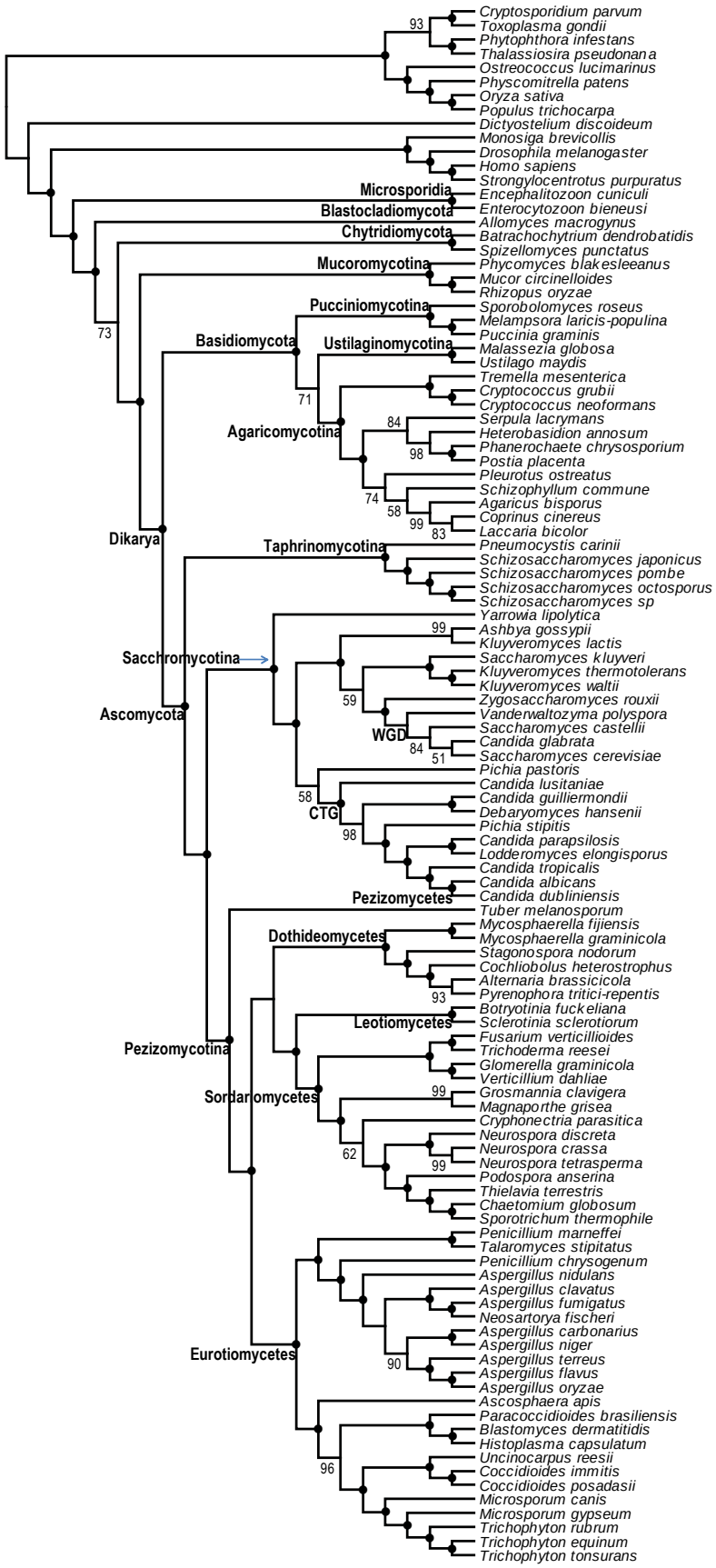
Appendix Figure 4.4 Cladogram of 70 fungal species that correspond to the 82 taxa analyzed by Wang et al [229]. The topology was estimated by the ML approach with support values calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support. Only support values greater than 50% were shown. Nodes different from previous study [229] are highlighted by thick arrows.


Cryptosporidium parvum
Toxoplasma gondii
Phytophthora infestans
Thalassiosira pseudonana
Ostreococcus lucimarinus
Physcomitrella patens
Oryza sativa
Populus trichocarpa
Dictyostelium discoideum
Monosiga brevicollis
Drosophila melanogaster
Homo sapiens
Strongylocentrotus purpuratus
Microsporidia
Encephalitozoon cuniculi
Enterocytozoon bieneusi
Chytridiomycota
Batrachochytrium dendrobatidis
Blastocladiomycota
Spizellomyces punctatus
Allomyces macrogynus
Mucoromycotina
Phycomyces blakesleeanus
Mucor circinelloides
Rhizopus oryzae
Pucciniomycotina
Sporobolomyces roseus
Melampsora laricis-populina
Puccinia graminis
Basidiomycota
Ustilaginomycotina
Malassezia globosa
Ustilago maydis
Tremella mesenterica
Cryptococcus grubii
Cryptococcus neoformans
Agaricomycotina
Serpula lacrymans
Heterobasidion annosum
Phanerochaete chrysosporium
Postia placenta
Pleurotus ostreatus
Schizophyllum commune
Agaricus bisporus
Coprinus cinereus
Laccaria bicolor
Dikarya
Taphrinomycotina
Pneumocystis carinii
Schizosaccharomyces japonicus
Schizosaccharomyces pombe
Schizosaccharomyces octosporus
Schizosaccharomyces sp
Yarrowia lipolytica
CTG
Candida lusitaniae
Candida guilliermondii
Saccharomycotina
Debaryomyces hansenii
Pichia stipitis
Candida parapsilosis
Lodderomyces elongisporus
Ascomycota
Candida tropicalis
Candida albicans
Candida dubliniensis
Pichia pastoris
Ashbya gossypii
Kluyveromyces lactis
Saccharomyces kluyveri
Kluyveromyces thermotolerans
Kluyveromyces waltii
Zygosaccharomyces rouxii
WGD
Vanderwaltozyma polyspora
Saccharomyces cerevisiae
Candida glabrata
Saccharomyces castellii
Pezizomycetes
Tuber melanosporum
Leotiomycetes
Botryotinia fuckeliana
Sclerotinia sclerotiorum
Fusarium verticillioides
Trichoderma reesei
Glomerella graminicola
Verticillium dahliae
Cryphonectria parasitica
Pezizomycotina
Grosmannia clavigera
Magnaporthe grisea
Sordariomycetes
Neurospora discreta
Neurospora crassa
Neurospora tetrasperma
Podospora anserina
Thielavia terrestris
Chaetomium globosum
Sporotrichum thermophile
Mycosphaerella fijiensis
Dothideomycetes
Mycosphaerella graminicola
Stagonospora nodorum
Cochliobolus heterostrophus
Alternaria brassicicola
Pyrenophora tritici-repentis
Ascosphaera apis
Paracoccidioides brasiliensis
Blastomyces dermatitidis
Histoplasma capsulatum
Uncinocarpus reesii
Coccidioides immitis
Coccidioides posadasii
Microsporum canis
Microsporum gypseum
Eurotiomycetes
Trichophyton rubrum
Trichophyton equinum
Trichophyton tonsurans
Penicillium marneffei
Talaromyces stipitatus
Penicillium chrysogenum
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
Aspergillus nidulans
Aspergillus carbonarius
Aspergillus niger
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae

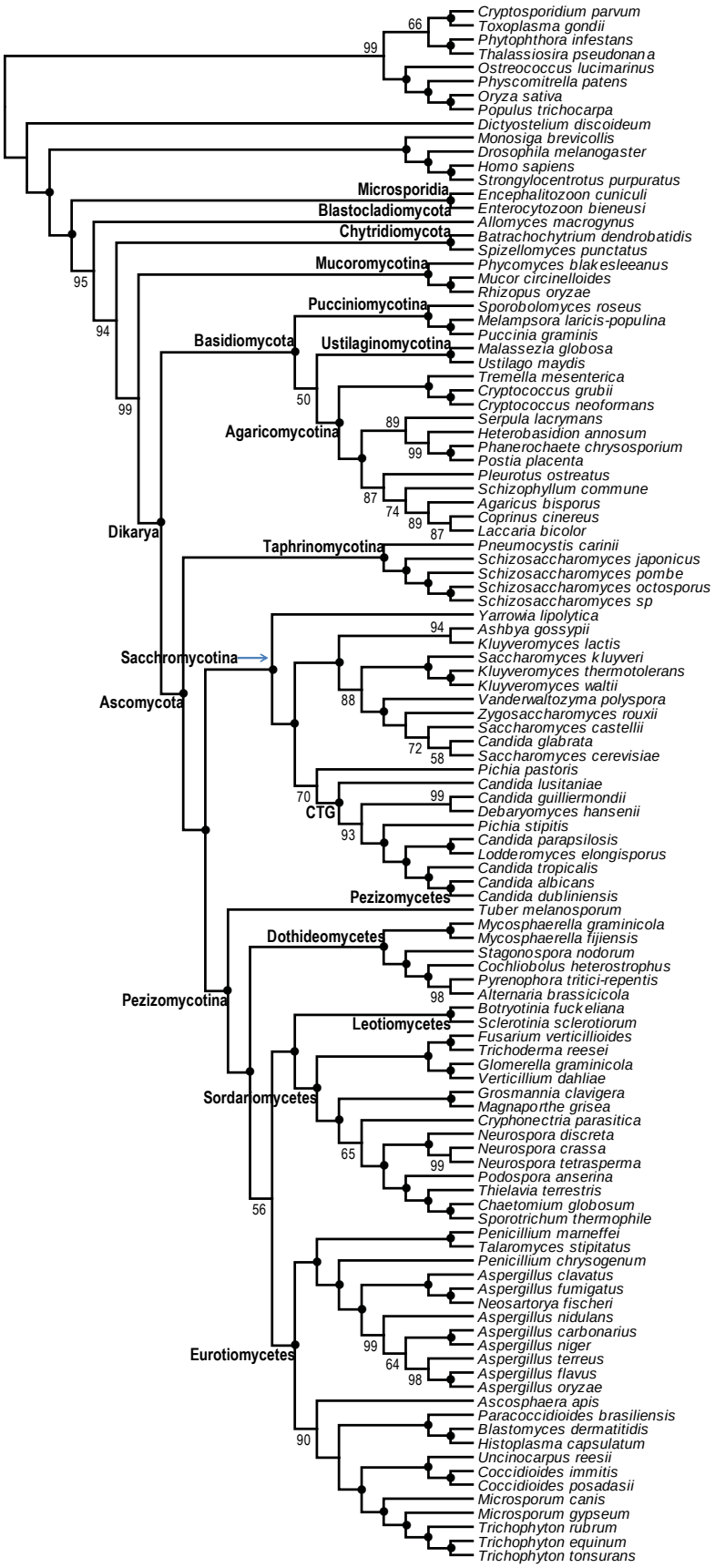
Appendix Figure 4.5 Cladogram of 98 fungal species using 29 genes (topology estimated by Bayesian approach). Black dots indicate 100% support from both Posterior Probability (PP). This analysis included all 29 marker genes that were used for Fig. 4.5.


Cryptosporidium parvum
Toxoplasma gondii
Phytophthora infestans
Thalassiosira pseudonana
Ostreococcus lucimarinus
Physcomitrella patens
Oryza sativa
Populus trichocarpa
Dictyostelium discoideum
Monosiga brevicollis
Drosophila melanogaster
Homo sapiens
Strongylocentrotus purpuratus
Microsporidia
Encephalitozoon cuniculi
Enterocytozoon bieneusi
Blastocladiomycota
Allomyces macrogynus
Chytridiomycota
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Mucoromycotina
Phycomyces blakesleeanus
Mucor circinelloides
Rhizopus oryzae
Pucciniomycotina
Sporobolomyces roseus
Melampsora laricis-populina
Puccinia graminis
Basidiomycota
Ustilaginomycotina
Malassezia globosa
Ustilago maydis
Tremella mesenterica
Cryptococcus grubii
Cryptococcus neoformans
Agaricomycotina
Serpula lacrymans
Heterobasidion annosum
Phanerochaete chrysosporium
Postia placenta
Pleurotus ostreatus
Schizophyllum commune
Agaricus bisporus
Coprinus cinereus
Laccaria bicolor
Dikarya
Taphrinomycotina
Pneumocystis carinii
Schizosaccharomyces japonicus
Schizosaccharomyces pombe
Schizosaccharomyces octosporus
Schizosaccharomyces sp
Yarrowia lipolytica
Ashbya gossypii
Kluyveromyces lactis
Saccharomycotina
Saccharomyces kluyveri
Kluyveromyces thermotolerans
Kluyveromyces waltii
Zygosaccharomyces rouxii
Vanderwaltozyma polyspora
Ascomycota
WGD
Candida glabrata
Saccharomyces castellii
Saccharomyces cerevisiae
Pichia pastoris
Candida lusitaniae
Candida guilliermondii
CTG
Debaryomyces hansenii
Pichia stipitis
Lodderomyces elongisporus
Candida parapsilosis
Candida tropicalis
Candida albicans
Candida dubliniensis
Pezizomycetes
Tuber melanosporum
Leotiomycetes
Botryotinia fuckeliana
Sclerotinia sclerotiorum
Fusarium verticillioides
Trichoderma reesei
Glomerella graminicola
Verticillium dahliae
Pezizomycotina
Grosmannia clavigera
Magnaporthe grisea
Sordariomycetes
Cryphonectria parasitica
Neurospora discreta
Neurospora crassa
Neurospora tetrasperma
Podospora anserina
Thielavia terrestris
Chaetomium globosum
Sporotrichum thermophile
Mycosphaerella fijiensis
Dothideomycetes
Mycosphaerella graminicola
Stagonospora nodorum
Cochliobolus heterostrophus
Alternaria brassicicola
Pyrenophora tritici-repentis
Penicillium marneffei
Talaromyces stipitatus
Penicillium chrysogenum
Aspergillus nidulans
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
Aspergillus carbonarius
Eurotiomycetes
Aspergillus niger
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Ascosphaera apis
Paracoccidioides brasiliensis
Blastomyces dermatitidis
Histoplasma capsulatum
Uncinocarpus reesii
Coccidioides posadasii
Coccidioides immitis
Microsporum canis
Microsporum gypseum
Trichophyton rubrum
Trichophyton equinum
Trichophyton tonsurans

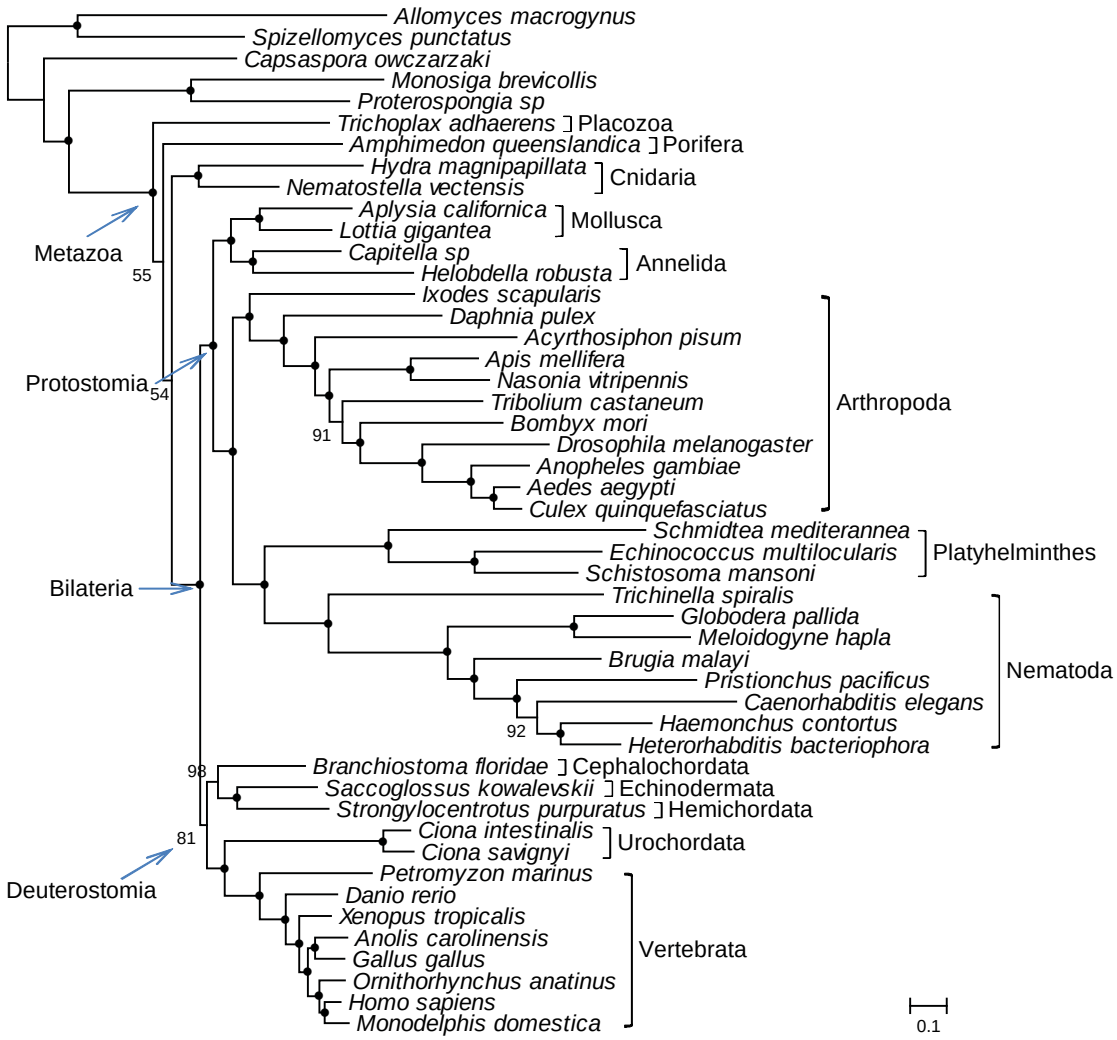
Appendix Figure 4.6 Cladogram of 98 fungal species using 29 genes (topology estimated by Maximum Likelihood approach). Black dots indicate 100% support from bootstrap test of 100 replicates. This analysis included all 29 marker genes that were used for Fig. 4.5.


Cryptosporidium parvum
Toxoplasma gondii
Phytophthora infestans
Thalassiosira pseudonana
Ostreococcus lucimarinus
Physcomitrella patens
Oryza sativa
Populus trichocarpa
Dictyostelium discoideum
Monosiga brevicollis
Drosophila melanogaster
Homo sapiens
Strongylocentrotus purpuratus
Microsporidia
Encephalitozoon cuniculi
Enterocytozoon bieneusi
Blastocladiomycota
Allomyces macrogynus
Chytridiomycota
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Mucoromycotina
Phycomyces blakesleeanus
Mucor circinelloides
Rhizopus oryzae
Pucciniomycotina
Sporobolomyces roseus
Melampsora laricis-populina
Puccinia graminis
Basidiomycota
Ustilaginomycotina
Malassezia globosa
Ustilago maydis
Tremella mesenterica
Cryptococcus grubii
Cryptococcus neoformans
Agaricomycotina
Serpula lacrymans
Heterobasidion annosum
Phanerochaete chrysosporium
Postia placenta
Pleurotus ostreatus
Schizophyllum commune
Agaricus bisporus
Coprinus cinereus
Laccaria bicolor
Dikarya
Taphrinomycotina
Pneumocystis carinii
Schizosaccharomyces japonicus
Schizosaccharomyces pombe
Schizosaccharomyces octosporus
Schizosaccharomyces sp
Saccharomycotina
Yarrowia lipolytica
Ashbya gossypii
Kluyveromyces lactis
Saccharomyces kluyveri
Kluyveromyces thermotolerans
Kluyveromyces waltii
Zygosaccharomyces rouxii
Vanderwaltozyma polyspora
WGD
Saccharomyces castellii
Candida glabrata
Saccharomyces cerevisiae
Ascomycota
Pichia pastoris
Candida lusitaniae
Candida guilliermondii
Debaryomyces hansenii
CTG
Pichia stipitis
Candida parapsilosis
Lodderomyces elongisporus
Candida tropicalis
Candida albicans
Candida dubliniensis
Pezizomycetes
Tuber melanosporum
Mycosphaerella fijiensis
Dothideomycetes
Mycosphaerella graminicola
Stagonospora nodorum
Cochliobolus heterostrophus
Alternaria brassicicola
Pyrenophora tritici-repentis
Botryotinia fuckeliana
Leotiomycetes
Sclerotinia sclerotiorum
Fusarium verticillioides
Pezizomycotina
Trichoderma reesei
Glomerella graminicola
Verticillium dahliae
Sordariomycetes
Grosmannia clavigera
Magnaporthe grisea
Cryphonectria parasitica
Neurospora discreta
Neurospora crassa
Neurospora tetrasperma
Podospora anserina
Thielavia terrestris
Chaetomium globosum
Sporotrichum thermophile
Penicillium marneffei
Talaromyces stipitatus
Penicillium chrysogenum
Aspergillus nidulans
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
Aspergillus carbonarius
Eurotiomycetes
Aspergillus niger
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Ascosphaera apis
Paracoccidioides brasiliensis
Blastomyces dermatitidis
Histoplasma capsulatum
Uncinocarpus reesii
Coccidioides immitis
Coccidioides posadasii
Microsporum canis
Microsporum gypseum
Trichophyton rubrum
Trichophyton equinum
Trichophyton tonsurans

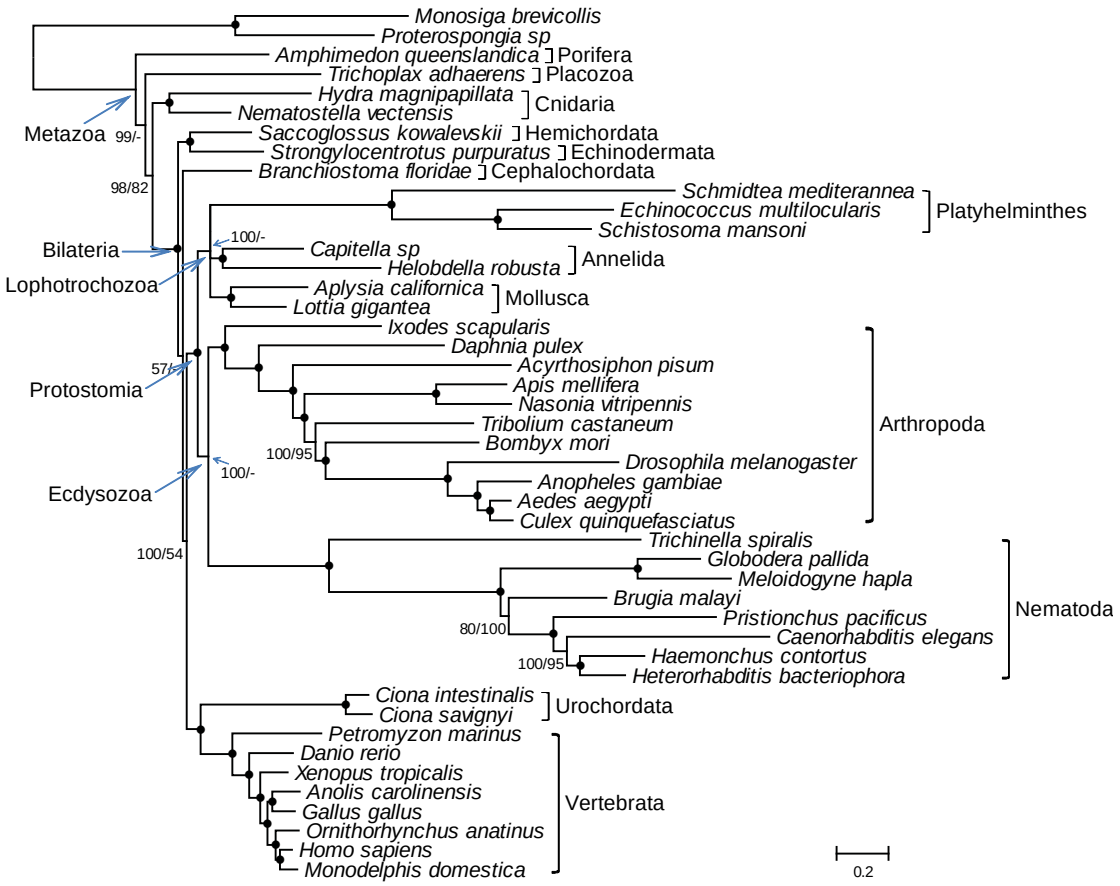
Appendix Figure 4.7 Cladogram of 98 fungal species using 24 genes. The topology was estimated by the Maximum Likelihood method with support values calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support. Only support values greater than 50% were shown. This analysis included 24 marker genes (*SMC1-6, MSH1-3, MSH6, MLH1, MLH4, RAD51, MCM2-7, RPA1, RPB1, RPC1, eIF1A* and *eIF5B*) which is a subset of the genes used for Fig. 4.5.


Cryptosporidium parvum
Toxoplasma gondii
Phytophthora infestans
Thalassiosira pseudonana
Ostreococcus lucimarinus
Physcomitrella patens
Oryza sativa
Populus trichocarpa
Dictyostelium discoideum
Monosiga brevicollis
Drosophila melanogaster
Homo sapiens
Strongylocentrotus purpuratus
Microsporidia
Encephalitozoon cuniculi
Enterocytozoon bieneusi
Blastocladiomycota
Allomyces macrogynus
Chytridiomycota
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Mucoromycotina
Phycomyces blakesleeanus
Mucor circinelloides
Rhizopus oryzae
Pucciniomycotina
Sporobolomyces roseus
Melampsora laricis-populina
Puccinia graminis
Basidiomycota
Ustilaginomycotina
Malassezia globosa
Ustilago maydis
Tremella mesenterica
Cryptococcus grubii
Cryptococcus neoformans
Agaricomycotina
Serpula lacrymans
Heterobasidion annosum
Phanerochaete chrysosporium
Postia placenta
Pleurotus ostreatus
Schizophyllum commune
Agaricus bisporus
Coprinus cinereus
Laccaria bicolor
Dikarya
Taphrinomycotina
Pneumocystis carinii
Schizosaccharomyces japonicus
Schizosaccharomyces pombe
Schizosaccharomyces octosporus
Schizosaccharomyces sp
Yarrowia lipolytica
Ashbya gossypii
Kluyveromyces lactis
Saccharomycotina
Saccharomyces kluyveri
Kluyveromyces thermotolerans
Kluyveromyces waltii
Ascomycota
Vanderwaltozyma polyspora
Zygosaccharomyces rouxii
Saccharomyces castellii
Candida glabrata
Saccharomyces cerevisiae
Pichia pastoris
Candida lusitaniae
Candida guilliermondii
CTG
Debaryomyces hansenii
Pichia stipitis
Candida parapsilosis
Lodderomyces elongisporus
Candida tropicalis
Candida albicans
Candida dubliniensis
Pezizomycetes
Tuber melanosporum
Mycosphaerella graminicola
Dothideomycetes
Mycosphaerella fijiensis
Stagonospora nodorum
Cochliobolus heterostrophus
Pyrenophora tritici-repentis
Pezizomycotina
Alternaria brassicicola
Botryotinia fuckeliana
Leotiomycetes
Sclerotinia sclerotiorum
Fusarium verticillioides
Trichoderma reesei
Glomerella graminicola
Verticillium dahliae
Sordariomycetes
Grosmannia clavigera
Magnaporthe grisea
Cryphonectria parasitica
Neurospora discreta
Neurospora crassa
Neurospora tetrasperma
Podospora anserina
Thielavia terrestris
Chaetomium globosum
Sporotrichum thermophile
Penicillium marneffei
Talaromyces stipitatus
Penicillium chrysogenum
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
Aspergillus nidulans
Aspergillus carbonarius
Eurotiomycetes
Aspergillus niger
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Ascosphaera apis
Paracoccidioides brasiliensis
Blastomyces dermatitidis
Histoplasma capsulatum
Uncinocarpus reesii
Coccidioides immitis
Coccidioides posadasii
Microsporum canis
Microsporum gypseum
Trichophyton rubrum
Trichophyton equinum
Trichophyton tonsurans

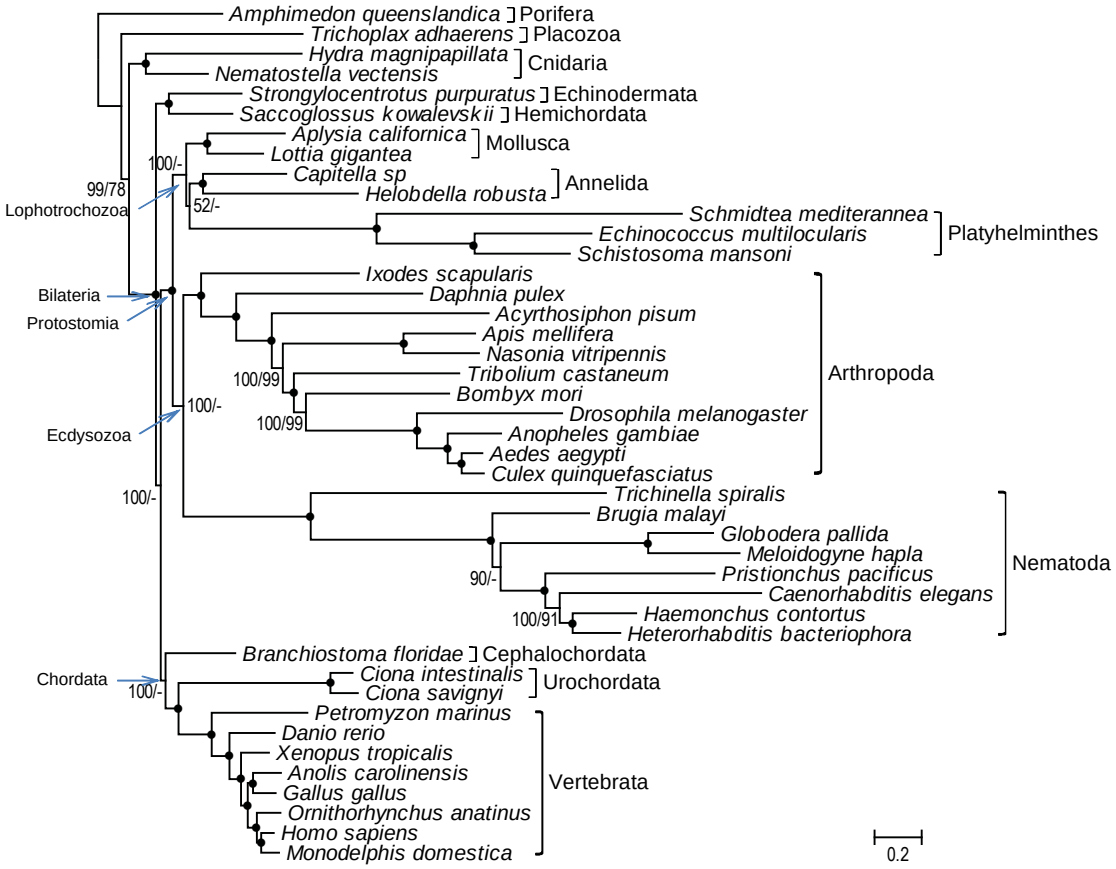
Appendix Figure 4.8 Cladogram of 98 fungal species using 19 genes. The topology was estimated by the Maximum Likelihood method with support values calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support. Only support values greater than 50% were shown. This analysis included 19 marker genes (*SMC1-6*, *MSH1-3*, *MSH6*, *MLH1*, *MLH4*, *RAD51*, *MCM2-7*) which is a subset of the genes used for Fig. 4.5.



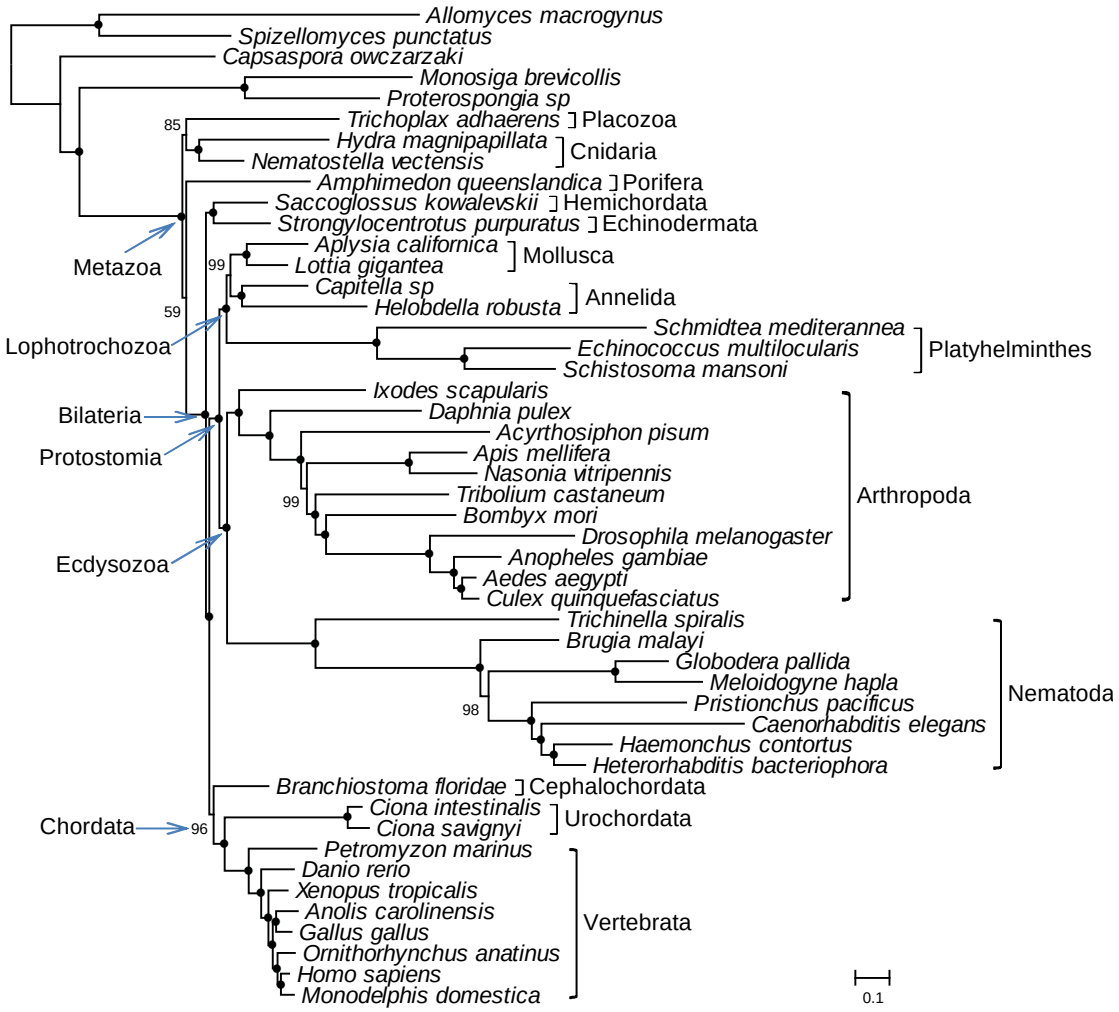
Appendix Figure 4.9 ML analysis of 43 animal and five outgroup species. Support values were calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap support.



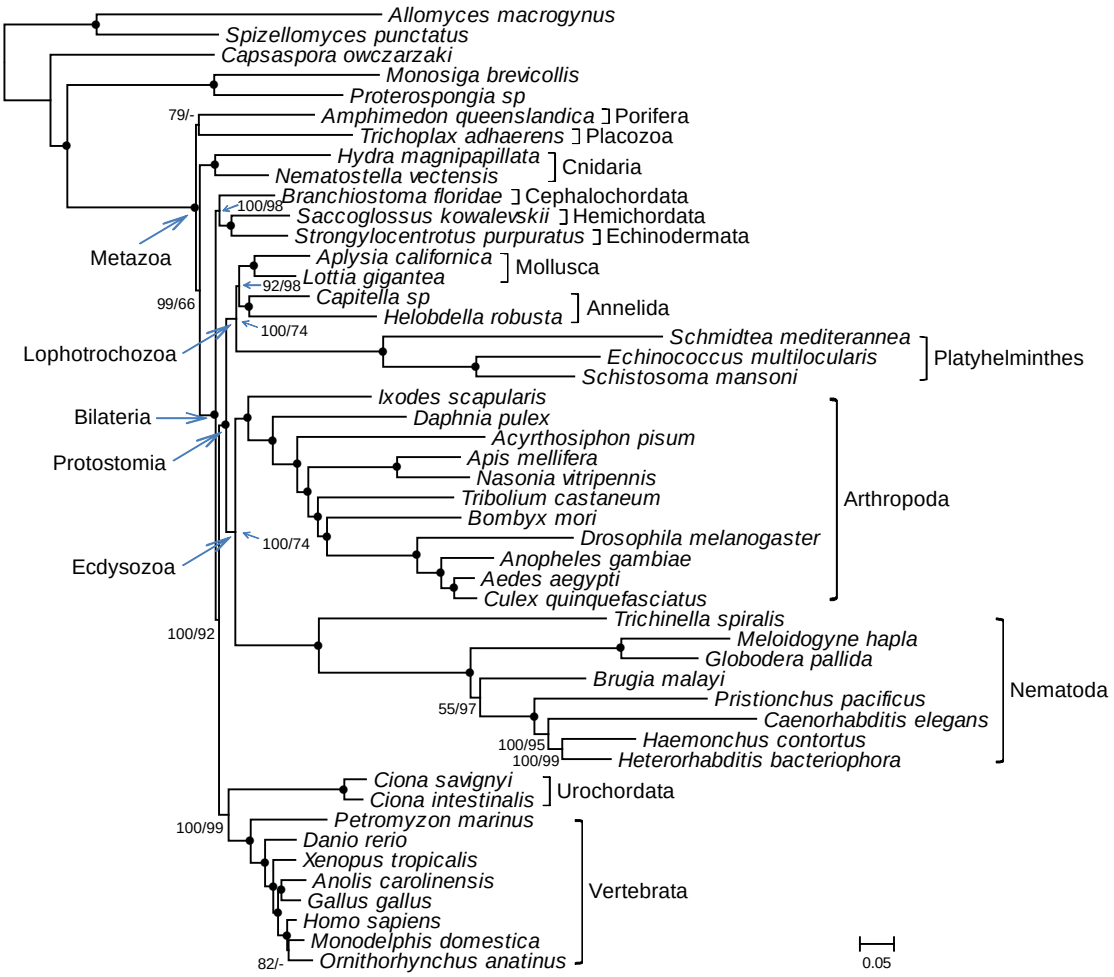
Appendix Figure 4.10 Phylogenetic analysis of 43 animal species by using closely related protists (*Monosiga* and *Proterospongia*) as outgroups. The tree was constructed by Phylobayes using the CAT model. Black dots indicate 100% support from both Posterior probability (PP) and bootstrap (BS). Support values from PP/BS are shown for nodes that do not receive 100% support. Dashes indicate <50 % support from PP or BS, or inconsistent topology between Bayesian and ML methods.



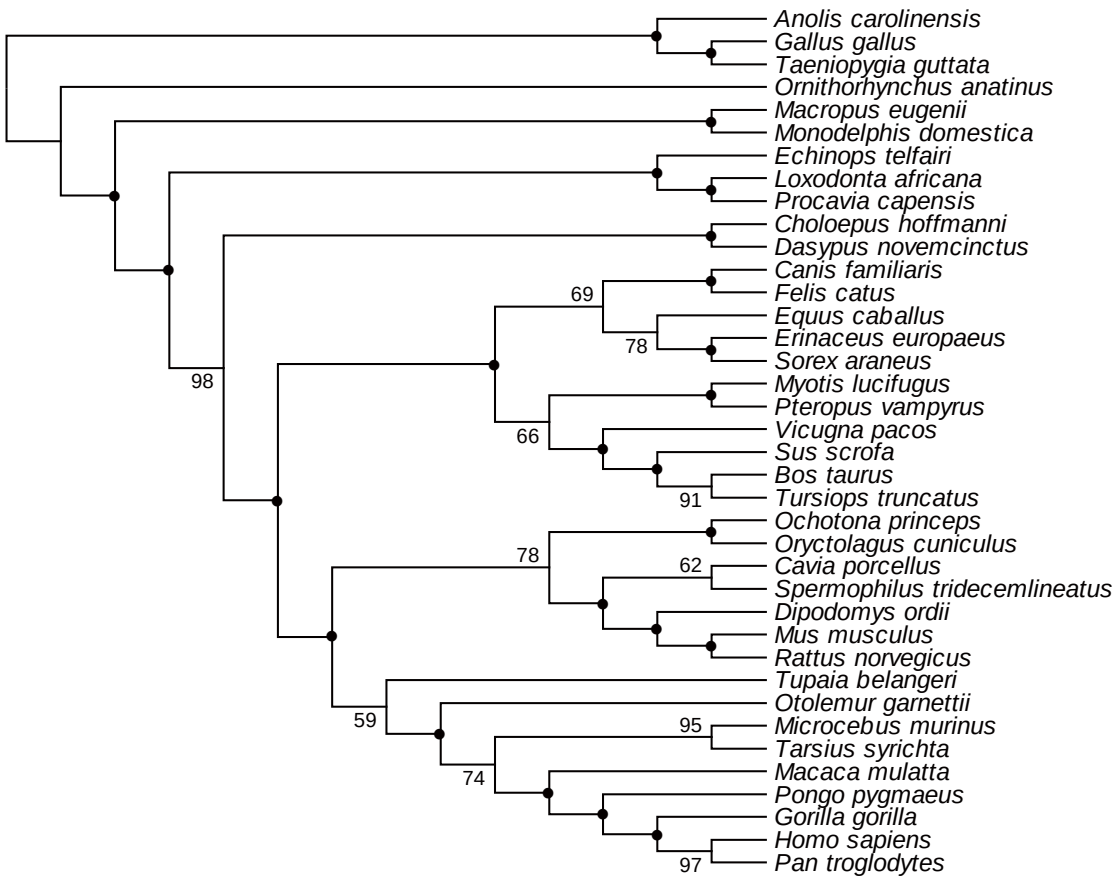
Appendix Figure 4.11 Phylogenetic analysis of 43 animal species. The shown topology was rooted with Porifera. The tree was constructed by Phylobayes using the CAT model. Black dots indicate 100% support from both Posterior probability (PP) and bootstrap (BS). Support values from PP/BS are shown for nodes that do not receive 100% support. Dashes indicate <50 % support from PP or BS, or inconsistent topology between Bayesian and ML methods.



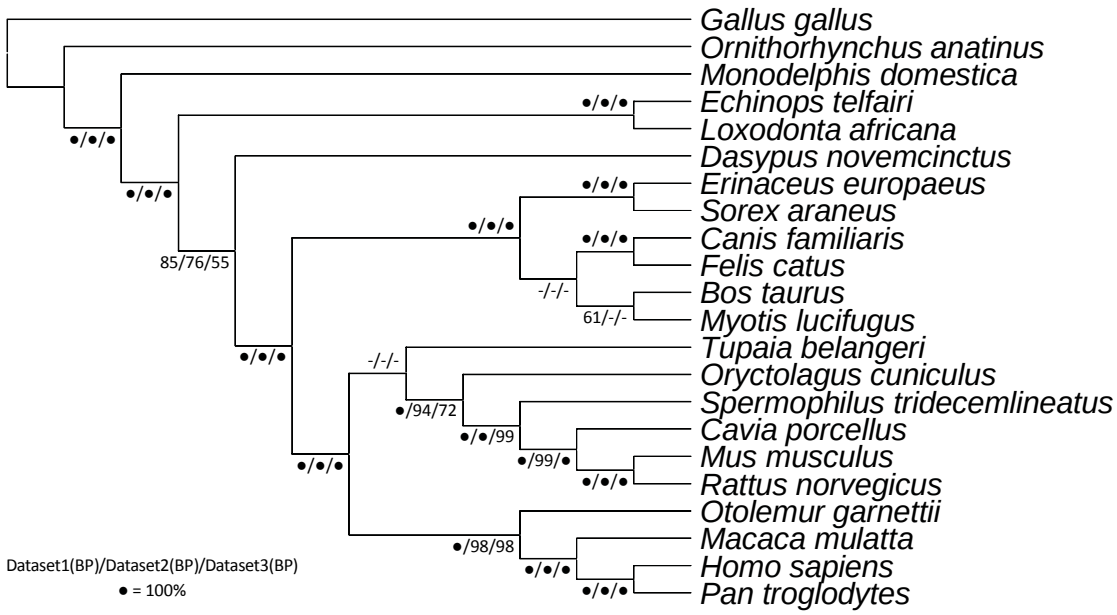
Appendix Figure 4.12 Bayesian analysis of animal phylogeny with amino acid recoded into functional categories according to six Dayhoff groups. Black dots indicate 100% posterior probability.



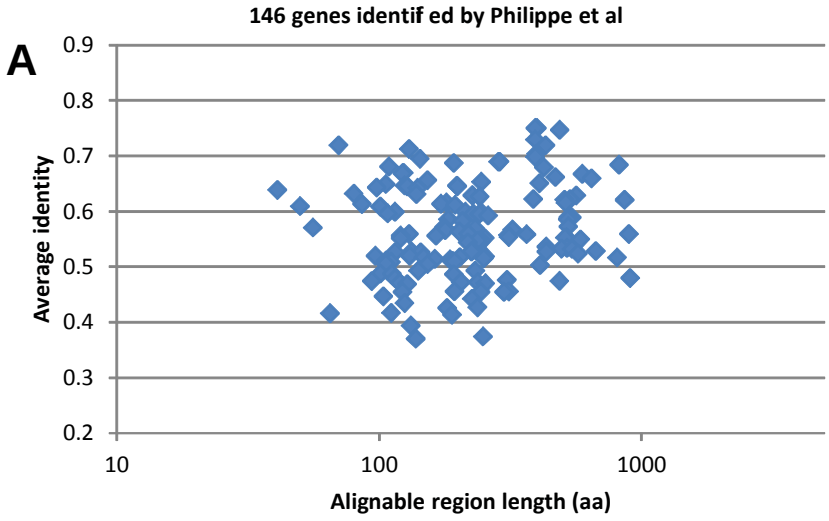
Appendix Figure 4.13 Phylogenetic analysis of animal phylogeny after removing fast-evolving sites. The dataset used in this analysis was created by removing the top three categories of fast-evolving sites from the dataset used for Fig. 4.6. The tree was constructed by Phylobayes using the CAT model. Black dots indicate 100% support from both Posterior probability (PP) and bootstrap (BS). Support values from PP/BS are shown for nodes that do not receive 100% support. Dashes indicate <50 % support from PP or BS, or inconsistent topology between Bayesian and ML methods.

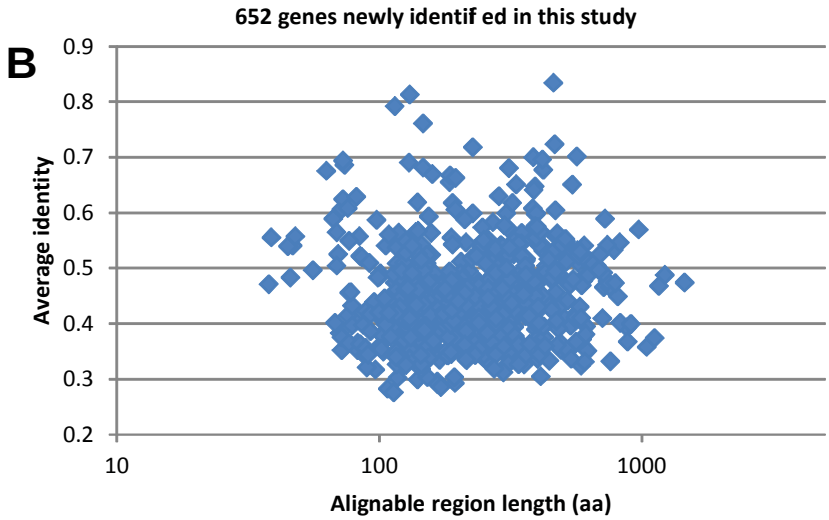


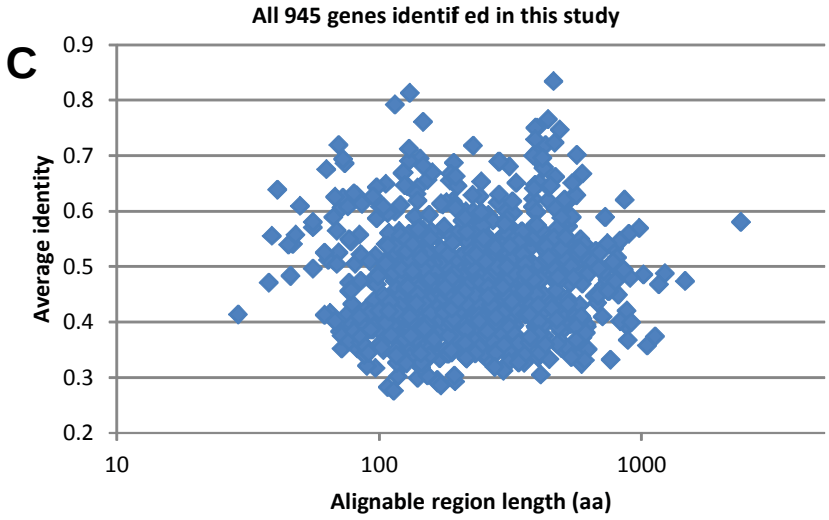
Appendix Figure 4.14 Cladogram of 35 mammalian species with green anole and birds as outgroup. The topology shown was revealed by the ML analysis of the protein dataset. Black dots indicate 100% bootstrap (BS) support. Only support values greater than 50% were shown.



Appendix Figure 4.15 Cladogram of 21 mammalian species with chicken as outgroup. The topology shown was revealed by the ML analysis of the complete dataset. The 22 species are the same set of species analyzed in Hallstrom *et al* [259] . Support values were calculated from bootstrap test of 100 replicates. Black dots indicate 100% bootstrap (BS) support and dashes indicate <50 % BS support or inconsistent topology between datasets.


146 genes identif ed by Philippe et al
A
Average identity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10
100
1000
Alignable region length (aa)


652 genes newly identif ed in this study
B
Average identity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10
100
1000
Alignable region length (aa)


All 945 genes identif ed in this study
C
Average identity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10
100
1000
Alignable region length (aa)

Appendix Figure 4.16 Distribution of sequence properties (alignable region length and average identity) of marker genes identified in this study. (A) The 146 genes identified by Philippe *et al* [89]. (B) The 652 genes newly identified in this study. (C) All the 945 genes identified in this study.

BIBLIOGRAPHY

1. Ohno, S., 1970, Evolution by Gene DuplicationBerlin-Heidelberg-NY: Springer-Verlag.
2. Tischler, G., 1915. Chromosomenzahl, Form und Individualitä in Planzenreiche. Progr Rei Bot, (5).
3. Bridges, C.B., 1936. The Bar "Gene" a Duplication. Science, **83**(2148): 210-211.
4. Taylor, J.S. and J. Raes, 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet, **38**: 615-643.
5. Stuber, C.W. and M.M. Goodman, 1983. Inheritance, intracellular localization, and genetic variation of phosphoglucomutase isozymes in maize (Zea mays L.). Biochem Genet, **21**(7-8): 667-689.
6. Ferris, S.D. and G.S. Whitt, 1979. Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol, **12**(4): 267-317.
7. Zhang, J., 2003. Evolution by gene duplication: an update. Trends Ecol Evol, **18**(6): 292-298.
8. Lynch, M. and J.S. Conery, 2000. The evolutionary fate and consequences of duplicate genes. Science, **290**(5494): 1151-1155.
9. Innan, H. and F. Kondrashov, 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet, **11**(2): 97-108.
10. Rokas, A., 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. Annu Rev Genet, **42**: 235-251.
11. Nam, J., et al., 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A, **101**(7): 1910-1915.
12. Amores, A., et al., 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res, **14**(1): 1-10.
13. Nei, M. and A.P. Rooney, 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet, **39**: 121-152.
14. Sun, Y., X. Zhou, and H. Ma, 2007. Genome-wide Analysis of Kelch Repeat-containing F-box Family. J Integr Plant Biol, **49**(6): 940-952.
15. Li, X., et al., 2006. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. Plant Physiol, **141**(4): 1167-1184.
16. Chaudhuri, I., J. Soding, and A.N. Lupas, 2008. Evolution of the beta-propeller fold. Proteins, **71**(2): 795-803.
17. Graur, D. and W.H. Li, 2000, Fundamentals of Molecular Evolution. 2nd edSunderland, MA: Sinauer.
18. Kong, H., et al., 2007. Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. Plant J,

50(5): 873-885.
19. Jiang, S.Y., et al., 2009. Expansion mechanisms and functional annotations of hypothetical genes in the rice genome. Plant Physiol, **150**(4): 1997-2008.
20. Wang, W., et al., 2006. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell, **18**(8): 1791-1802.
21. Zhang, J., et al., 2004. Evolving protein functional diversity in new genes of Drosophila. Proc Natl Acad Sci U S A, **101**(46): 16246-16250.
22. Bowers, J.E., et al., 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature, **422**(6930): 433-438.
23. Kellis, M., B.W. Birren, and E.S. Lander, 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature, **428**(6983): 617-624.
24. Cui, L., et al., 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res, **16**(6): 738-749.
25. Jiao, Y., et al., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature, **473**(7345): 97-100.
26. Cannon, S.B., et al., 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol, **4**: 10.
27. Freeling, M. and B.C. Thomas, 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res, **16**(7): 805-814.
28. Semon, M. and K.H. Wolfe, 2007. Consequences of genome duplication. Curr Opin Genet Dev, **17**(6): 505-512.
29. Koszul, R. and G. Fischer, 2009. A prominent role for segmental duplications in modeling eukaryotic genomes. C R Biol, **332**(2-3): 254-266.
30. Dehal, P. and J.L. Boore, 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol, **3**(10): e314.
31. Jaillon, O., et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature, **431**(7011): 946-957.
32. Wolfe, K.H. and D.C. Shields, 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature, **387**(6634): 708-713.
33. Tang, H., et al., 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res, **18**(12): 1944-1954.
34. Aury, J.M., et al., 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature, **444**(7116): 171-178.
35. Van de Peer, Y., S. Maere, and A. Meyer, 2009. The evolutionary significance of

ancient genome duplications. Nat Rev Genet, **10**(10): 725-732.
36. Taylor, J.S., Y. Van de Peer, and A. Meyer, 2001. Genome duplication, divergent resolution and speciation. Trends Genet, **17**(6): 299-301.
37. De Bodt, S., S. Maere, and Y. Van de Peer, 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol, **20**(11): 591-597.
38. Zhang, J., H.F. Rosenberg, and M. Nei, 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A, **95**(7): 3708-3713.
39. Yokoyama, S. and R. Yokoyama, 1989. Molecular evolution of human visual pigment genes. Mol Biol Evol, **6**(2): 186-197.
40. Zhang, J., Y.P. Zhang, and H.F. Rosenberg, 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet, **30**(4): 411-415.
41. Blomme, T., et al., 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol, **7**(5): R43.
42. Otto, S.P. and J. Whitton, 2000. Polyploid incidence and evolution. Annu Rev Genet, **34**: 401-437.
43. Force, A., et al., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics, **151**(4): 1531-1545.
44. Des Marais, D.L. and M.D. Rausher, 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature, **454**(7205): 762-765.
45. Hittinger, C.T. and S.B. Carroll, 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature, **449**(7163): 677-681.
46. He, X. and J. Zhang, 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics, **169**(2): 1157-1164.
47. Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. Syst Zool, **19**(2): 99-113.
48. Webber, C. and C.P. Ponting, 2004. Genes and homology. Curr Biol, **14**(9): R332-333.
49. Van de Peer, Y., 2006. Evolutionary genetics: when duplicated genes don't stick to the rules. Heredity, **96**(3): 204-205.
50. Lin, Z., et al., 2006. Origins and evolution of the *recA/RAD51* gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. Proc Natl Acad Sci USA, **103**(27): 10328-10333.
51. Lin, Z., M. Nei, and H. Ma, 2007. The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution. Nucleic Acids Res, **35**(22): 7591-7603.
52. Surcel, A., et al., 2008. Long-term maintenance of stable copy number in the

eukaryotic *SMC* family: origin of a vertebrate meiotic *SMC1* and fate of recent segmental duplicates. J Syst Evol, **46**(3): 405-423.
53. Liu, Y., T.A. Richards, and S.J. Aves, 2009. Ancient diversification of eukaryotic MCM DNA replication proteins. BMC Evol Biol, **9**: 60.
54. Fares, M.A. and K.H. Wolfe, 2003. Positive selection and subfunctionalization of duplicated CCT chaperonin subunits. Mol Biol Evol, **20**(10): 1588-1597.
55. Xu, G., et al., 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci U S A, **106**(3): 835-840.
56. Cardozo, T. and M. Pagano, 2004. The SCF ubiquitin ligase: insights into a molecular machine. Nat Rev Mol Cell Biol, **5**(9): 739-751.
57. Xu, G., et al., 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proceedings of the National Academy of Sciences of the United States of America, **106**(3): 835-840.
58. Pagel, M., 1999. Inferring the historical patterns of biological evolution. Nature, **401**(6756): 877-884.
59. Wang, R.L., et al., 1999. The limits of selection during maize domestication. Nature, **398**(6724): 236-239.
60. Yates, T.L., et al., 2002. The ecology and evolutionary history of an emergent disease: Hantavirus pulmonary syndrome. Bioscience, **52**(11): 989-998.
61. Gaucher, E.A., J.T. Kratzer, and R.N. Randall, 2010. Deep Phylogeny—How a Tree Can Help Characterize Early Life on Earth. Cold Spring Harb Perspect Biol, **2**(1).
62. Chatton, E., 1925. Pansporella perplexa. Re ﬂexions sur la biologie et la phyloge ́nie des protozoaires. Ann Sci Nat Zool (Ser 10), **8**: 80.
63. Baldauf, S.L., et al., 2004, The Tree of Life: An Overview, in Assembling the Tree of Life, J. Cracraft and M.J. Donoghue, Editors. Oxford University Press, Inc.: New York, NY.
64. Woese, C.R. and G.E. Fox, 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A, **74**(11): 5088-5090.
65. Woese, C.R., O. Kandler, and M.L. Wheelis, 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA, **87**(12): 4576-4579.
66. Whittaker, R.H., 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science, **163**(863): 150-160.
67. Knoll, A.H., 1992. The early evolution of eukaryotes: a geological perspective. Science, **256**(5057): 622-627.

68. Sogin, M.L., 1991. Early evolution and the origin of eukaryotes. Curr Opin Genet Dev, **1**(4): 457-463.
69. Adl, S.M., et al., 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol, **52**(5): 399-451.
70. Simpson, A.G. and A.J. Roger, 2004. The real 'kingdoms' of eukaryotes. Curr Biol, **14**(17): R693-696.
71. Cavalier-Smith, T., 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol, **52**(Pt 2): 297-354.
72. Sogin, M.L. and J.D. Silberman, 1998. Evolution of the protists and protistan parasites from the perspective of molecular systematics. Int J Parasitol, **28**(1): 11-20.
73. Hampl, V., et al., 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci USA, **106**(10): 3859-3864.
74. Stechmann, A. and T. Cavalier-Smith, 2002. Rooting the eukaryote tree by using a derived gene fusion. Science, **297**(5578): 89-91.
75. Richards, T.A. and T. Cavalier-Smith, 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature, **436**(7054): 1113-1118.
76. Rogozin, I.B., et al., 2009. Analysis of Rare Genomic Changes Does Not Support the Unikont–Bikont Phylogeny and Suggests Cyanobacterial Symbiosis as the Point of Primary Radiation of Eukaryotes. Genome Biol Evol, **1**(1): 99-113.
77. Arisue, N., M. Hasegawa, and T. Hashimoto, 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol Biol Evol, **22**(3): 409-420.
78. Koonin, E.V., 2007. The Biological Big Bang model for the major transitions in evolution. Biol Direct, **2**: 21.
79. Arisue, N., et al., 2002. The phylogenetic position of the pelobiont Mastigamoeba balamuthi based on sequences of rDNA and translation elongation factors EF-1alpha and EF-2. J Eukaryot Microbiol, **49**(1): 1-10.
80. Burki, F., K. Shalchian-Tabrizi, and J. Pawlowski, 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol Lett, **4**(4): 366-369.
81. Sogin, M.L., G. Hinkle, and D.D. Leipe, 1993. Universal tree of life. Nature, **362**(6423): 795.
82. Kamaishi, T., et al., 1996. Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, Glugea plecoglossi: implications for the deepest branching of eukaryotes. J Biochem, **120**(6): 1095-1103.
83. Nakamura, Y., et al., 1996. Phylogenetic position of kinetoplastid protozoa

inferred from the protein phylogenies of elongation factors 1alpha and 2. J Biochem, **119**(1): 70-79.
84. Delsuc, F., H. Brinkmann, and H. Philippe, 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet, **6**(5): 361-375.
85. Jeffroy, O., et al., 2006. Phylogenomics: the beginning of incongruence? Trends Genet, **22**(4): 225-231.
86. Kallersjo, M., et al., 1998. Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst Evol, **213**(3-4): 259-287.
87. Rokas, A. and S.B. Carroll, 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol, **22**(5): 1337-1344.
88. Hillis, D.M., et al., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol, **52**(1): 124-126.
89. Philippe, H., N. Lartillot, and H. Brinkmann, 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol, **22**(5): 1246-1253.
90. Dunn, C.W., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature, **452**(7188): 745-749.
91. Liu, Y., et al., 2009. Phylogenomic analyses support the monophyly of Taphrinomycotina, including Schizosaccharomyces fission yeasts. Mol Biol Evol, **26**(1): 27-34.
92. Liu, Y., et al., 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. BMC Evol Biol, **9**: 272.
93. Zwickl, D.J. and D.M. Hillis, 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol, **51**(4): 588-598.
94. Parfrey, L.W., et al., 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol, **59**(5): 518-533.
95. Rokas, A., et al., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature, **425**(6960): 798-804.
96. Kouzarides, T., 2007. Chromatin modifications and their function. Cell, **128**(4): 693-705.
97. Martin, C. and Y. Zhang, 2005. The diverse functions of histone lysine methylation. Nat Rev Mol Cell Biol, **6**(11): 838-849.
98. Murray, K., 1964. The occurrence of Epsilon-N-methyl lysine in histones. Biochemistry, **3**: 10-15.
99. Rea, S., et al., 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature, **406**(6796): 593-599.

100. Shi, Y., et al., 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell, **119**(7): 941-953.
101. Stavropoulos, P., G. Blobel, and A. Hoelz, 2006. Crystal structure and mechanism of human lysine-specific demethylase-1. Nat Struct Mol Biol, **13**(7): 626-632.
102. Tochio, N., et al., 2006. Solution structure of the SWIRM domain of human histone demethylase LSD1. Structure, **14**(3): 457-468.
103. Chen, Y., et al., 2006. Crystal structure of human histone lysine-specific demethylase 1 (LSD1). Proc Natl Acad Sci U S A, **103**(38): 13956-13961.
104. Shi, Y. and J.R. Whetstine, 2007. Dynamic regulation of histone lysine methylation by demethylases. Mol Cell, **25**(1): 1-14.
105. Metzger, E., et al., 2005. LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. Nature, **437**(7057): 436-439.
106. Tsukada, Y., et al., 2006. Histone demethylation by a family of JmjC domain-containing proteins. Nature, **439**(7078): 811-816.
107. Clissold, P.M. and C.P. Ponting, 2001. JmjC: cupin metalloenzyme-like domains in jumonji, hairless and phospholipase A2beta. Trends Biochem Sci, **26**(1): 7-9.
108. Agger, K., et al., 2008. The emerging functions of histone demethylases. Curr Opin Genet Dev, **18**(2): 159-168.
109. Chang, B., et al., 2007. JMJD6 is a histone arginine demethylase. Science, **318**(5849): 444-447.
110. Wang, J., et al., 2007. Opposing LSD1 complexes function in developmental gene activation and repression programmes. Nature, **446**(7138): 882-887.
111. Di Stefano, L., et al., 2007. Mutation of Drosophila Lsd1 disrupts H3-K4 methylation, resulting in tissue-specific defects during development. Curr Biol, **17**(9): 808-812.
112. Jiang, D., et al., 2007. Arabidopsis relatives of the human Lysine-Specific Demethylase1 repress the expression of FWA and FLOWERING LOCUS C and thus promote the floral transition. Plant Cell, **19**(10): 2975-2987.
113. Liu, F., et al., 2007. The Arabidopsis RNA-binding protein FCA requires a Lysine-Specific Demethylase 1 homolog to downregulate FLC. Mol Cell, **28**(3): 398-407.
114. Krichevsky, A., et al., 2007. C2H2 zinc finger-SET histone methyltransferase is a plant-specific chromatin modifier. Dev Biol, **303**(1): 259-269.
115. He, Y., S.D. Michaels, and R.M. Amasino, 2003. Regulation of flowering time by histone acetylation in Arabidopsis. Science, **302**(5651): 1751-1754.
116. Agger, K., et al., 2007. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature, **449**(7163): 731-734.
117. Klose, R.J., et al., 2007. The retinoblastoma binding protein RBP2 is an H3K4 demethylase. Cell, **128**(5): 889-900.

118. Jepsen, K., et al., 2007. SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron. Nature, **450**(7168): 415-419.
119. Iwase, S., et al., 2007. The X-linked mental retardation gene SMCX/JARID1C defines a family of histone H3 lysine 4 demethylases. Cell, **128**(6): 1077-1088.
120. Tahiliani, M., et al., 2007. The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. Nature, **447**(7144): 601-605.
121. Loh, Y.H., et al., 2007. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. Genes Dev, **21**(20): 2545-2557.
122. Noh, B., et al., 2004. Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time. Plant Cell, **16**(10): 2601-2613.
123. Saze, H., et al., 2008. Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. Science, **319**(5862): 462-465.
124. Klose, R.J., E.M. Kallin, and Y. Zhang, 2006. JmjC-domain-containing proteins and histone demethylation. Nat Rev Genet, **7**(9): 715-727.
125. Schultz, J., et al., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A, **95**(11): 5857-5864.
126. Finn, R.D., et al., 2006. Pfam: clans, web tools and services. Nucleic Acids Res, **34**(Database issue): D247-251.
127. Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, **32**(5): 1792-1797.
128. Tamura, K., et al., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol, **24**(8): 1596-1599.
129. Nicholas, K.B., Nicholas H.B. Jr., and Deerfield, D.W. II., 1997. GeneDoc: Analysis and Visualization of Genetic Variation. EMBNETNEWS, **4**: 14.
130. Thompson, J.D., et al., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, **25**(24): 4876-4882.
131. Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol, **24**(8): 1586-1591.
132. Guindon, S. and O. Gascuel, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, **52**(5): 696-704.
133. Abascal, F., R. Zardoya, and D. Posada, 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics, **21**(9): 2104-2105.
134. Tuskan, G.A., et al., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, **313**(5793): 1596-1604.
135. Lan, F., et al., 2007. Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. Nature, **448**(7154): 718-722.

136. Gordon, M., et al., 2007. Genome-wide dynamics of SAPHIRE, an essential complex for gene activation and chromatin boundaries. Mol Cell Biol, **27**(11): 4058-4069.
137. Rusche, L.N., A.L. Kirchmaier, and J. Rine, 2003. The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. Annu Rev Biochem, **72**: 481-516.
138. Yang, M., et al., 2006. Structural basis for CoREST-dependent demethylation of nucleosomes by the human LSD1 histone demethylase. Mol Cell, **23**(3): 377-387.
139. Nicolas, E., et al., 2006. Fission yeast homologs of human histone H3 lysine 4 demethylase regulate a common set of genes with diverse functions. J Biol Chem, **281**(47): 35983-35988.
140. Hahn, M.W., M.V. Han, and S.G. Han, 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet, **3**(11): e197.
141. Wapinski, I., et al., 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature, **449**(7158): 54-61.
142. Godmann, M., et al., 2007. Dynamic regulation of histone H3 methylation at lysine 4 in mammalian spermatogenesis. Biol Reprod, **77**(5): 754-764.
143. Perry, J. and Y. Zhao, 2003. The CW domain, a structural module shared amongst vertebrates, vertebrate-infecting parasites and higher plants. Trends Biochem Sci, **28**(11): 576-580.
144. Springer, N.M., et al., 2003. Comparative analysis of SET domain proteins in maize and Arabidopsis reveals multiple duplications preceding the divergence of monocots and dicots. Plant Physiol, **132**(2): 907-925.
145. Mouradov, A., F. Cremer, and G. Coupland, 2002. Control of flowering time: interacting pathways as a basis for diversity. Plant Cell, **14 Suppl**: S111-130.
146. Zimmermann, P., et al., 2004. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol, **136**(1): 2621-2632.
147. Zhang, X., et al., 2005. Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. Plant Mol Biol, **58**(3): 401-419.
148. Long, M., et al., 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet, **4**(11): 865-875.
149. Aravind, L. and L.M. Iyer, 2002. The SWIRM domain: a conserved module found in chromosomal proteins points to novel chromatin-modifying activities. Genome Biol, **3**(8): RESEARCH0039.
150. Jaillon, O., et al., 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, **449**(7161): 463-467.
151. Hirschberg, J., 2001. Carotenoid biosynthesis in flowering plants. Curr Opin Plant Biol, **4**(3): 210-218.

152. Timmis, J.N., et al., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet, **5**(2): 123-135.
153. Armbrust, E.V., et al., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science, **306**(5693): 79-86.
154. Sandmann, G., 2002. Molecular evolution of carotenoid biosynthesis from bacteria to plants. Physiol Plant, **116**: 431-440.
155. Malik, H.S. and S. Henikoff, 2003. Phylogenomics of the nucleosome. Nat Struct Biol, **10**(11): 882-891.
156. Allis, C.D., et al., 2007. New nomenclature for chromatin-modifying enzymes. Cell, **131**(4): 633-636.
157. Yamane, K., et al., 2006. JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. Cell, **125**(3): 483-495.
158. Christensen, J., et al., 2007. RBP2 belongs to a family of demethylases, specific for tri-and dimethylated lysine 4 on histone 3. Cell, **128**(6): 1063-1076.
159. Yamane, K., et al., 2007. PLU-1 is an H3K4 demethylase involved in transcriptional repression and breast cancer cell proliferation. Mol Cell, **25**(6): 801-812.
160. Lee, M.G., et al., 2007. Physical and functional association of a trimethyl H3K4 demethylase and Ring6a/MBLR, a polycomb-like protein. Cell, **128**(5): 877-887.
161. Fuchs, J., et al., 2006. Chromosomal histone modification patterns--from conservation to diversity. Trends Plant Sci, **11**(4): 199-208.
162. Cloos, P.A., et al., 2006. The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. Nature, **442**(7100): 307-311.
163. Fodor, B.D., et al., 2006. Jmjd2b antagonizes H3K9 trimethylation at pericentric heterochromatin in mammalian cells. Genes Dev, **20**(12): 1557-1562.
164. Klose, R.J., et al., 2006. The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. Nature, **442**(7100): 312-316.
165. Whetstine, J.R., et al., 2006. Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. Cell, **125**(3): 467-481.
166. Pandey, R., et al., 2002. Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes. Nucleic Acids Res, **30**(23): 5036-5055.
167. Garcia-Bassets, I., et al., 2007. Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors. Cell, **128**(3): 505-518.
168. Okada, Y., et al., 2007. Histone demethylase JHDM2A is critical for Tnp1 and Prm1 transcription and spermatogenesis. Nature, **450**(7166): 119-123.
169. Lan, F., et al., 2007. A histone H3 lysine 27 demethylase regulates animal

posterior development. Nature, **449**(7163): 689-694.
170. Basu, M.K., et al., 2008. Evolution of protein domain promiscuity in eukaryotes. Genome Res, **18**(3): 449-461.
171. Alvarez-Buylla, E.R., et al., 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci U S A, **97**(10): 5328-5333.
172. Miki, H., Y. Okada, and N. Hirokawa, 2005. Analysis of the kinesin superfamily: insights into structure and function. Trends Cell Biol, **15**(9): 467-476.
173. Zhou, X. and H. Ma, 2008. Evolutionary history of histone demethylase families: distinct evolutionary patterns suggest functional divergence. BMC Evol Biol, **8**: 294.
174. Makarova, K.S., et al., 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res, **33**(14): 4626-4638.
175. Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, **25**(17): 3389-3402.
176. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, **30**(7): 1575-1584.
177. Zdobnov, E.M. and R. Apweiler, 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics, **17**(9): 847-848.
178. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, **25**(15): 1972-1973.
179. Jones, D.T., W.R. Taylor, and J.M. Thornton, 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci, **8**(3): 275-282.
180. Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, **5**: 164-166.
181. Le, S.Q. and O. Gascuel, 2008. An improved general amino acid replacement matrix. Mol Biol Evol, **25**(7): 1307-1320.
182. Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, **22**(21): 2688-2690.
183. Anisimova, M. and O. Gascuel, 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol, **55**(4): 539-552.
184. Guindon, S., et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol, **59**(3): 307-321.
185. Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology.

The Gene Ontology Consortium. Nat Genet, **25**(1): 25-29.
186. Ontology, G. *www.geneontology.org*. 2009.
187. Bauer, S., et al., 2008. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics, **24**(14): 1650-1651.
188. Tatusov, R.L., et al., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, **4**: 41.
189. Wang, G., et al., 2004. Genome-wide analysis of the cyclin family in Arabidopsis and comparative phylogenetic analysis of plant cyclin-like proteins. Plant Physiol, **135**(2): 1084-1099.
190. Hillis, D.M. and J.J. Bull, 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Syst Biol, **42**(2): 182-192.
191. Hedges, S.B., et al., 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol, **4**: 2.
192. Blair Hedges, S. and S. Kumar, 2003. Genomic clocks and evolutionary timescales. Trends Genet, **19**(4): 200-206.
193. Keeling, P.J., et al., 2005. The tree of eukaryotes. Trends Ecol Evol, **20**(12): 670-676.
194. Simpson, A.G., et al., 2004. Early evolution within kinetoplastids (euglenozoa), and the late emergence of trypanosomatids. Protist, **155**(4): 407-422.
195. Marques-Bonet, T., S. Girirajan, and E.E. Eichler, 2009. The origins and impact of primate segmental duplications. Trends Genet, **25**(10): 443-454.
196. Van de Peer, Y., 2004. Computational approaches to unveiling ancient genome duplications. Nat Rev Genet, **5**(10): 752-763.
197. Seoighe, C., 2003. Turning the clock back on ancient genome duplication. Curr Opin Genet Dev, **13**(6): 636-643.
198. Blanc, G. and K.H. Wolfe, 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell, **16**(7): 1667-1678.
199. Scannell, D.R., et al., 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature, **440**(7082): 341-345.
200. Ciccarelli, F.D., et al., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, **311**(5765): 1283-1287.
201. Parfrey, L.W., et al., 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS Genet, **2**(12): e220.
202. Hedtke, S.M., T.M. Townsend, and D.M. Hillis, 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst Biol, **55**(3): 522-529.
203. Townsend, J.P. and F. Lopez-Giraldez, 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst Biol, **59**(4): 446-457.

204. Sanderson, M.J. and A.C. Driskell, 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci, **8**(8): 374-379.
205. Keeling, P.J. and Y. Inagaki, 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc Natl Acad Sci USA, **101**(43): 15380-15385.
206. Simpson, A.G., T.A. Perley, and E. Lara, 2008. Lateral transfer of the gene for a widely used marker, alpha-tubulin, indicated by a multi-protein study of the phylogenetic position of Andalucia (Excavata). Mol Phylogen Evol, **47**(1): 366-377.
207. Aguileta, G., et al., 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst Biol, **57**(4): 613-627.
208. Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst Biol, **56**(2): 222-231.
209. Yoon, H.S., et al., 2008. Broadly sampled multigene trees of eukaryotes. BMC Evol Biol, **8**: 14.
210. Tekle, Y.I., et al., 2010. Identification of new molecular markers for assembling the eukaryotic tree of life. Mol Phylogen Evol, **55**(3): 1177-1182.
211. Duarte, J.M., et al., 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol Biol, **10**: 61.
212. Chen, F., et al., 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res, **34**(Database issue): D363-368.
213. Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol, **17**(4): 540-552.
214. Nicholas, K.B., H.B. Nicholas, and D.W. Deerfield, 1997. GeneDoc: analysis and visualization of genetic variation. EMBNETNEWS, **4**: 1-4.
215. Lartillot, N., T. Lepage, and S. Blanquart, 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics, **25**(17): 2286-2288.
216. Quang le, S., O. Gascuel, and N. Lartillot, 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics, **24**(20): 2317-2323.
217. Regier, J.C., et al., 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst Biol, **57**(6): 920-938.
218. Ronquist, F. and J.P. Huelsenbeck, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, **19**(12): 1572-1574.
219. Shimodaira, H. and M. Hasegawa, 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics, **17**(12): 1246-1247.
220. Philippe, H., et al., 2011. Resolving difficult phylogenetic questions: why more

sequences are not enough. PLoS Biol, **9**(3): e1000602.
221. Andersson, J.O., 2009. Gene transfer and diversification of microbial eukaryotes. Annu Rev Microbiol, **63**: 177-193.
222. Baurain, D., H. Brinkmann, and H. Philippe, 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol, **24**(1): 6-9.
223. Rodriguez-Ezpeleta, N., et al., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol, **56**(3): 389-399.
224. Baurain, D., et al., 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol Biol Evol, **27**(7): 1698-1709.
225. Simpson, A.G., Y. Inagaki, and A.J. Roger, 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol Biol Evol, **23**(3): 615-625.
226. Stechmann, A. and T. Cavalier-Smith, 2003. The root of the eukaryote tree pinpointed. Curr Biol, **13**(17): R665-666.
227. Cavalier-Smith, T., 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett, **6**(3): 342-345.
228. Fitzpatrick, D.A., et al., 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol, **6**: 99.
229. Wang, H., et al., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol, **9**: 195.
230. James, T.Y., et al., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature, **443**(7113): 818-822.
231. Sogin, M.L. and J.D. Silberman, 1998. Evolution of the protists and protistan parasites from the perspective of molecular systematics. Int J Parasitol, **28**(1): 11-20.
232. Fischer, W.M. and J.D. Palmer, 2005. Evidence from small-subunit ribosomal RNA sequences for a fungal origin of Microsporidia. Mol Phylogen Evol, **36**(3): 606-622.
233. Gill, E.E. and N.M. Fast, 2006. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. Gene, **375**: 103-109.
234. Keeling, P.J., 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet Biol, **38**(3): 298-309.
235. Katinka, M.D., et al., 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature, **414**(6862): 450-453.
236. Hirt, R.P., et al., 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci USA,

96(2): 580-585.
237. Schoch, C.L., et al., 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst Biol, **58**(2): 224-239.
238. Lumbsch, H.T., et al., 2005. Performance of four ribosomal DNA regions to infer higher-level phylogenetic relationships of inoperculate euascomycetes (Leotiomyceta). Mol Phylogen Evol, **34**(3): 512-524.
239. Spatafora, J.W., et al., 2006. A five-gene phylogeny of Pezizomycotina. Mycologia, **98**(6): 1018-1028.
240. Lutzoni, F., et al., 2004. Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. Am J Bot, **91**(10): 1446-1480.
241. Halanych, K.M., 2004. The new view of animal phylogeny. Annu Rev Ecol, Evol Syst, **35**(1): 229-256.
242. Philippe, H., et al., 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol, **19**(8): 706-712.
243. Schierwater, B., et al., 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol, **7**(1): e1000020.
244. Pick, K.S., et al., 2010. Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships. Mol Biol Evol, **27**(9): 1983-1987.
245. Telford, M.J., et al., 2008. The evolution of the Ecdysozoa. Philos Trans R Soc Lond, Ser B: Biol Sci, **363**(1496): 1529-1537.
246. Lartillot, N. and H. Philippe, 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol, **21**(6): 1095-1109.
247. Lartillot, N. and H. Philippe, 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond, Ser B: Biol Sci, **363**(1496): 1463-1472.
248. Delsuc, F., et al., 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature, **439**(7079): 965-968.
249. Nesnidal, M.P., et al., 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol, **27**(9): 2095-2104.
250. Peterson, K.J. and D.J. Eernisse, 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev, **3**(3): 170-205.
251. Cameron, C.B., J.R. Garey, and B.J. Swalla, 2000. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. Proc Natl Acad Sci USA, **97**(9): 4469-4474.
252. Delsuc, F., et al., 2008. Additional molecular support for the new chordate

phylogeny. Genesis, **46**(11): 592-604.
253. Philippe, H., et al., 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS One, **2**(1): e717.
254. Halanych, K.M., et al., 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science, **267**(5204): 1641-1643.
255. Arendt, D., et al., 2008. The evolution of nervous system centralization. Philos Trans R Soc Lond, Ser B: Biol Sci, **363**(1496): 1523-1528.
256. Holland, N.D., 2003. Early central nervous system evolution: an era of skin brains? Nat Rev Neurosci, **4**(8): 617-627.
257. Nomaksteinsky, M., et al., 2009. Centralization of the deuterostome nervous system predates chordates. Curr Biol, **19**(15): 1264-1269.
258. Springer, M.S., et al., 2004. Molecules consolidate the placental mammal tree. Trends Ecol Evol, **19**(8): 430-438.
259. Hallstrom, B.M. and A. Janke, 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evol Biol, **8**: 162.
260. Nishihara, H., S. Maruyama, and N. Okada, 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci USA, **106**(13): 5235-5240.
261. Churakov, G., et al., 2009. Mosaic retroposon insertion patterns in placental mammals. Genome Res, **19**(5): 868-875.
262. Murphy, W.J., et al., 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res, **17**(4): 413-421.
263. Janecka, J.E., et al., 2007. Molecular and genomic data identify the closest living relative of primates. Science, **318**(5851): 792-794.
264. Kriegs, J.O., et al., 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol, **4**(4): e91.
265. Kullberg, M., et al., 2006. Housekeeping genes for phylogenetic analysis of eutherian relationships. Mol Biol Evol, **23**(8): 1493-1503.
266. Hallstrom, B.M. and A. Janke, 2010. Mammalian evolution may not be strictly bifurcating. Mol Biol Evol, **27**(12): 2804-2816.
267. Prasad, A.B., M.W. Allard, and E.D. Green, 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol, **25**(9): 1795-1808.
268. Nishihara, H., N. Okada, and M. Hasegawa, 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol, **8**(9): R199.
269. Nikolaev, S., et al., 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet, **3**(1): e2.
270. Belinky, F., O. Cohen, and D. Huchon, 2010. Large-scale parsimony analysis of

metazoan indels in protein-coding genes. Mol Biol Evol, **27**(2): 441-451.
271. McKenna, M.C., 1975, Toward a phylogenetic classification of the Mammalia, in Phylogeny of the Primates: A Multidisciplinary Approach, W.P. Luckett and F.S. Szalay, Editors. Plenum: New York. 21-46.
272. Philippe, H., et al., 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature, **470**(7333): 255-258.
273. Rodriguez-Ezpeleta, N., et al., 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol, **15**(14): 1325-1330.
274. Rodriguez-Ezpeleta, N., et al., 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol, **17**(16): 1420-1425.
275. Bapteste, E., et al., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci USA, **99**(3): 1414-1419.
276. Blair, J.E., et al., 2002. The evolutionary position of nematodes. BMC Evol Biol, **2**: 7.
277. Wolf, Y.I., I.B. Rogozin, and E.V. Koonin, 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res, **14**(1): 29-36.
278. Holton, T.A. and D. Pisani, 2010. Deep genomic-scale analyses of the metazoa reject coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biol Evol, **2**: 310-324.
279. Archibald, J.M., 2009. The puzzle of plastid evolution. Curr Biol, **19**(2): R81-88.
280. Kim, E. and L.E. Graham, 2008. EEF2 analysis challenges the monophyly of Archaeplastida and Chromalveolata. PLoS One, **3**(7): e2621.
281. Malik, S.B., et al., 2007. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. Mol Biol Evol, **24**(12): 2827-2841.
282. Zong, J., et al., 2009. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. Gene, **447**(1): 29-39.
283. Archambault, J. and J.D. Friesen, 1993. Genetics of eukaryotic RNA polymerases I, II, and III. Microbiol Rev, **57**(3): 703-724.
284. Filee, J., et al., 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol, **54**(6): 763-773.
285. Kearsey, S.E. and K. Labib, 1998. MCM proteins: evolution, properties, and role in DNA replication. Biochim Biophys Acta, **1398**(2): 113-136.
286. Gupta, R.S., 1995. Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol Microbiol, **15**(1): 1-11.

287. Hughes, A.L., 1997. Evolution of the proteasome components. Immunogenetics, **46**(2): 82-92.
288. Wang, H., et al., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol, **9**: 195.

VITA

Xiaofan Zhou

Intercollege Graduate Program in Cell and Developmental Biology
The Pennsylvania State University, University Park, PA 16802, USA
Phone: 814-865-3438, Email:xxz135@psu.edu

Education

Sep 2006 – Present	Ph.D. candidate; The Pennsylvania State University, USA
Sep 2001 – Jul 2005	B.S.; Biotechnology; Shanghai Jiao Tong University, P.R. China

Honors and Awards

2010	The Pennsylvania State University Braddock Graduate Fellowship
2008	The J. Ben and Helen D. Hill Memorial Fund Award
2008	ASPB 2008 Meeting Travel Grant
2007	The Pennsylvania State University Graduate Fellowship
2003	Excellent Academic Scholarship, Class B
2002	Excellent Academic Scholarship, Class C
2002	Exceptional Student of Shanghai Jiao Tong University

Publications

Zhou X, Sun Y, Ma H. Robust eukaryotic phylogenies using a moderate number of single-copy genes: from the root to tips. *In prep,* October 2010.

Zhou X, Lin Z, Ma H. (2010). Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants. *Genome Biology*, 11(4):R38.

Feng B, Li L, **Zhou X**, Stanley B, Ma H. (2009). Analysis of the Arabidopsis floral proteome: detection of over 2000 proteins and evidence for posttranslational modifications. *Journal of Integrative Plant Biology*, 51(2):207-223.

Zhou X, Ma H. (2008). Evolutionary history of histone demethylase families: distinct evolutionary patterns suggest functional divergence. *BMC Evolutionary Biology*, 8:294.

Surcel A, **Zhou X**, Quan L, Ma H. (2008). Long-term maintenance of stable copy number in the eukaryotic SMC family: origin of a vertebrate meiotic SMC1 and fate of recent segmental duplicates. *Journal of Systematics and Evolution*, 46(3):405-423.

Sun Y, **Zhou X**, Ma H. (2007). Genome-wide analysis of Kelch Repeat-containing F-box family. *Journal of Integrative Plant Biology*, 49(6):940-952.

Jiang D, Yin C, Yu A, **Zhou X**, Liang W, Yuan Z, Xu Y, Yu Q, Wen T, Zhang D. (2006). Duplication and expression analysis of multicopy miRNA gene family members in Arabidopsis and rice. *Cell Research*, 16(5):507-518.

Selected presentations

Zhou X, Ma H. (2009). The relationship of evolutionary pattern and gene function: histone demethylase families as a case study. *2009 Annual Meeting of the Society for Molecular Biology and Evolution*, June 3-7, Iowa City, USA. Oral presentation.

Zhou X, Sun Y, Surcel A, Ma H. (2008). The relationship of gene evolutionary pattern and function of eukaryotic gene family. *Joint Annual Meeting of the American Society of Plant Biologists and the Sociedad Mexicana De Bioquimica 2008*, June 26-July 1, Merida, Mexico. Oral presentation.