MOLECULAR EVOLUTION OF SET-DOMIN PROTEIN FAMILIES IN EUKARYOTES

by

Chendhore S. Veerappan

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master in Science

Major: Biological Sciences

Under the supervision of Professors Etsuko N. Moriyama, Zoya Avramova and Catherine Chia

Lincoln, Nebraska

July, 2007

MOLECULAR EVOLUTION OF SET-DOMIN PROTEIN FAMILIES IN EUKARYOTES

Chendhore S. Veerappan, M.S.

University of Nebraska, 2007

Advisor: Etsuko N. Moriyama, Zoya Avramova, Catherine Chia

SET-domain proteins are histone methyl transferases. These proteins methylate histone tails of nucleosomes, which aid in the modification of chromatin. SET-domain proteins are responsible for transcriptional activation or repression of genes in eukaryotes. Understanding the distribution of these proteins among the three eukaryotic kingdoms would shed lights on the evolution and specialization of eukaryotic functions in the plant, animal, and fungal kingdoms. In this thesis, extensive analyses were conducted on the fungal kingdom and searches were conducted on eleven fungal genomes which included representatives from filamentous, non-filamentous and dimorphic fungal species. *Arabidopsis thaliana*, *Mus musculus* and *Drosophila melanogaster* were chosen to represent the plant and animal kingdoms. Thorough examination of the SET-domain protein distribution in fourteen genomes across fungal, plant, and animal kingdoms was conducted. One hundred and forty four or more of new SET-domain proteins were identified. Phylogenetic analyses of these SET-domain sequences showed the consistent relationships with those based on the internal structure of the SET-domains and with the domain architecture of the entire proteins. Some SET-domain protein groups were

represented by all three kingdoms; others were specific to fungi or animals. While the majority of SET-domain protein groups were found both in filamentous and non-filamentous fungi, a few groups were represented by only a limited fungal species. In conclusion, results show a wide distribution of SET-domain proteins in fungi, plants, and animals. Conservation of family-specific domain architectures suggests that other domains associated with the SET-domain might also play an important role in the functions of these proteins.

ACKNOWLEDGEMENTS

This thesis is dedicated to my parents Padma Veerappan and V.R. Veerappan, whose love for life, sincerity and dedication for education serves as my constant guiding beacon. My heartfelt gratitude to my sister Shridhevi, who has always been my constant companion and supporter. I am very grateful to Krishna Mohanraj for all the support and encouragement.

This work would have not been possible without the constant guidance, patience and support of Dr. Etsuko Moriyama. I am extremely indebted to her in shaping a lost and confused undergraduate student into a more complete and dedicated scholar. She has instilled in me the drive to push towards perfection with utmost dedication and sincerity. I am very grateful to Dr. Zoya Avramova for all the interesting talks about Epigenetics. Her excitement and enthusiasm has been instrumental in shaping my thesis.

I am very grateful to Dr. Hideaki Moriyama, Dr. Alan Christensen and Dr. Heriberto Cerutti for the opportunity to work, teach and gain valuable research experience.

My special gratitude to Ivybelle Palu for helping me realize the meaning of Oneness and the many profound realities and revelations I have come to see.

I am very grateful to Srikanth Anumalla for all the talks and ideas about the Vedas, AI and for accidentally discovering the “Vedic junk DNA”. I would like to thank my lab mates, Cory and Pooja Strope, Stephen Opiyo and Mamta Bajaj for the support and team work in the Moriyama Lab.

Last but not least, I owe it all to the inherent randomness in the Universe without which, nothing is possible.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Histone Modification

Histones are proteins which are found associated with DNA in Eukaryotic genomes. This association between histones and DNA form structures called chromatin. Chromatin is composed of repeating basic units called nucleosomes. Nucleosomes consist of around 200 base pairs (bp) of DNA wrapped around a histone ocatamer. The histone octamer consists of two copes of histone proteins H3, H4, H2A, and H2B. Repeating nucleosomal sub-units along with histone protein H1, form higher order chromatin structures which effectively package DNA in the nucleus, as shown in Figure 1.1.

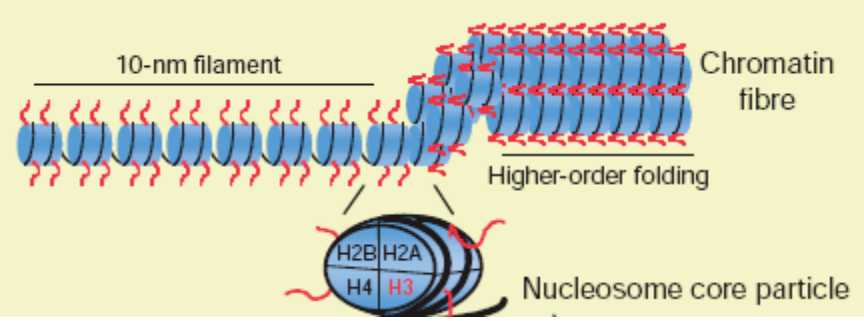


Figure 1.1: Chromatin fibre structure along with higher-order folding. The nucleosome core particle is shown with two copies of four core histone proteins (H3, H4, H2A, and H2B). The N-terminal tail domain of the histone proteins are exposed and extended indicated in red (courtesy [1]).

The N-terminus domain of histone core proteins commonly referred to as ‘histone tails’, protrude from the surface of the chromatin complex (Figure 1.1). These structures

are centres of various covalent modifications such as methylation, acetylation, phosphorylation, ubiquitination and ADP-ribosylation [1]. Histone modification largely seen on particular substrates on histone tails has been shown to regulate DNA transcription by alteration of chromatin structure. These alterations may expose DNA previously unavailable for transcription by the transcription machinery. Alternatively, the changes may render DNA unavailable for transcription. Studies have shown covalent modification of histones to cause inheritable changes in DNA transcription states [2].

Recent studies have revealed some of the high complexity of histone modifications. Various combinations of specific covalent modifications on specific sites on histone tails result in a certain transcription state. These changes constitute a sort of language which has been termed as the 'histone code' [1]. Figure 1.2 shows the various sites on the tails of histones H3 and H4 observed with different covalent modifications.

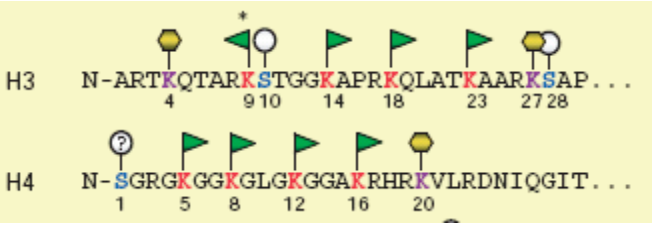


Figure 1.2: N-terminal tails of H3 and H4. Substrates shown to be methylated are indicated in purple; phosphorylation shown in blue and acetylation in red. The asterisk on residue lysine 9 on H3 indicates that this residue is a substrate for acetylation as well as methylation.

Histone modifications along with other epigenetic changes are very important areas of study to further understand the complexity of biological species as well as their evolution. Differential expression and regulation of genes in various organisms by these

modifications may provide an explanation for the extreme physiological and anatomical complexities seen in nature. Such complexities do not seem to correspond directly with the number of genes observed in a genome [2].

One important covalent modification of histones tails is histone methylation. Histone methylation in eukaryotes is carried out by a set of enzymes called histone methyl transferases (HMTs). Histone methylation target lysine or arginine residues. With the exception of a DOT1 family protein [3, 4], all lysine methyltransferase enzymes contain a highly conserved peptide sequence called the SET-domain. SET-domain proteins transfer a methyl group to the amino group of lysine using the co-factor methyl donor from *S*-adeno-syl-L-methionine (AdoMet) [5].

1.2 The SET-domain

The SET-domain is approximately a 130 amino acid motif initially found in three *Drosophila melanogaster* proteins: Suppressor of variegation 3-9 (Su(var)3-9) [6], Enhancer of Zeste (E(z)) [7], and Trithorax (Trx) [8]. This conserved sequence has been identified in plants, animals, and yeast. The absence of SET-domain proteins in prokaryotic genomes indicates that this protein family is involved in eukaryotic functions. Some SET-domain proteins function as histone-modifying enzymes. They have methyltransferase activities; they methylate specific lysines on histone-tails. It has been widely accepted that chromatin remodelling caused by histone-modifying enzymes, which alter the structure of nucleosomes or influence the binding of other molecules,

affects either gene expression or repression. The mechanism by which methylation results in expression or repression of genes is, however, yet to be resolved. In Drosophila, for example, Su(var)3-9 proteins are responsible for the di- and tri- methylation of the Histone 3 lysine 9 and position effect variegation, and mediate gene repression [9]. However, many SET-domain proteins have been identified only based on their sequence similarities and their functions still remain uncharacterized. SET-domain proteins have recently become a focal point of study in cancer proliferation. Mis-regulations of human homologues of E(z) and Trx SET families have been implicated in the development of human cancer [10, 11]. It is, therefore, important to understand the functional and evolutionary relationships of these protein families from multitudes of perspectives, from eukaryotic evolution, epigenetic gene regulations, to cancer genetics.

Within its ~130 amino acid sequence, two non-contiguous conserved regions are recognized: SET-N at the N-terminal and SET-C at the C-terminal ends. The insert region between SET-N and SET-C is highly variable in length (ranging from 23 to 361 amino acids) and called SET-I [12]. Three particularly highly conserved regions, one in SET-N and two in SET-C, are considered to play important roles in cofactor-binding, substrate-binding, and catalytic activity, respectively [12, 13]. A sample multiple sequence alignment with insert and conserved sequence highlighted is shown in Appendix A.

The SET-domain containing proteins, recognized as the SET-domain protein super family, were initially classified under three families: Su(var)3-9, E(z), and Trx. The Su(var)3-9 and E(z) families are mainly involved in gene repression, while the Trx family is involved in positive regulation of homeotic genes [14]. The Su(var)3-9 group methylates the Histone 3 lysine 9 that creates a binding site for the heterochromatin

protein 1 (HP1), which results in heterochromatin formation [15]. The E(z) proteins were shown to have no methylase activity but use the N-terminal region of the SET-domain to bind to the *Drosophila* Extra sex combs protein (Esc) to form a protein complex required for the maintenance of a previous repression pattern [7]. An intriguing point about these families is that they all share the SET-domain, but have counteractive functions; some resulting in gene expression and some in gene repression. Another SET-domain protein family called Ash 1 distinguishes itself from the other families by internal location of the SET-domain within the protein [16]. Trx and Ash 1 proteins are part of the triothorax group (trx-G) genes that interact together using their SET-domains and are shown to be involved in homeotic gene expression in *Drosophila* [17]. Ash 1 protein in *Drosophila* has been shown to methylate lysine residues 4 and 9 in Histone 3 and lysine 20 of Histone 4. It targets genes and provide a binding site for a chromatin remodelling complex and subsequent activation [18].

In *Saccharomyces cerevisiae*, a Trithorax-related protein called SET 1 was found to be part of a protein complex (Complex Proteins Associated with SET 1), which is involved in a gene expression [19]. SET 2 in *S. cerevisiae* methylates Histone 3 lysine 36 and is associated with transcription elongation [20]. SET 2 was also shown to be essential for normal cellular growth and development in *Neurospora crassa* [21]. Subsequently other related proteins, SET 3, SET 4, SET 5, SET 6 have been identified in *Saccharomyces cerevisiae* [22]. SET 7/9 was isolated in human [23] and SET 8 has been have been identified in metazoans [24].

In many proteins, the SET-domain usually exists with other domains forming multi-domain protein families [25]. Single-domain SET proteins are known only in some

bacteria and they are considered to be the results of horizontal transfers from eukaryotes to pathogenic bacteria [26, 27]. Due to their involvements in fundamentally important eukaryotic functions involving histone modification and chromatin remodelling, it is of a great interest to examine how different types of SET-domain proteins are distributed across eukaryotes. Work conducted on comparative analysis of SET-domain proteins in plants [28, 29] showed similarity of some SET-domain families in animals and plants, suggesting ancient origins of these proteins. SET-domain proteins of E(z), Trx, Ash 1, and Su(var)3-9 families are evolutionarily conserved in animals, plants, and yeast [30]. Alvarez-Venegas and Avramova [25] showed that Su(var), Trx, and E(z) family proteins exist across plant, animal, and fungal species.

1.3 Goals of this study

With the discovery of many new SET-domain proteins in various eukaryotes, thorough examination of SET-domain protein repertoire from the three major kingdoms is in order. The main goals of this study are as follows:

1. To trace the evolution of the SET-domain families in eukaryotic genomes and in particular, the fungal kingdom.
2. To gain more insight in the evolution of cellular complexity and specialization in fungi to further understand the overall evolution of multicellularity in eukaryotes.
3. To get an overall picture of the interaction between the different SET domain families and other related domains present in the gene.

Only few studies have been done on the distribution of SET-domain proteins especially among fungi. This thesis primarily addresses the distribution of the SET-domain genes in Ascomycota (or sac fungi). The Ascomycota sub-phylum forms a monophyletic group which include sister clades, Pezizomycotina and Saccharomycotina (Figure 1.3). The Saccharomycotina group has several interesting sub-groups. There is a well supported branch leading to Candida species which translate CTG as serine instead of leucine. The sister group of this Candida group is another well supported group which consists of closely related species of bakers yeast, *Saccharomyces cerevisiae*. This group includes members (see Figure 1.3) which have undergone whole genome duplication (WGD) and those which have not. Studies have shown that the most common ancestor of this group may have undergone speciation and the subsequent duplication event lead to the evolution of the baker's yeast [31]. It is interesting to note that members of this group, filamentous fungus *Ashbya gossypii* shares homology with single-celled *Saccharomyces cerevisiae* with over 90% of its genes [31]. The Pezizomycotina group includes dominant filamentous, ascoma producing fungi [32].

The members of Ascomycota and sister group Basidiomycota are highly diverse in their morphology ranging from yeasts, dimorphic fungi to complex mycelial forms (filamentous fungi). The Filamentous fungi present in Ascomycota, as well as Basidiomycota, generally are characterized by the presence of septa (cross-walls) and the formation of unfused nuclei after mating [33]. The basal group to the sister clades Ascomycota and Basidiomycota is the Zygomycota [34]. The members of Zygomycota such as *Rhizopus oryzae* are filamentous fungi, which lack cross-walls or septa [35] and

thus represent a rudimentary form of a filamentous fungus as compared to the well defined 'higher' filamentous fungi of Basidiomycota and Ascomycota.

Ascomycota being a monophyletic group is a good candidate to study the evolution of various fungal morphology types. Therefore, it will be very interesting to further understand the transitional evolution of the most common ancestor of Ascomycota into filamentous, non-filamentous and dimorphic fungi. The distribution of the SET-domain families in this group may reflect this transition. By comparing the presence or absence of SET-domain groups in fungal genomes as well as higher eukaryotes, it may be possible to further understand the role of overall evolution of multicellular species or specialization of SET-domain families for specific functions in plants, fungi and animal evolution.

In this study, SET-domain proteins were identified from ten Ascomycota fungal species (including filamentous, non-filamentous, and dimorphic fungi). A non-filamentous fungus, fission yeast Sc*hizosaccharomyces pombe* was chosen. This species forms a very well supported clade outside the Ascomycota group (see Figure 1.3). *Drosophila melanogaster*, *Mus musculus*, and *Arabidopsis thaliana* were chosen as the representatives of animals and plants.

Results showed that some SET-domain families and their domain architectures are conserved across all three kingdoms, and others are conserved and/or duplicated in only some kingdoms or species. Their implication to the organismal and protein evolution will be discussed.

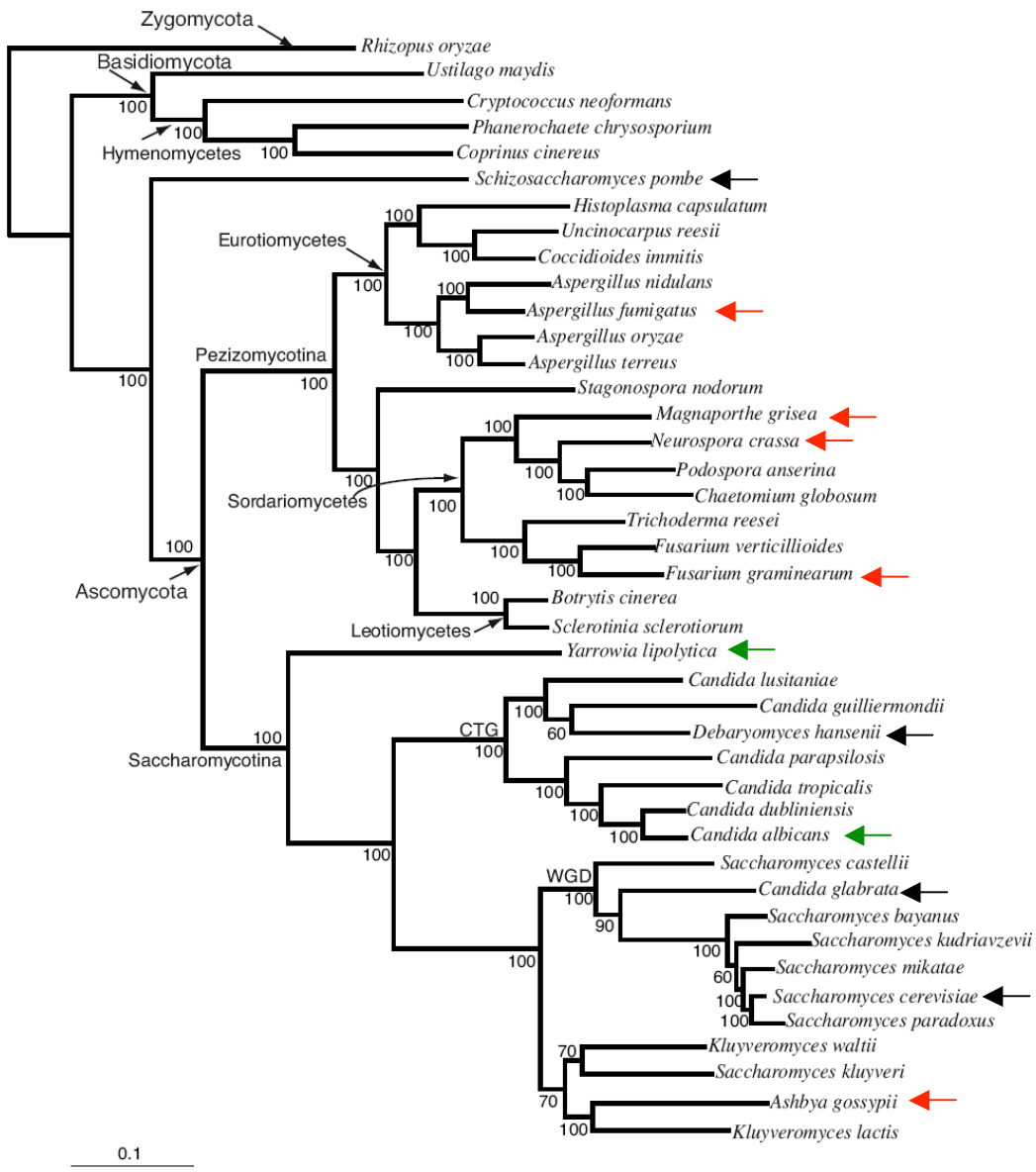


Figure 1.3: Maximum likelihood phylogeny reconstructed using concatenated alignment of 153 universally distributed fungal genes (courtesy [34]). Bootstrap scores for all clades are displayed. Filamentous fungi used for the study are indicated by red arrows; dimorphic fungi with green arrows and non-filamentous fungi with black arrows.

2 Materials and Methods

2.1 Sequences used

Known SET-domain sequences were collected from various sources. Appendix B lists the 24 sequences used for similarity searches along with their source. The SET-domain sequences were extracted from these proteins and used as the query sequences for the search against twelve complete genomes listed in Table 3.1. They represent the three kingdoms including eleven fungi (five filamentous, four non-filamentous, and two dimorphic species), two animals (*Mus musculus* and *Drosophila melanogaster*), and one plant (*Arabidopsis thaliana*). All genomic sequences were current as of May 20, 2005.

2.2 SET-domain protein mining

The following four similarity search methods were used to mine new SET-domain proteins from the twelve genomes.

2.2.1 BLAST protein similarity searches

The initial protein similarity search was conducted with BLASTP [36] using the 24 SET-domain sequences as the queries against the non-redundant database available at National Center for Biotechnology Information (NCBI) with the default settings. In order to find any similar protein regions from unannotated genomic regions, we also used

TBLASTN [37] to perform similarity searches against nucleotide sequences of the twelve genomes, translated in all six frames.

More sensitive searches were performed using the position specific iteration BLAST (PSI-BLAST) [37]. Each query was used against individual genomes with the inclusion E-value threshold of 0.001 and four search iterations.

2.2.2 Profile hidden Markov model searches

Profile hidden Markov models (HMMs), probabilistic models of multiple alignments, were built and used to search for sequences with remote similarities [38]. Using the sequences obtained from the above-mentioned BLAST searches and the query sequences, 27 well-aligned sequences were selected which are shown in Appendix C. A profile HMM was built using these sequences with the Sequence Alignment and Modelling System (SAM; [39]). Two programs of the SAM package were used: *buildmodel* for building the profile HMMs and *hmmscore* (with *-sw 2* and *–calibrate 1* options) for searching similar protein sequences from the genomes. The searches were conducted in each genome individually. The resulting hits from each organism were analyzed for the presence of the SET-domain and previously unidentified sequences were collected. An E-value threshold was not used, but rather each hit was examined for the SET-domain.

After the four similarity searches, 214 non-redundant hits were compiled from the twelve genomes (data not shown). Each of these 214 sequences was examined to confirm the presence of the SET-domain by searching the Conserved Domain Database (CDD) available from NCBI [40], as well as the Simple Modular Architecture Research Tool

(SMART) database [41, 42]. Some dubious hits including too highly diverged sequences and those with very short SET-domain like conserved sequences were removed. After these analyses, we obtained 183 non-redundant SET-domain sequences. These sequences were used in further analyses.

2.2.3 Multiple alignments of SET-domain sequences

CLUSTALX (version 1.83; [43]) was used to generate multiple alignments of SET-domain sequences (with the GONNET series protein weight matrices and gap opening penalty = 10 and gap extension penalty = .20). Due to the highly variable length of the SET-I region, poorly conserved sites across the sequences were removed. Some other highly variable positions were also removed and the alignments were adjusted manually.

2.2.4 Phylogenetic analyses

A draft phylogeny was reconstructed using all of the SET-domain sequences found in this study using the maximum likelihood method implemented in PHYML Version 2.4.4 [44]. This draft phylogeny (shown in Appendix D) was used in the further analyses.

In order to produce a more reliable multiple alignment and phylogenies, the number of sequences was reduced by choosing representative SET-domain sequences using the draft phylogeny as a guide tree. Poorly aligned sequences and those not clustering clearly with any known major SET-domain families were removed. All fungal sequences were retained, while only one each representative hit from plants and animals

was chosen from each SET-domain cluster of SET 1, SET 2, Su(var), and E(z). The selected representative sequences are indicated with red fonts in the complete dataset phylogeny (see Appendix D). These sequences were selected based on the multiple alignment of the entire dataset as well as the protein architectures. The multiple alignment of these sequences is shown in Appendix E. The final multiple alignment including 114 sequences is shown in Appendix F.

Phylogenetic reconstruction was done using the maximum likelihood method (implemented in PHYML Version 2.4.4) [45] and the maximum parsimony method (implemented in PHYLIP 3.65) [46]. The neighbour-joining method was not used because the estimated distance matrix using the JTT substitution model (implemented in PHYLIP 3.65) generated estimation errors due to too many substitutions. For the maximum likelihood method, two sets of trees were reconstructed: one with no invariable site and a constant substitution rate among sites, and the other with the proportion of invariable sites and the gamma shape parameter estimated from the data. Trees reconstructed using the estimated proportion of invariable and gamma shape parameter was used for our analysis. The estimated values are indicated on the respective phylogeny figures. For the maximum parsimony method, the input sequence order was jumbled 10 times. Phylogenetic confidence was estimated by the bootstrap analysis [47] with 500 pseudoreplicates.

2.2.5 *Yarrowia lipolytica* and *Dabaryomyces hansenii* Genomes

The Genomes of *Y. lipolytica* and *D. hansenii* were not included in the initial searches and the phylogenetic reconstructions of the data set. In order to increase the resolution of the Saccharomycotina group in Table 3.1, detailed searches were conducted with the previously described profile HMM in the two genomes. The presence of the SET-domain was confirmed for each hit, using SMART and CDD, and SET-domain families were assigned using protein similarity searches (using BLASTP) in the non-redundant (NR) database of NCBI. Dimorphic fungus *Y. lipolytica* did not fall within CTG or the Saccharomyces clade and was used to represent an ancient divergence within the Saccharomycotina clade. *D. hansenii* was picked to represent a non-filamentous fungus from the CTG clade (Figure 1.3).

3 Results and Discussion

3.1 Searching and phylogenetic analysis of SET-domain proteins

Starting with 24 known SET-domain protein sequences, shown in Appendix B, and performing a series of similarity searches from fourteen complete genomes including eleven fungi, two animals, and one plant as listed in Table 3.1. 183 SET- domain sequences were identified. Detailed phylogenetic analysis was conducted using 114 representative sequences. Figure 3.1 shows the phylogeny reconstructed using the maximum likelihood method with the proportion of invariant sites and gamma shape parameter estimated. The maximum parsimony phylogeny was also largely consistent with this phylogeny (see Appendix G). Thirteen distinct SET-domain protein clusters were identified. Each family has a distinct gene architecture distribution. SET-domain proteins identified were classified based on the multiple alignment (see Appendix E), Protein architectures (see Appendix H), and phylogenetic analyses (see Appendix D).

3.2 Distribution of the SET-domain proteins identified across twelve genomes

Table 3.2 summarizes the distribution of the SET-domain families identified in this study based on the representative phylogeny, along with the bootstrap values.

Sequences which could not be classified definitively into a specific family, based on alignment and phylogeny, are not included in the table. These sequences are listed in Appendix I; the accession numbers of these proteins with tentative SET-domain family assignment based only on similarity with other SET-domain families are shown. BLASTP against the NR database of NCBI was used for the similarity searches of the unknown SET-domain sequences.

The multiple sequence alignment of the representative phylogeny is shown in Appendix F. Three SET-domain families, SET 1, SET 2, and SET 3/4, were found across all genomes examined. It suggests that these three families may have been present in the common ancestor of the three kingdoms. Animals and *A. thaliana* (a plant) have many more SET 1 and SET 2 proteins than fungi. This may imply specialization of these proteins for more complex multicelluar functions. In fungi, a single copy of the SET 1 protein is present in all fungal genomes used in the study; however, multiple copies of the SET 2 proteins were found in all filamentous fungi with the exception of *Magnaporthe griesa*. Non-filamentous and dimorphic fungi only have a single copy of the SET 2 protein.

The SET 2 family also includes the Ash 1 protein from *D. melanogaster*, which is a well supported clade along with single copies of the gene from filamentous fungi representing Pezizomycotina. It is interesting to note that SET 2 proteins and Ash 1 have different methylation specificities but share the same basic gene architecture. Previously classified Ash 1 and SET 2 proteins are indicated in Figure 3.1. The branch leading to the SET 2 family is well supported (91% bootstrap value by maximum likelihood phylogeny) and the basic domain architecture found in the *Saccharomyces cerevisiae* SET 2

(Genbank: NP012367) are conserved among all SET 2 proteins and the Ash 1 proteins (see Appendix I for domain architectures).

E(z) family appears to be specific to higher eukaryotes and filamentous fungi from Pezizomycotina. SET 7/9 is found only in the mouse genome, and SET 8 is animal specific. Plant specific SET-domain proteins were not found. Su(var)3-9 is distributed among filamentous fungi from Pezizomycotina, *Schizosaccharomyces pombe,* and higher eukaryotes, but missing from non-filamentous fungi, *S. cerevisiae* and *Candida glabrata*. Su(var) 4-20 is found in filamentous fungi from Pezizomycotina, *Schizosaccharomyces pombe,* and animal representatives. SET MYND family is also distributed among filamentous fungi from Pezizomycotina, *S. pombe*,and higher eukaryotes with the exception of *D. melanogaster.* SET 5, SET 6, and SET JmjC are fungal specific, but they do not exist across all fungal species that were examined. SET 5 and SET 6 were found only from non-filamentous and dimorphic fungi and SET JmjC was found only in filamentous fungi but appears to be absent from the *Fusarium graminearum* genome.

More detailed description of each of the SET-domain protein families are described next.

Table 3.1: Twelve genomes used in the study

Species[a]	**Size/genes**	**Sources and Genbank accession mumbers**
Fungi [filamentous]		
Neurospora crassa (Nc)	40 Mb/10,082	Fungal Genome Initiative [48, 49]
Aspergillus fumigatus (Af)	29.4 Mb/9,926	SANGER Institute [50]
Fusarium graminearum[b] (Fg)	40 Mb/11,640	Fungal Genome Initiative [51]
Magnaporthe grisea (Mg)	37.8 Mb/11,109	Fungal Genome Initiative [52]
Eremothecium gossypii[b] (Eg)	9.2 Mb/ 4,718	NCBI: NC005782, NC005788 (Chromosomes 1-7) [31]
Fungi [dimorphic]		
Yarrowia lipolytica (Yl)	20.5 Mb/6700	European Molecular Biology Labaratory (EMBL) [53]
Candida albicans (Ca)	15.6 Mb/6,090	CandidaDB [54]
Fungi [non-filamentous]		
Candida glabrata (Cg)	12.3 Mb/5,283	NCBI: NC005967, NC005968, NC006026, NP006037 (Chromosomes A-M) [55]
Schizosaccharomyces pombe (Sp)	14 Mb/4,824	NCBI: NC003424, NC003423, NC003421 (Chromosomes 1-3) [55]
Debaryomyces hansenii (Dh)	12 Mb/7000	EMBL [56]
Saccharomyces cerevisiae (Sc)	12.1 Mb/6,294	*Saccharomyces* Genome Database [52, 57]
Plant		
Arabidopsis thaliana (At)	120 Mb/25,498	MIPS *Arabidopsis thaliana Genome Project* [58]
Animals		
Mus musculus (Mm)	2.5 Gb/24,174	NCBI [59]
Drosophila melanogaster (Dm)	165 Mb/ 13,600	FlyBase [60]

Table 3.1 continued
[a]Species abbreviations used in this study are shown in parentheses.
[b]*Fusarium graminearum* is also known as *Gibberella zeae*. *Eremothecium gossypii* is also known as *Ashbya gossypii*.
The Table tabulates the different genomes used in this study along with the genomes sizes and number of predicted genes. The sources are also indicated along with available accession numbers.

Table 3.2: Distribution of SET-domain families in fourteen genomes

SET-domain protein families	Pezizomycotina species					Saccharomycotina species								
	Sp	Nc	Af	Fg	Mg	Yl	Ca	Eg	Dh	Cg	Sc	Mm	Dm	At
SET 1	1	1	1	1	1	1	1	1	1	1	1	6	2	5
SET 2	1	2	2	2	2	1	1	1	1	1	1	5	3	5
SET 3	1	1	1	1	1	1	1	1	1	1	1	1	1	2
SET 4	0	0	0	0	0	0	0	0	0	1	1	0	0	0
SET 5	0	0	0	0	0	1	1	1	1	1	1	0	0	0
SET 6	0	0	0	0	0	0	1	1	1	1	1	0	0	0
Su(var) 4-20	1	1	1	1	1	1	0	0	0	0	0	2	1	0
SET MYND	1	1	1	1	1	0	0	0	0	0	0	4	2[a]	1
Su(var) 3-9	1	1	1	1	1	0	0	0	0	0	0	7	4	15
E(z)	0	1	0	1	1	0	0	0	0	0	0	2	1	3
SET JmjC	0	1	1	0	2	0	0	0	0	0	0	0	0	0
SET TPR	0	1	1	2	2	0	0	0	0	0	0	0	0	1
SET Mg	0	0	0	0	5	0	0	0	0	0	0	0	0	0
SET 7/9	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SET 8	0	0	0	0	0	0	0	0	0	0	0	2	1	0
SET Dm	0	0	0	0	0	0	0	0	0	0	0	0	7	0
[Total]	**6**	**10**	**9**	**10**	**17**	**5**	**5**	**5**	**5**	**6**	**6**	**28**	**22**	**31**

[a] Two D. melanogaster SET MYND genes shown in this table are included after BLASTP similarity searches of previously unknown *D. melanogaster* SET-domain genes (see Appendix I)

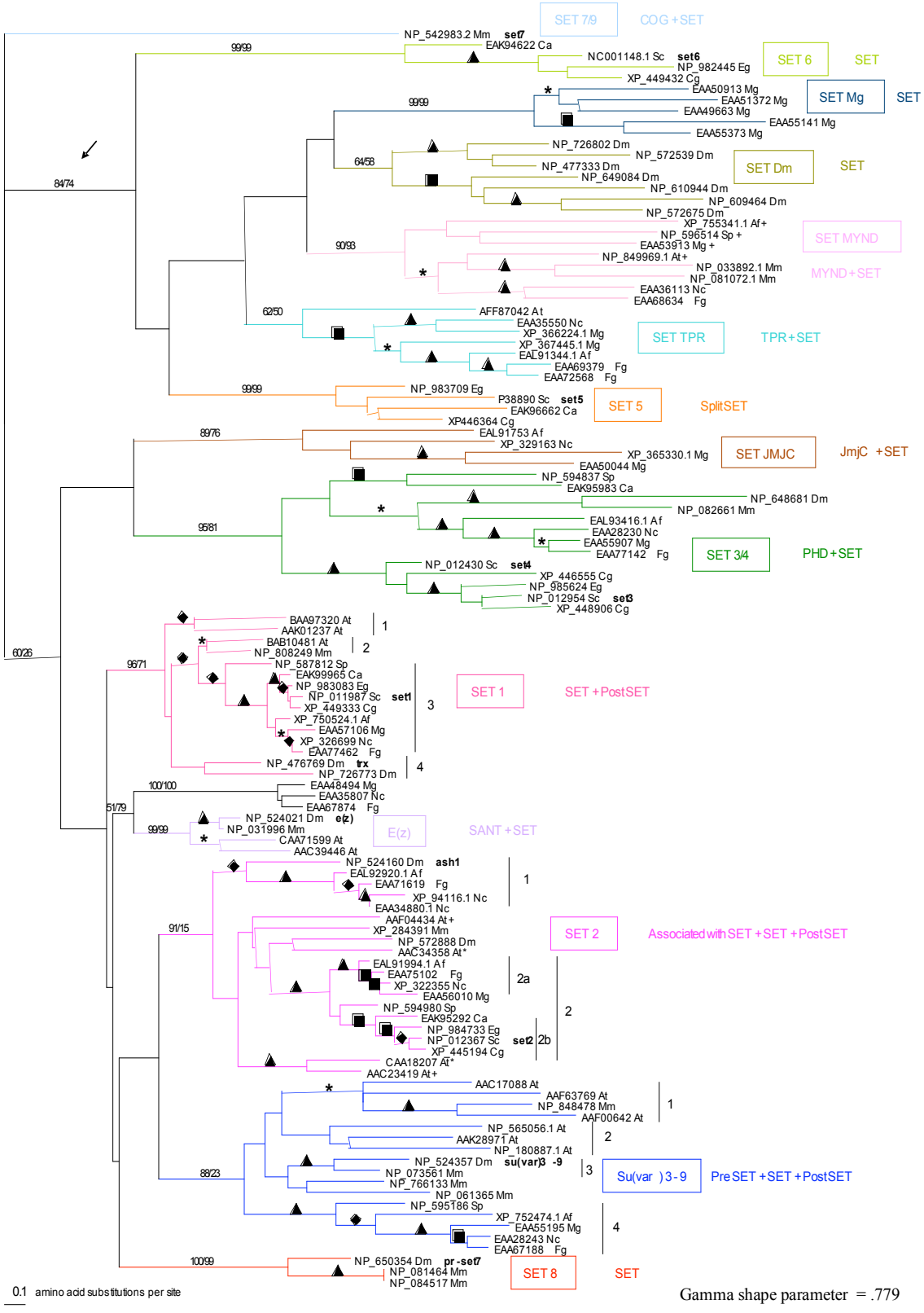

SET 7/9
COG + SET
NP_542983.2 Mm set7
EAK94622 Ca
9999
NC001148.1 Sc set6
NP_982445 Eg
XP_449432 Cg
SET 6
SET
EAA50913 Mg
EAA51372 Mg
EAA49663 Mg
EAA55141 Mg
EAA55373 Mg
SET Mg
SET
9999
NP_726802 Dm
NP_572539 Dm
NP_477333 Dm
NP_649084 Dm
NP_610944 Dm
NP_609464 Dm
NP_572675 Dm
64/58
SET Dm
SET
84/74
XP_755341.1 Af+
NP_596514 Sp+
EAA53913 Mg+
SET MYND
9093
NP_849969.1 At+
NP_033892.1 Mm
NP_081072.1 Mm
EAA36113 Nc
EAA68634 Fg
MYND + SET
AFF87042 At
EAA35550 Nc
XP_366224.1 Mg
XP_367445.1 Mg
EAL91344.1 Af
EAA69379 Fg
EAA72568 Fg
62/50
SET TPR
TPR + SET
NP_983709 Eg
P38890 Sc set5
EAK96662 Ca
XP446364 Cg
9999
SET 5
SplitSET
EAL91753 Af
XP_329163 Nc
XP_365330.1 Mg
EAA50044 Mg
8976
SET JMJC
JmjC + SET
NP_594837 Sp
EAK95983 Ca
NP_648681 Dm
NP_082661 Mm
EAL93416.1 Af
EAA28230 Nc
EAA55907 Mg
EAA77142 Fg
9581
SET 3/4
PHD + SET
NP_012430 Sc set4
XP_446555 Cg
NP_985624 Eg
NP_012954 Sc set3
XP_448906 Cg
BAA97320 At
AAK01237 At
BAB10481 At
NP_808249 Mm
1
2
6026
9671
NP_587812 Sp
EAK99965 Ca
NP_983083 Eg
NP_011987 Sc set1
XP_449333 Cg
XP_750524.1 Af
EAA57106 Mg
XP_326699 Nc
EAA77462 Fg
3
SET 1
SET + PostSET
NP_476769 Dm trx
NP_726773 Dm
4
100/100
EAA48494 Mg
EAA35807 Nc
EAA67874 Fg
51/79
9999
NP_524021 Dm e(z)
NP_031996 Mm
CAA71599 At
AAC39446 At
E(z)
SANT + SET
NP_524160 Dm ash1
EAL92920.1 Af
EAA71619 Fg
XP_94116.1 Nc
EAA34880.1 Nc
1
91/15
AAF04434 At+
XP_284391 Mm
NP_572888 Dm
AAC34358 At*
SET 2
Associated with SET + SET + PostSET
EAL91994.1 Af
EAA75102 Fg
XP_322355 Nc
EAA56010 Mg
2a
NP_594980 Sp
EAK95292 Ca
NP_984733 Eg
NP_012367 Sc set2
XP_445194 Cg
2b
2
CAA18207 At*
AAC23419 At+
AAC17088 At
AAF63769 At
NP_848478 Mm
AAF00642 At
1
NP_565056.1 At
AAK28971 At
NP_180887.1 At
2
NP_524357 Dm su(var)3 -9
NP_073561 Mm
NP_766133 Mm
3
Su(var) 3 - 9
Pre SET + SET + PostSET
8823
NP_061365 Mm
NP_595786 Sp
XP_752474.1 Af
EAA55195 Mg
EAA28243 Nc
EAA67188 Fg
4
10099
NP_650354 Dm pr -set7
NP_081464 Mm
NP_084517 Mm
SET 8
SET
0.1 amino acid substitutions per site
Gamma shape parameter = .779

Figure 3.1: **Maximum likelihood phylogeny of 114 representative SET-domain sequences**. Bootstrap values of major branches leading to SET-domain families that are higher than 60% by either of the maximum likelihood (ML) or the maximum parsimony (MP) methods are shown on the branches using numerical values (ML/MP). Within the major SET-domain protein groups, bootstrap values by the ML analysis greater than 60%, 70%, 80%, and 90% are indicated by a star, circle, square, and triangle, respectively. Sub-groups from each family discussed in the text are indicated by numbers. Further down the nested hierarchy, sub-clades of numbered groups are indicated by lower case letters after the number. Some sub-clades lacking sufficient boostrap supports are also marked for the sake of discussion. Vertical lines with numbers after the sequence names indicate the SET-domain protein subgroups discussed in the text. Different SET-protein families are shown in different colors. Plant genes indicated by either + or * symbols were classified as Ash 1 homologues or Ash-1-related, respectively The domain architecture representing each family is shown next to the SET family labels. The arrow indicates the branch leading to the SET 5/6 'Super' group. See Appendix I for the domain names. See Table 3.1 for species abbreviations. Gamma shape parameter is .779 and proportion of invariant sites is 0.

3.2 Evolution and characteristics of SET-domain protein families

3.2.1 SET 1 Family

This family consists of a fungal specific protein group (clade 3) along with plant and animal specific clades (clades 1 and 4). Clade 2 with plant and animal sequences is supported only weakly with bootstrap values between 60% and 70%. Proteins of the fungal specific clade 3 are single SET-domain proteins. All SET 1 proteins are characterized by the presence of a post SET motif in addition to the SET-domain. The post SET motif is cysteine rich and is shown to be important in the HMTase activity of the protein. However this motif may not play a direct role in the HMTase activity since there are functional SET-domain proteins that do not carry this motif [61]. Plant and

animal sequences also have multiple copies of the Plant Homeo Domain (PHD) finger motif in addition to the post SET motif. This zinc-finger like cysteine-rich motif maybe involved in DNA or, protein-protein interaction required for chromatin-mediated transcriptional regulation [62]. Additionally, RING, FY-rich N terminal region (FYRN, SMART accession number: SM00541), FY-rich C terminal region (FYRC, SMART accession number: SM00542) and bromodomain motifs can be seen in different plant and animal sequences (see Appendix H). The RING finger motif is a cysteine rich zinc-binding domain, between 40-60 amino acids, and maybe involved in protein-protein interaction [63] The FYRN and FYRC motifs are associated with chromatin associated proteins. The bromodomain motif, named after the *Drosophila* brahma, recognizes acetylated lysine residues and involved in chromatin remodelling and transcriptional activation [64].

3.2.2 SET 2 Family

SET 2 family proteins are characterized by the presence of three motifs: the SET-domain, an associated with SET (AWS) motif (SMART accession number: SM00570), and a post SET motif. The SET 2 proteins are present in all fungal, animal, and plant genomes used in this study suggesting the ancient origin of these proteins before the divergence of the three major kingdoms. Figure 3.1 shows the presence of several distinct sub-groups indicated by numbered vertical lines. Clade 1 consists of the *D. melanogaster* Ash1 protein along with filamentous fungi from Pezizomycotina. Clade 2 is fungi specific, and is split into two subgroups: one includes the second copy of the gene from Pezizomycotina representatives (2a) and another consists of representatives from

Saccharomycotina (2b). A WW motif (InterPro accession number: IPR001202) is conserved in all sequences except in *E. gossypii*. The WW domain contains about 40 amino acids and is characterized by the presence of two conserved tryptophan residues. This domain recognizes specific proline containing motifs or phosphoserine-phosphothreonine containing motifs. A *S. pombe* sequence (Genbank: NP594980) without the WW motif is present and could not be placed in any one of the two sub-clades (2a or 2b). In addition to these fungal proteins, *D. melanogaster* protein NP572888 also has a WW motif. In general, plant and animal SET 2 family proteins have multiple copies of PHD finger motifs and poorly conserved motifs which bind to AT-rich DNA, known as AT hooks (InterPro accession number:IPR000637) (see gene architectures in Appendix H). Many animal proteins also have one or more of the PWWP motif which may play a role in protein-protein interaction in nuclear proteins involved in differentiation [65]. It is interesting to note that from a previous study, plant genes indicated by either + or * symbols in Figure 3.1, were classified as Ash 1 homologues or Ash-1-related, respectively [30]. However, our results show that these putative plant Ash 1 proteins do not form a well supported group with the known *Drosophila* Ash1 gene and could not be placed in the Ash1 subgroup (1) or the SET 2 subgroup (2).

3.2.3 SET 3/4 family

S. cerevisiae is known to have two highly similar SET-domain proteins: SET 3 and SET 4 [22]. Springer *et al*. [29] showed that SET 3 and SET 4 proteins from *S. cerevisiae* cluster with two plant proteins (CAB89351.1 and NP197821). However, these proteins lack several amino acids required for HMTase activity. The yeast SET 3 protein

has been shown to be a part of a multi-protein complex that includes histone deacetylases but lacks histone methylase activity *in vitro* [22]. Their comparative analysis of the protein sequences of the yeast and plant sequences revealed no significant sequence conservation. Evidence points to independent duplication events in Arabidopsis and yeast leading to the evolution of the SET 3 and SET 4 families. The phylogenetic trees reconstructed using the maximum likelihood method using the all non-redundant SET-domain sequences revealed that the two plant sequences do not cluster with the clade containing yeast SET 3 and SET 4 sequences (see Appendix D). It was therefore excluded the two plant sequences from the representative SET-domain sequence phylogeny (Figure 3.1).

SET 3 and SET 4 families share an identical domain architecture, which includes a PHD-finger domain in addition to the SET-domain (see gene architectures in Appendix I). Based on such similarity, it is likely that the two families have emerged by duplication. As shown in Figure 3.1, with the exception of *S. cerevisiae* among the nine fungal genomes, a duplication only in the *C. glabrata* genome was found. The clade leading to this clade is well supported. Therefore these two SET 4 sequences are shown as a separate family in Table 3.2. Only a single copy was found in dimorphic species *E. gossypii* and *C. albicans*. From all filamentous fungi as well as *S. pombe*, and animal sequences, only a single representative of SET 3/4 related proteins was found and they form a highly supported protein group (95% bootstrap value in Figure 3.1).

3.2.3.1 SET 3/4 duplication in Saccharaomyces and closely related species

Since both of non-filamentous fungi used, *S. cerevisiae* and *C. glabrata*, have duplicated SET 3/4 proteins, there is a possibility that the duplicated genes are required in the evolution of non-filamentous fungi. Kellis *et al.* [66] showed that *Saccharomyces* species underwent whole genome duplication during its evolution. In order to elucidate the timing of the duplication event during the SET 3/4 family evolution, similarity searches in *S. cerevisiae* and nine closely related *Saccharomyces* species were performed. Duplicated SET 3/4 proteins were found in *S. bayanus*, *S. castelli,* in addition to *S. cerevisiae* and *C. glabrata*. The result is consistent with the study by Kellis *et al* [66]. It places the whole genome duplication after the divergence of the yeast related species: *Eremothecium gossypii*, *Kluyveromyces lactis*, and most recently *Kluyveromyces waltii* [66]. Figure 3.2 shows the maximum likelihood phylogenetic tree with bootstrap values higher than 70%. It indicates that SET 3/4 duplication occurred in the ancestral lineage leading to *Saccharomyces* species and *Candida glabrata* if duplicated genes common to both genomes arose from the whole genome duplication explained in Kellis *et al.* [66]. It is interesting to note that the SET 3 and SET 4 sequences are highly conserved between the *S. cerevisiae* and *Saccharomyces bayanus* species compared to *Saccharomyces castelli* (Figure 3.2). Out results provide evidence that this gene duplication can place the whole genome duplication event before the divergence of *C. glabrata* and the *Saccharomyces* species.

3.2.4 Su(var) 3-9

The Su(var)3-9 family is found in all filamentous fungi from Pezizomycotina, *S. pombe*, plants, and animals. The distinctive feature of this family is the presence of the pre-SET (PRS) motif (InterPro accession number: IPR003606). This motif is found associated to the N- terminal end of the SET-domain and contains nine invariant cysteine residues which co-ordinate three zinc ions. The function of the PRS motif is in the structural stability of the SET-domain [13]. This motif is found in all of the proteins with the exception of the *Arabidopsis* sequence (AAF00642, clade 1). This protein has an "Associated with SET" (AWS) motif (InterPro accession number: IPR006560) instead of the PRS motif. The fungal species also carry the post-SET (PS) domain and the *S. pombe* sequence also carries the CHromatin Organization MOdifier (CHROMO) domain (clade 4). The CHROMO domain (InterPro accession number: IPR000953) is also found in the upstream of the SET-domain motif of animal representative proteins (group 3: NP524357 and NP073561). One animal protein (NP766133) also carries contiguous Ankyrin (ANK) repeats (InterPro accession number: IPR002110) and lacks the PS domain. Proteins in clade 2 (unsupported by bootstrap analysis) carry also the SET and RING finger Associated (SRA) domain in addition to PRS, SET, and PS. Clade 2 appears to be plant specific.

3.2.5 Su(var) 4-20

The Su(var) 4-20 family of SET-domain proteins are involved in pericentric heterochromatin formation along with Su(var)3-9 and Heterochromatin Protein 1 (HP1)

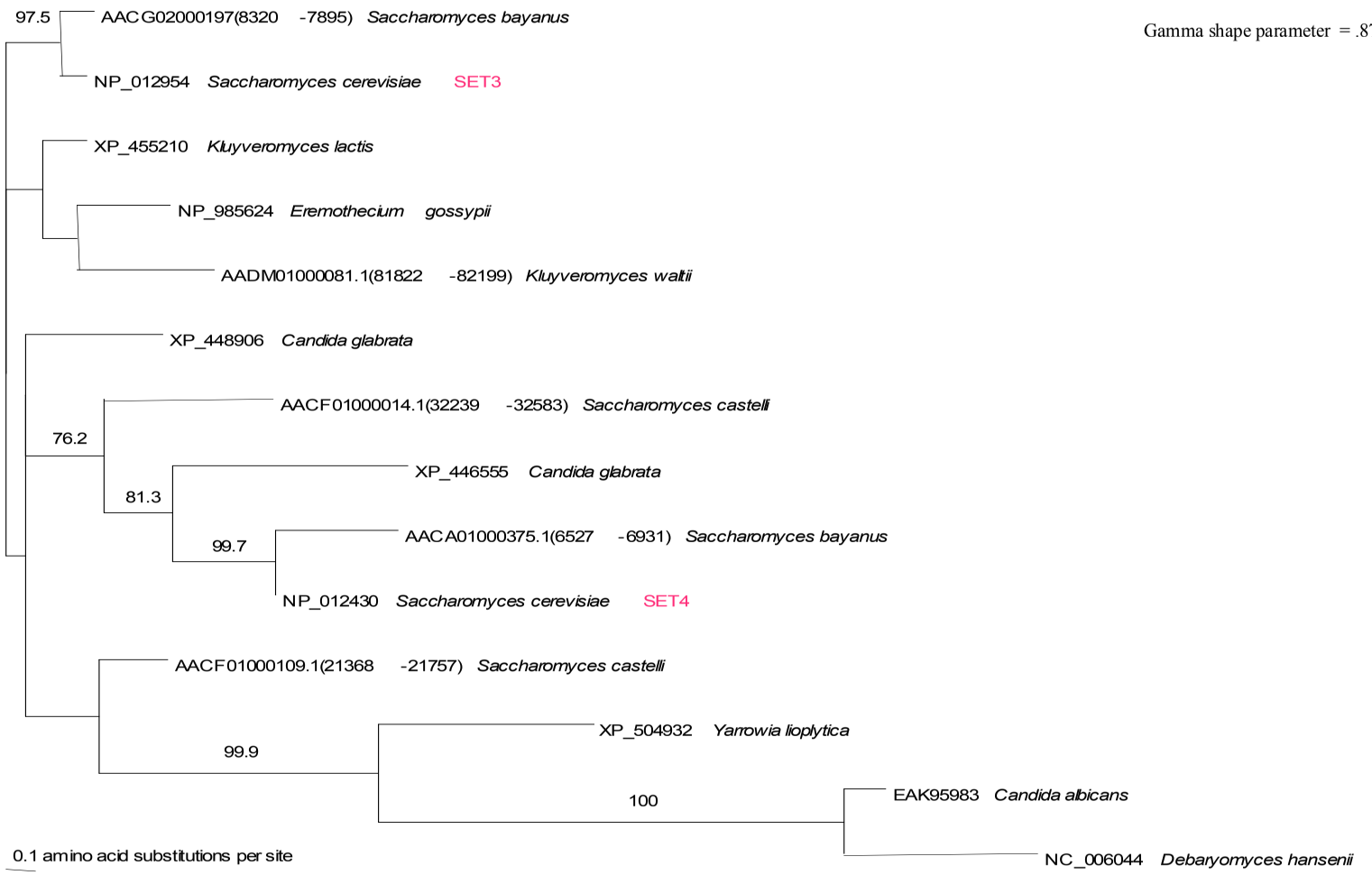

Gamma shape parameter = .879
97.5
AACG02000197(8320 -7895) Saccharomyces bayanus
NP_012954 Saccharomyces cerevisiae SET3
XP_455210 Kluyveromyces lactis
NP_985624 Eremothecium gossypii
AADM01000081.1(81822 -82199) Kluyveromyces waltii
XP_448906 Candida glabrata
AACF01000014.1(32239 -32583) Saccharomyces castelli
76.2
XP_446555 Candida glabrata
81.3
AACA01000375.1(6527 -6931) Saccharomyces bayanus
99.7
NP_012430 Saccharomyces cerevisiae SET4
AACF01000109.1(21368 -21757) Saccharomyces castelli
XP_504932 Yarrowia lioplytica
99.9
100
EAK95983 Candida albicans
NC_006044 Debaryomyces hansenii
0.1 amino acid substitutions per site

Figure 3.2: Maximum likelihood phylogeny of the SET 3/4 family among closely related yeast species. Fourteen SET 3/4 sequences identified from ten genomes of closely related yeast species are included. Bootstrap values are shown if they are greater than 60%. The genomic position (in bp) is indicated in parenthesis for the sequences from unannotated genomes. Gamma shape parameter is .879 and proportion of invariant sites is 0.

[67]. *S. pombe*, filamentous fungi from Pezizomycotina and animals carry this protein family. These proteins carry only the SET-domain. This family is not represented in Figure 3.1 due to sequence divergence. The candidate proteins are indicated by the '*' sign in the draft phylogeny in Appendix D.

3.2.6 SET JmjC

The SET JmjC family proteins have the Jumanji C domain (JmjC domain), which is associated with the jumanji protein required for the neural tube formation shown in mice [68]. The SET JmjC family was found in all filamentous fungi from Pezizomycotina with the exception of *F. graminearum*. The JmjC domain has been shown to be associated with histone demethylase activities. It is very interesting to observe two domains with counteracting functions on the same protein sequence. To speculate a possible function, demethylation of certain residues maybe required prior to methylation by the SET-domain.

3.2.7 SET 5/6 'Super' group

In Figure 3.1, indicated by an arrow sign, there is a well supported clade (84% support) which leads to at least six other relatively well supported SET-domain protein

families. This clade is therefore defined as the SET 5/6 'Super' group. These families include SET 5, SET 6, SET Mg, SET Dm, SET MYND and SET TPR. Multiple alignment data and gene architecture information supports a possible common ancestor for these protein families (see Appendix E and Appendix H). The SET-domain sequence of these families is relatively longer (greater than 200 amino acids) or carry a split SET-domain, as characterized by *S. cerevisiae* SET 6 and SET 5 proteins respectively. This group also includes highly duplicated species specific clusters, SET Mg and SET Dm. The following sub-sections provide a more detailed explanation of the six families along with their distribution in various genomes.

3.2.7.1 SET 5 and SET 6

The SET 5 proteins are usually about 500 amino acids (aa) long and have only the SET-domain. The *S. cerevisiae* SET 5 (P38890) has a split SET-domain (see Appendix H) with a long (~125 aa) SET-I region. The other SET 5 proteins found in the search have similar and shorter SET-domain length and their SET-domains are not the split SET-domain. The SET 5 protein is found all representatives from Saccharomycotina.

The lengths of SET 6 proteins are around 370 amino acids with the SET-domain being approximately 330 amino acids. The SET 6 family proteins have large SET-I region similar to SET 5 proteins. This family is found in all representatives of Sachharomycotina with the exception of *Y. lipolytica*.

3.2.7.2 SET Mg and SET Dm

Two highly duplicated species-specific SET-domain protein families were found: one in *Magnaporthe grisea* (SET Mg) and the other in *D. melanogaster* (SET Dm). Both proteins have only the SET-domain and seem to have undergone multiple genome specific duplications. The SET Mg cluster is highly supported (99% bootstrap value). Although the support for the SET Dm cluster is only about 60%, some of the internal clusters have much higher supports. Some members of the SET Dm group carry the zinc finger MYND motif (see Appendix H for gene architectures). Both of SET Mg and SET Dm are clustered with the larger cluster of SET 5 and SET 6 (Figure 3.1). These SET-domain protein families appear to be prone to duplication.

3.2.7.3 SET MYND

The SET MYND family was found in filamentous fungi from Pezizomycotina, a plant, and animals. This family is characterized by the presence of the MYND (myeloid, nervy, and DEAF-1) motif with some exceptions where the Zinc finger (ZnF) motif is included instead (the *N. crassa, F. graminearum,* and animal proteins). Both motifs are found in the insert region of the SET-domain. The MYND consists of cysteins and histidine residues present in a specific order, which could form a zinc-binding motif [69]. However unlike the DNA binding function of the zinc-binding motif, this motif has also been suggested to have protein-protein interaction [69]. The MYND motif has been shown to recruit co-repressors [70, 71].

3.2.7.4 SET TPR

The main feature of the SET TPR family is the presence of the TPR (tetratrico peptide repeat) motif upstream of the SET-domain. This family was found in plant and filamentous fungi from the Pezizomycotina group. Two copies of this family were found in *M. grisea* and *F. graminearum*. The TPR motif is known to be involved in protein-protein interaction. This may allow assembly of transcription regulatory factors [72]. The exact function of this family is yet to be elucidated.

3.2.8 SET-domain families specific to animals and plants

The SET 7/9 family protein was found only from the mouse genome. Protein similarity search using this sequence as a query found no hit from *D. melanogaster* or any other invertebrates. All hits found were from vertebrates. This suggests a specialization of this gene towards functions related to the presence of vertebrae in species. This protein has, in addition to the SET-domain, multiple MORN (membrane occupation and recognition nexus) motif (InterPro accession number: IPR003409) repeats. The function of this motif is not known.

The SET 8 family is an animal specific group and consists of only the SET-domain in its sequence.

Animal and plant E(z) proteins cluster together with very high bootstrap values (99%). The animal/plant E(z) group is characterized by the presence of the SANT motif (InterPro accession number: IPR001005). Animal proteins have two copies and plant proteins have a single copy of the SANT motif. The SANT motif recognizes the DNA sequence YAAC(G/T)G and takes part in the transcriptional repression activity [73, 74].

Three proteins found from *F. graminearum*, *N. crassa* and *M. griesa* (EAA67874, EAA35807, and XP36972) seem to cluster with this animal/plant E(z) family. However, the bootstrap value was not significant (only 51% in the maximum likelihood method and 79% in the maximum parsimony method). Furthermore, the fungal proteins do not carry the SANT motif. However, these sequences are included under the E(z) family in Table 3.2 based on sequence similarity as shown under the E(z) group in Appendix F.

4 Conclusion and Future Work

Using the initial SET-domain sequence queries to search in fourteen genomes, it is evident that the SET-domain has diversified into the three major eukaryotic kingdoms. There are also species and kingdom specific SET-domain families and these specializations may provide information about the evolution of the three eukaryotic kingdoms. More specifically, analyzing and comparing genes regulated by SET-domain families specific to filamentous and non-filamentous fungi may shed more light on the evolution of multi-cellular fungal species.

4.1 Evolution of the SET-domain in Eukaryotes

Several interesting observations can be made regarding the distribution of SET-domain proteins in fungi and higher eukaryotes.

SET 1, SET 2 and SET 3 are present in all genomes. This indicates that these families are ancient in origin and probably were present before the divergence of the three eukaryotic kingdoms. However, no reliable evolutionary information regarding the evolution of these three families was observed. Figure 3.1 shows a 60% bootstrap support for the branch leading to the three families. However, this branch is not monophyletic. The three families are probably involved in the basic function of a single cell. Higher eukaryotes carry multiple copies of the SET 1 and SET 2 protein. These extra copies also carry various other domains in addition to the SET-domain and this maybe due to specialization for more complex cellular functions.

Focusing on SET 2 proteins in fungi, the evolution of the gene in the fungal species used in this study follows the overall evolution of the genomes themselves. Clade 2 of SET 2 in Figure 3.1 and Figure 1.3 are consistent with the exception of the placement of *S. pombe*. In the fungal tree made from 153 genes (Figure 1.3), *S. pombe* is placed as the outgroup of Ascomycota. However, the general distribution of the family is consistent with the ancient origin and specialization of the SET 2 gene before the divergence of the fungal genomes. Another clade in addition to the *Drosophila* ash1 protein contains filamentous fungi from Pezizomycotina.

Another example of gene evolution following genome evolution of fungi can be found in Clade 3 of SET 1 in Figure 3.1. In this case, there is no discrepancy over the placement of *S. pombe*.

A single copy of the SET 3 gene is found in all genomes with the exception of *A. thaliana* (two copies). Evidence of animal specific specialization of the SET domain can be observed in SET 7/9 and SET 8. Species specific specialization represented by SET Mg and SET Dm is seen in *M. griesa* and *D. melanogaster*.

4.1.1 Evolution of Saccharomycotina genomes

Several transitions and losses of SET-domain proteins are observed in the evolution of the Saccharomyotina sub-group. SET MYND and Su(var)3-9 families are found in all filamentous from Pezizomycotina, *S. pombe* and higher eukaryotes. These two families have been lost during the divergence of the Saccharomycotina sub-group. Su(var)4-20 was lost after the divergence of *Y. lipolytic*a. The presence of Su(var)4-20 in this dimorphic fungus may indicate a transition in the Saccharomycotina sub-group,

towards the evolution of the *Candida* and *Saccharmyces* clades. The explanation for the absence of these families as losses in the evolution of the Saccharomycotina sub-group is most parsimonious, since these families are found in higher eukaryotes and other fungi.

SET 4 is found only in the closely related yeast species *S. cerevisiae*, *S. bayanus*, *S. castelli* and *C. glabrata* (see Figure 3.2). The conservation this gene copy after the whole genome duplication event maybe due to the relatively rapid evolution [75] and specialization of the extra copy in the evolution of the highly efficient fermentative capabilities of *Saccharomyces* yeasts [76]. SET 5 is found in Saccharomycotina species and SET 6 was not found in dimorphic *Y. lipolytica*. SET 6 may have appeared later in branch leading to the evolution of the *Candida* and *Saccharomyces* species.

4.1.2 Evolution of Pezizomycotina species

There are several SET-domain families (E(z), SET JmjC, SET TPR and SET Mg) which are only specific to Pezizomycotina among fungi. E(z) is also found in plants and animals and a single copy of SET TPR in *A. thaliana* (see Table 3.2). This may indicate specialization of the SET-domain in the evolution of filamentous fungi specific to this group.

4.2 Role of the SET-domain proteins in the evolution of Multicellularity

One of the major goals of this study is to further understand the evolution of mulicellularity in Eukaryotes and more specifically the role of the SET-domain in this transition. It was expected that the distribution of the SET-domain among filamentous fungi, dimorphic fungi and higher eukaryotes may have suggested a specialization of the SET-domain for multicellularity. However, Table 3.2 does show any such relationship was found. E(z) is found in all filamentous fungi from Pezizomycotina and higher eukaryotes. However, this gene was not found in *A. gossypii* which is a filamentous fungus. This family may play a role in multicellular functions but is not required for multicellularity.

On the question of placing the fungal kingdom closer to plants or animals in the Tree of Life, our study does not resolve the question based on the distribution of the SET-domain families. SET TPR is found only in fungi and *A .thaliana*, whereas, Su(var)4-20 is found only found in fungi and animals. There is no consistent relationship of fungi with either plants or animals.

4.3 Role of other domains carried by SET-domain proteins

One of the interesting results of the analyses shows the conservation of other domains in the different SET-domain protein families. Various SET-domain families have in addition to the SET-domain, other domains which are conserved in plants,

animals and fungi. A minimum set of domains seem to be responsible for the basic function of a specific family; however, shuffling, replications and addition of new domains are also seen in some proteins. The next step is to further the understanding of the evolution and function of these families by analyzing certain DNA and protein binding domains which may be critical for function.

Apart from the SET-domain sequence, one particular type of domain which is largely distributed in many SET families, such as SET 1, SET 2, SET 3/4, SET MYND, SET Dm and SET 8, is the presence of the zinc finger and zinc finger like domains such as the PHD finger, RING finger and MYND finger. CDD and SMART searches of all 183 SET-domain hits revealed many zinc finger and other zinc finger related sequences found associated with the SET-domain (some of which are shown in Appendix I). The wide distribution of these sequences along with the SET-domain sequences in various SET families suggests an ancient origin of this association.

4.4 Future Work

Appendix I contain the accession numbers of SET-domain sequences which were not included in the analyses. One very interesting feature of the many of these sequences is that their SET-domain share similarity with mostly the C-terminal region of the SET-domain. They are similar to the SET MYND proteins. However most of them lack a MYND finger (using CDD and SMART to detect conserved domains). These may represent a family of genes related to the SET MYND family. These sequences were initially removed from the representative phylogeny in Figure 3.1 because of a lack of a

reliable multiple alignment which included these sequences. The high sequence divergence is seen mainly in the SET5/6 'Super' group. Most of the genes with similarity to the SET MYND domain are found in this group (see arrow indicating SET5/6 'Super' group in Appendix D). In order to overcome to problem of high sequence divergence, a separate multiple alignment was generated using only sequences from the SET5/6 'Super' group from the draft phylogeny. Appendix K shows the phylogeny reconstructed using previously described methods for multiple alignment and maximum likelihood phylogeny generation. To test the reliability of the branches, bootstrap analyses using 500 pseudoreplicates was conducted. The phylogeny obtained, failed to reliably support the previously well classified groups within the SET5/6 'Super' group in Figure 3.1. SET 5 and SET Dm are the only families with marginally acceptable bootstrap values of 60.2% and 71.8% respectively. Further analysis will be conducted on this clade by including more fungal and other eukaryotic species to increase the resolution of the multiple alignment.

It will also be interesting to further understand the association of the zinc finger with the SET-domain in the twelve genomes by performing more sensitive searches on the existing 183 SET-domain sequence dataset. Profile hidden Markov models will be constructed using the zinc finger sequences along with their relatives like the RING, MYND and PHD domains. Performing searches using the SAM package on each sequence may reveal more diverged zinc finger like sequences around or within the SET-domain which were not detected by SMART and CDD. Using this new information, phylogenetic analysis maybe conducted with these sequences to test their co-evolution with the SET-domain.

5 Bibliography

1. Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**(6765):41-45.
2. Jenuwein T, Allis CD: **Translating the Histone Code**. *Science* 2001, **298**(5532):1074.
3. van Leeuwen F, Gafken PR, Gottschling DE: **Dot1p modulates silencing in yeast by methylation of the nucleosome core**. *Cell* 2002, **109**(6):745-756.
4. Feng Q, Wang H, Ng HH, Erdjument-Bromage H, Tempst P, Struhl K, Zhang Y: **Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain**. *Curr Biol* 2002, **12**(12):1052-1058.
5. Dillon SC, Zhang X, Trievel RC, Cheng X: **The SET-domain protein superfamily: protein lysine methyltransferases**. *Genome Biol* 2005, **6**(8):227.
6. Tschiersch B, Hofmann A, Krauss V, Dorn R, Korge G, Reuter G: **The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes**. *Embo J* 1994, **13**(16):3822-3831.
7. Jones RS, Gelbart WM: **The Drosophila Polycomb-group gene Enhancer of zeste contains a region with sequence similarity to trithorax**. *Mol Cell Biol* 1993, **13**(10):6357-6366.
8. Stassen MJ, Bailey D, Nelson S, Chinwalla V, Harte PJ: **The Drosophila trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins**. *Mech Dev* 1995, **52**(2-3):209-223.
9. Ebert A, Schotta G, Lein S, Kubicek S, Krauss V, Jenuwein T, Reuter G: **Su(var) genes regulate the balance between euchromatin and heterochromatin in Drosophila**. *Genes Dev* 2004, **18**(23):2973-2983.
10. Rozovskaia T, Rozenblatt-Rosen O, Sedkov Y, Burakov D, Yano T, Nakamura T, Petruck S, Ben-Simchon L, Croce CM, Mazo A *et al*: **Self-association of the SET domains of human ALL-1 and of Drosophila TRITHORAX and ASH1 proteins**. *Oncogene* 2000, **19**(3):351-357.
11. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP *et al*: **The polycomb group protein EZH2 is involved in progression of prostate cancer**. *Nature* 2002, **419**(6907):624-629.
12. Marmorstein R: **Structure of SET domain proteins: a new twist on histone methylation**. *Trends Biochem Sci* 2003, **28**(2):59-62.
13. Zhang X, Yang Z, Khan SI, Horton JR, Tamaru H, Selker EU, Cheng X: **Structural basis for the product specificity of histone lysine methyltransferases**. *Mol Cell* 2003, **12**(1):177-185.

14. Breen TR, Harte PJ: **Molecular characterization of the trithorax gene, a positive regulator of homeotic gene expression in Drosophila**. *Mech Dev* 1991, **35**(2):113-127.
15. Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD *et al*: **Regulation of chromatin structure by site-specific histone H3 methyltransferases**. *Nature* 2000, **406**(6796):593-599.
16. Tripoulas N, LaJeunesse D, Gildea J, Shearn A: **The Drosophila ash1 gene product, which is localized at specific sites on polytene chromosomes, contains a SET domain and a PHD finger**. *Genetics* 1996, **143**(2):913-928.
17. Rozovskaia T, Tillib S, Smith S, Sedkov Y, Rozenblatt-Rosen O, Petruk S, Yano T, Nakamura T, Ben-Simchon L, Gildea J *et al*: **Trithorax and ASH1 interact directly and associate with the trithorax group-responsive bxd region of the Ultrabithorax promoter**. *Mol Cell Biol* 1999, **19**(9):6441-6447.
18. Beisel C, Imhof A, Greene J, Kremmer E, Sauer F: **Histone methylation by the Drosophila epigenetic transcriptional regulator Ash1**. *Nature* 2002, **419**(6909):857-862.
19. Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF, Shilatifard A: **COMPASS: a complex of proteins associated with a trithorax-related SET domain protein**. *Proc Natl Acad Sci U S A* 2001, **98**(23):12902-12907.
20. Krogan NJ, Kim M, Tong A, Golshani A, Cagney G, Canadien V, Richards DP, Beattie BK, Emili A, Boone C *et al*: **Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II**. *Mol Cell Biol* 2003, **23**(12):4207-4218.
21. Adhvaryu KK, Morris SA, Strahl BD, Selker EU: **Methylation of histone H3 lysine 36 is required for normal development in Neurospora crassa**. *Eukaryot Cell* 2005, **4**(8):1455-1464.
22. Pijnappel WW, Schaft D, Roguev A, Shevchenko A, Tekotte H, Wilm M, Rigaut G, Seraphin B, Aasland R, Stewart AF: **The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program**. *Genes Dev* 2001, **15**(22):2991-3004.
23. Nishioka K, Chuikov S, Sarma K, Erdjument-Bromage H, Allis CD, Tempst P, Reinberg D: **Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation**. *Genes Dev* 2002, **16**(4):479-489.
24. Fang J, Feng Q, Ketel CS, Wang H, Cao R, Xia L, Erdjument-Bromage H, Tempst P, Simon JA, Zhang Y: **Purification and functional characterization of SET8, a nucleosomal histone H4-lysine 20-specific methyltransferase**. *Curr Biol* 2002, **12**(13):1086-1099.
25. Alvarez-Venegas R, Avramova Z: **SET-domain proteins of the Su(var)3-9, E(z) and trithorax families**. *Gene* 2002, **285**(1-2):25-37.
26. Alvarez-Venegas R, Sadder M, Tikhonov A, Avramova Z: **Origin of the Bacterial SET Domain Genes: Vertical or Horizontal?** *Mol Biol Evol* 2007, **24**(2):482-497.

27. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al*: **Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis**. *Science* 1998, **282**(5389):754-759.
28. Yu Y, Dong A, Shen WH: **Molecular characterization of the tobacco SET domain protein NtSET1 unravels its role in histone methylation, chromatin binding, and segregation**. *Plant J* 2004, **40**(5):699-711.
29. Springer NM, Napoli CA, Selinger DA, Pandey R, Cone KC, Chandler VL, Kaeppler HF, Kaeppler SM: **Comparative analysis of SET domain proteins in maize and Arabidopsis reveals multiple duplications preceding the divergence of monocots and dicots**. *Plant Physiol* 2003, **132**(2):907-925.
30. Baumbusch LO, Thorstensen T, Krauss V, Fischer A, Naumann K, Assalkhou R, Schulz I, Reuter G, Aalen RB: **The Arabidopsis thaliana genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes**. *Nucleic Acids Res* 2001, **29**(21):4319-4333.
31. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S *et al*: **The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome**. *Science* 2004, **304**(5668):304-307.
32. Alexopolous CJ, Mims CW, Blackwell M: **Introductory Mycology**, 4 edn: John Wiley & Sons, New York, New York, USA; 1996.
33. Schaffer RL: **The major groups of Basidiomycetes**. *Mycologia* 1975, **66**:1-18.
34. Fitzpatrick DA, Logue ME, Stajich JE, Butler G: **A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis**. *BMC Evol Biol* 2006, **6**:99.
35. Webster J: ***Introduction to fungi***, 2 edn: Cambridge: Cambridge University Press; 1980.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
38. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.
39. Krogh RHaA: **SAM: SEQUENCE ALIGNMENT AND MODELING SOFTWARE SYSTEM**. In. Santa Cruz: University of California, Santa Cruz; 1995.
40. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z *et al*: **CDD: a Conserved Domain Database for protein classification**. *Nucleic Acids Res* 2005, **33**(Database issue):D192-196.
41. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: **SMART: a web-based tool for the study of genetically mobile domains**. *Nucleic Acids Res* 2000, **28**(1):231-234.

42. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32**(Database issue):D142-144.
43. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res* 1997, **25**(24):4876-4882.
44. Guindon S, Lethiec F, Duroux P, Gascuel O: **PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W557-559.
45. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**(5):696-704.
46. **PHYLIP (Phylogeny Inference Package) Version 3.6** [http://evolution.genetics.washington.edu/phylip.html]
47. Felsenstein J: **Confidence limits on phylogenies: An approach using the bootstrap**. *Evolution* 1985, **39**(4):783-791.
48. **Fungal Genome Initiative** [http://www.broad.mit.edu/annotation/]
49. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*: **The genome sequence of the filamentous fungus Neurospora crassa**. *Nature* 2003, **422**(6934):859-868.
50. ***Aspergillus fumigatus* Genome Project** [http://www.sanger.ac.uk/Projects/Fungi/]
51. ***Fusarium graminearun* Sequencing Project** [http://www.broad.mit.edu]
52. ***Magnaporthe grisea* Sequencing project** [http://www.broad.mit.edu]
53. ***Yarrowia Lipolytica* : Chromosomes** [URL: http://cbi.labri.fr/Genolevures/download/YALI_chromosomes.php]
54. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT *et al*: **The diploid genome sequence of Candida albicans**. *Proc Natl Acad Sci U S A* 2004, **101**(19):7329-7334.
55. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E *et al*: **Genome evolution in yeasts**. *Nature* 2004, **430**(6995):35-44.
56. ***Debaryomyces hansenii* : Chromosomes**.
57. ***Saccharomyces* Genome Database** [ftp://ftp.yeastgenome.org/yeast/]
58. Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF: **MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics**. *Nucleic Acids Res* 2004, **32**(Database issue):D373-376.
59. **Mouse Genome Sequencing** [http://www.ncbi.nlm.nih.gov/genome/seq/MmHome.html ftp://ftp.ncbi.nih.gov/genomes/M_musculus]
60. Grumbling G, Strelets V: **FlyBase: anatomical data, images and queries**. *Nucleic Acids Res* 2006, **34**(Database issue):D484-488.
61. Min J, Zhang X, Cheng X, Grewal SI, Xu RM: **Structure of the SET domain histone lysine methyltransferase Clr4**. *Nat Struct Biol* 2002, **9**(11):828-832.

62. Aasland R, Gibson TJ, Stewart AF: **The PHD finger: implications for chromatin-mediated transcriptional regulation**. *Trends Biochem Sci* 1995, **20**(2):56-59.
63. Torii KU, Stoop-Myer CD, Okamoto H, Coleman JE, Matsui M, Deng XW: **The RING finger motif of photomorphogenic repressor COP1 specifically interacts with the RING-H2 motif of a novel Arabidopsis protein**. *J Biol Chem* 1999, **274**(39):27674-27681.
64. Zeng L, Zhou MM: **Bromodomain: an acetyl-lysine binding domain**. *FEBS Lett* 2002, **513**(1):124-128.
65. Stec I, Nagl SB, van Ommen GJ, den Dunnen JT: **The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation?** *FEBS Lett* 2000, **473**(1):1-5.
66. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae**. *Nature* 2004, **428**(6983):617-624.
67. Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, Reuter G, Reinberg D, Jenuwein T: **A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin**. *Genes Dev* 2004, **18**(11):1251-1262.
68. Takeuchi T, Yamazaki Y, Katoh-Fukui Y, Tsuchiya R, Kondo S, Motoyama J, Higashinakagawa T: **Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation**. *Genes Dev* 1995, **9**(10):1211-1222.
69. Gross CT, McGinnis W: **DEAF-1, a novel protein that binds an essential region in a Deformed response element**. *Embo J* 1996, **15**(8):1961-1970.
70. Lutterbach B, Westendorf JJ, Linggi B, Patten A, Moniwa M, Davie JR, Huynh KD, Bardwell VJ, Lavinsky RM, Rosenfeld MG *et al*: **ETO, a target of t(8;21) in acute leukemia, interacts with the N-CoR and mSin3 corepressors**. *Mol Cell Biol* 1998, **18**(12):7176-7184.
71. Lutterbach B, Sun D, Schuetz J, Hiebert SW: **The MYND motif is required for repression of basal transcription from the multidrug resistance 1 promoter by the t(8;21) fusion protein**. *Mol Cell Biol* 1998, **18**(6):3604-3611.
72. D'Andrea LD, Regan L: **TPR proteins: the versatile helix**. *Trends Biochem Sci* 2003, **28**(12):655-662.
73. Biedenkapp H, Borgmeyer U, Sippel AE, Klempnauer KH: **Viral myb oncogene encodes a sequence-specific DNA-binding activity**. *Nature* 1988, **335**(6193):835-837.
74. Aasland R, Stewart AF, Gibson T: **The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB**. *Trends Biochem Sci* 1996, **21**(3):87-88.
75. Wolfe K: **Evolutionary genomics: yeasts accelerate beyond BLAST**. *Curr Biol* 2004, **14**(10):R392-394.
76. Houghton-Larsen J, Brandt A: **Fermentation of high concentrations of maltose by Saccharomyces cerevisiae is limited by the COMPASS methylation complex**. *Appl Environ Microbiol* 2006, **72**(11):7176-7182.

Appendix A

Multiple sequence alignment of the conserved and SET-I regions
CLUSTAL X (1.82) multiple alignment of the entire SET-domain region of five sample sequences. The SET-domain region with the conserved catalytic and binding sites highlighted in yellow and the insert region highlighted with turquoise.

```
NP031997Mm          ---------SD-VAGWGIFIKDP----VQKNEFISEYCGEIISQDEDDRRGKVYDKYMCS
NP595186Sp          --LPLEIFKTK-EKGWGVRSLRF----APAGTFITCYLGEVITSAEAAKRDKNYDDDGIT
NP594980Sp          --AKVDVFLTE-KKGFGLRADAN----LPKDTFVYEYIGEVIPEQKFRKRMRQYDSEGIK
EAA30745Nc          KWLEIRHTGTA-EKGYGVFVKTNGRYEIPKGAYLGHYVGEIIPG--PTKGYNANNKYLYE
NP012430Sc          --ADIEVRKSSNERDFGVFAADS----CVKGELIQEYLGKIDFQKNYQTDPNNDYRLMGT

NP031997Mm          FLFNLN-----NDFVVDATRKGNKIRFANHSVNPNCYAKVMMVNGDHR---------IGI
NP595186Sp          YLFDLDMFDDASEYTVDAQNYGDVSRFFNHSCSPNIAIYSAVRNHGFRT-----IYDLAF
NP594980Sp          HFYFMMLQ---KGEYIDATKRGSLARFCNHSCRPNCYVDKWMVGDKLR---------MGI
EAA30745Nc          PQVGLEWDT--DRPIIDAGKMGNWTRFMNSSCDPNVSESIMQIG-KVR--------VIAF
NP012430Sc          TKPKVLFHP-HWPLYIDSRETGGLTRYIRRSCEPNVELVTVRPLDEKPRGDNDCRVKFVL

NP031997Mm          FAKRAIQTGEELFFDYRYSQAD-
NP595186Sp          FAIKDIQPLEELTFDYAGAKDFS
NP594980Sp          FCKRDIIRGEELTFDYNVDRYGA
EAA30745Nc          FANKTLKSGDELCIYYGDDYFRG
NP012430Sc          RAIRDIRKGEEISVEWQWDLRNP
```

Appendix B

Table B.1: SET-domain query sequences used to search new SET-domain proteins

Accession numbers[a]	Species	References
Su(var)3-9		
NP726483	*Schizosaccharomyces pombe*	Alvarez and Avramova (2002) Gene 285:25-37
NP524357	*Drosophila melanogaster*	Alvarez and Avramova (2002) Gene 285:25-37
NP665829	*Mus musculus*	Alvarez and Avramova (2002) Gene 285:25-37
AAK28966	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
AAD10665	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
E(z)		
AAC39446	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
AAC27124	*Caenorhabditis elegans*	Alvarez and Avramova (2002) Gene 285:25-37
NP031997	*Mus musculus*	Alvarez and Avramova (2002) Gene 285:25-37
Trithorax		
NP011987	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev. 15:2991-3004
AAF29390	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
BAA97320	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
BAB10481	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
NP589812	*Schizosaccharomyces pombe*	Alvarez and Avramova (2002) Gene 285:25-37
XP11067	*Mus musculus*	Alvarez and Avramova (2002) Gene 285:25-37
SET 2		
NP012367	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev.15:2991-3004
AAC23419	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
CAA18207	*Arabidopsis thaliana*	Baumbusch et. al. (2001) Nucleic Acids Res. Vol. 29 No. 21 4319-4333
NP524160	*Drosophila melanogaster*	Schotta et. al. (2002) Genes Dev. 18 (11): 1251-1262

Table continued

SET 3		
NP012954	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev.15:2991-3004
SET 4		
NP012430	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev. 15:2991-3004
SET 5		
P38890	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev. 15:2991-3004
SET 6		
NC001148	*Saccharomyces cerevisiae*	Pijnappel et. al. (2001) Genes Dev. 15:2991-3004
SET 8		
AAH50346	*Homo sapiens*	Fang et. al. (2002) Curr. Biol.12**:** 1086-1099
SET 7/9		
Q8WTS6	*Homo sapiens*	Nishioka et. al. (2002) Genes Dev. 16:479-489

Appendix C

Table C.1: SET-domain proteins used for building the Profile Hidden Markov model

SET-domain family	Accession numbers[a]	Species
E(z)	NP031996	*Mus musculus*
E(z)	AAC39446	*Arabidopsis thaliana*
E(z)	AAF00642	*Arabidopsis thaliana*
SET 1	NP587812	*Schizosaccharomyces pombe*
SET 1	NP011987	*Saccharomyces cerevisiae*
SET 1	NP808249	*Mus musculus*
SET 1	CAB71104	*Arabidopsis thaliana*
SET 1	NP726773	*Drosophila melanogaster*
SET 2	NP032765	*Mus musculus*
SET 2	NP012367	*Saccharomyces cerevisiae*
SET 2	CAA18207	*Arabidopsis thaliana*
SET 2	NP619620	*Mus musculus*
SET 3	NP012954	*Saccharomyces cerevisiae*
SET 3/4	NP594837	*Schizosaccharomyces pombe*
SET 4	NP012430	*Saccharomyces cerevisiae*
SET 5	P38890	*Saccharomyces cerevisiae*
SET6	NP001148	*Saccharomyces cerevisiae*
SET 7/9	NP542983	*Mus musculus*
SET 8	NP081464	*Mus musculus*
Su(var)3-9	NP726483	*Drosophila melanogaster*
Su(var)3-9	AAK28966	*Arabidopsis thaliana*
Su(var)3-9	NP524357	*Drosophila melanogaster*
Su(var)3-9	NP595186	*Schizosaccharomyces pombe*
Su(var)3-9	EAA28243	*Neurospora crassa*
Unknown	EAK94622	*Candida albicans*
Unknown	EAA58228	*Aspergillus nidulans*
Unknown	EAA71535	*Fusarium graminearum*

[a]All accession numbers are from National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/).

Appendix D

Maximum likelihood phylogeny reconstructed using 183 non-redundant SET-domain sequences. Red font indicates representative sequences used to reconstruct the primary phylogeny for analysis. Sequences with known SET-domain family affiliation are indicated with blue font. SET5/6 'Super' group indicated by arrow.

.

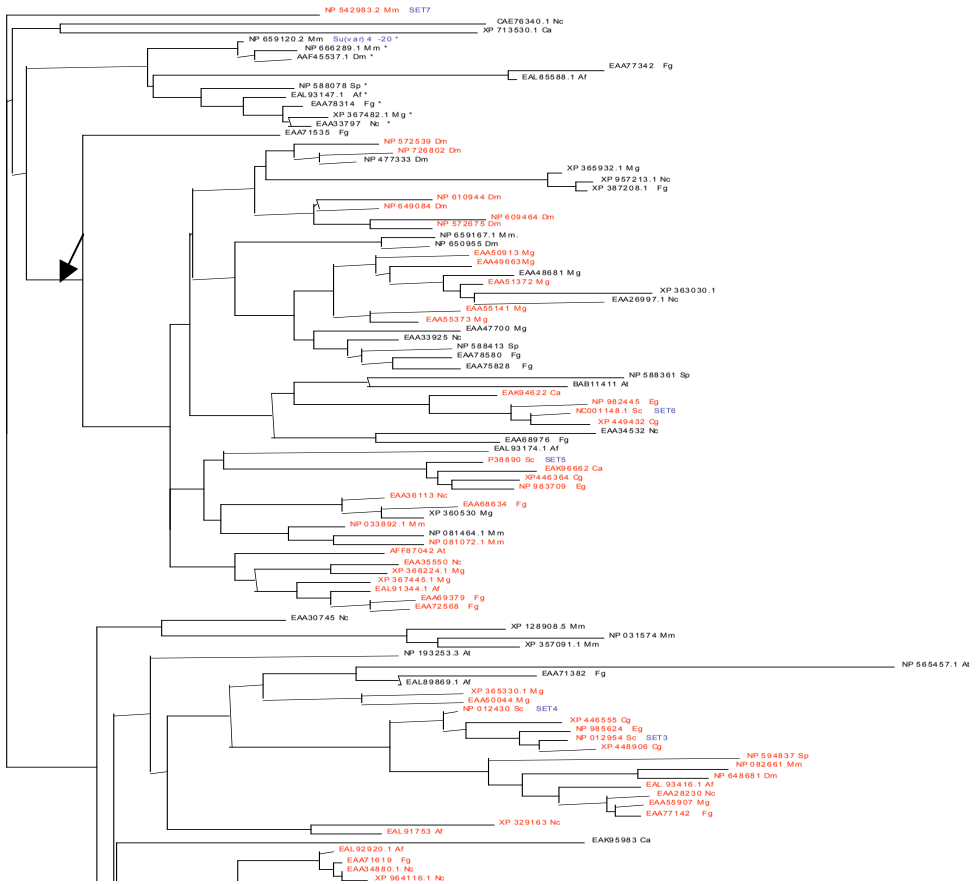
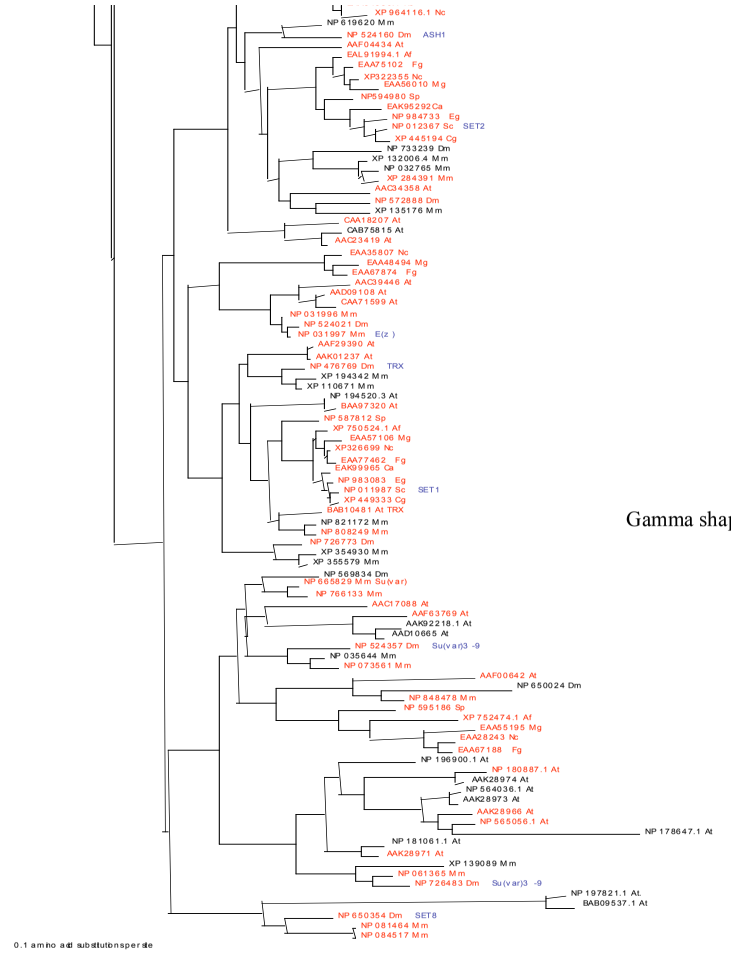


Figure continued…

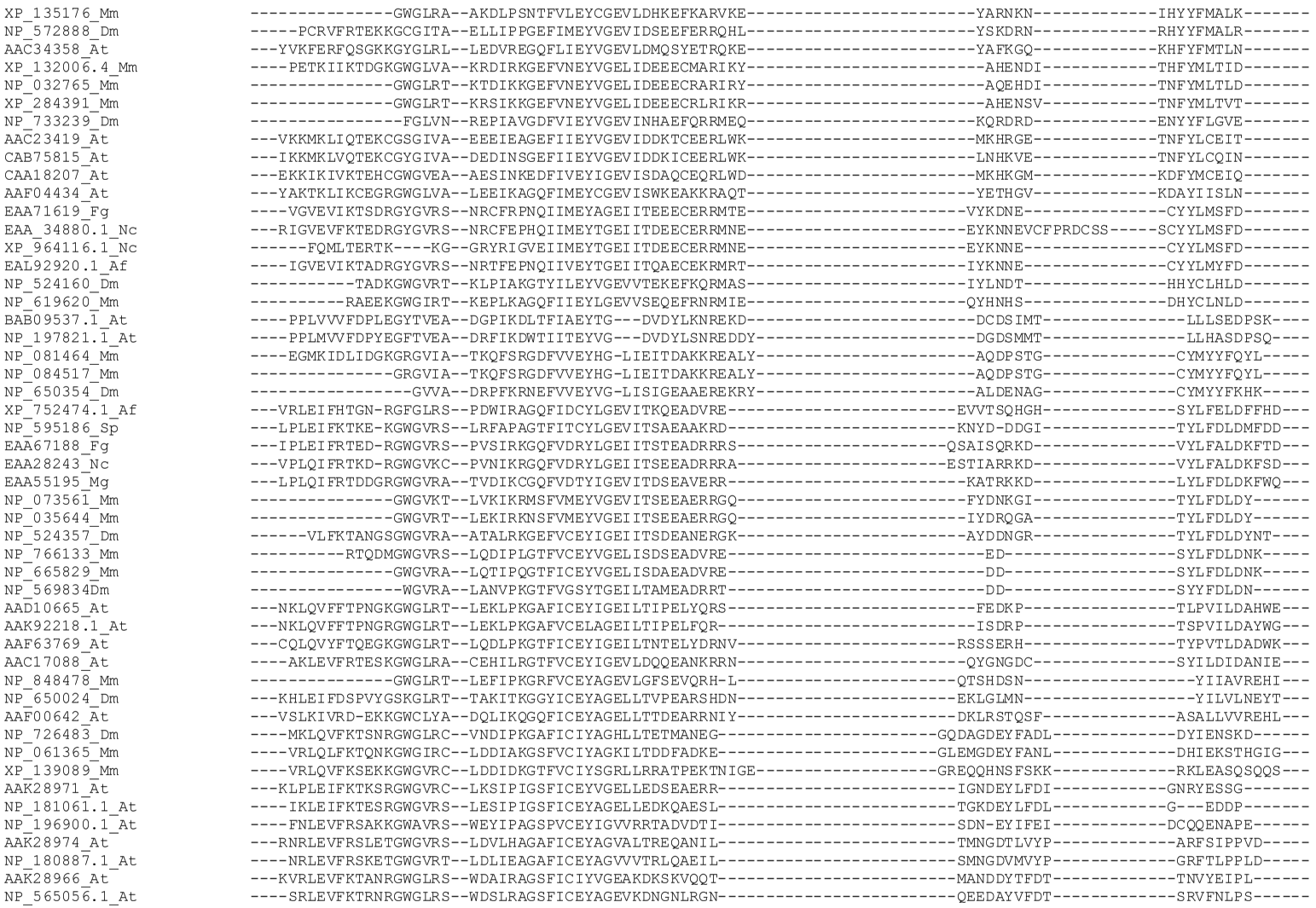


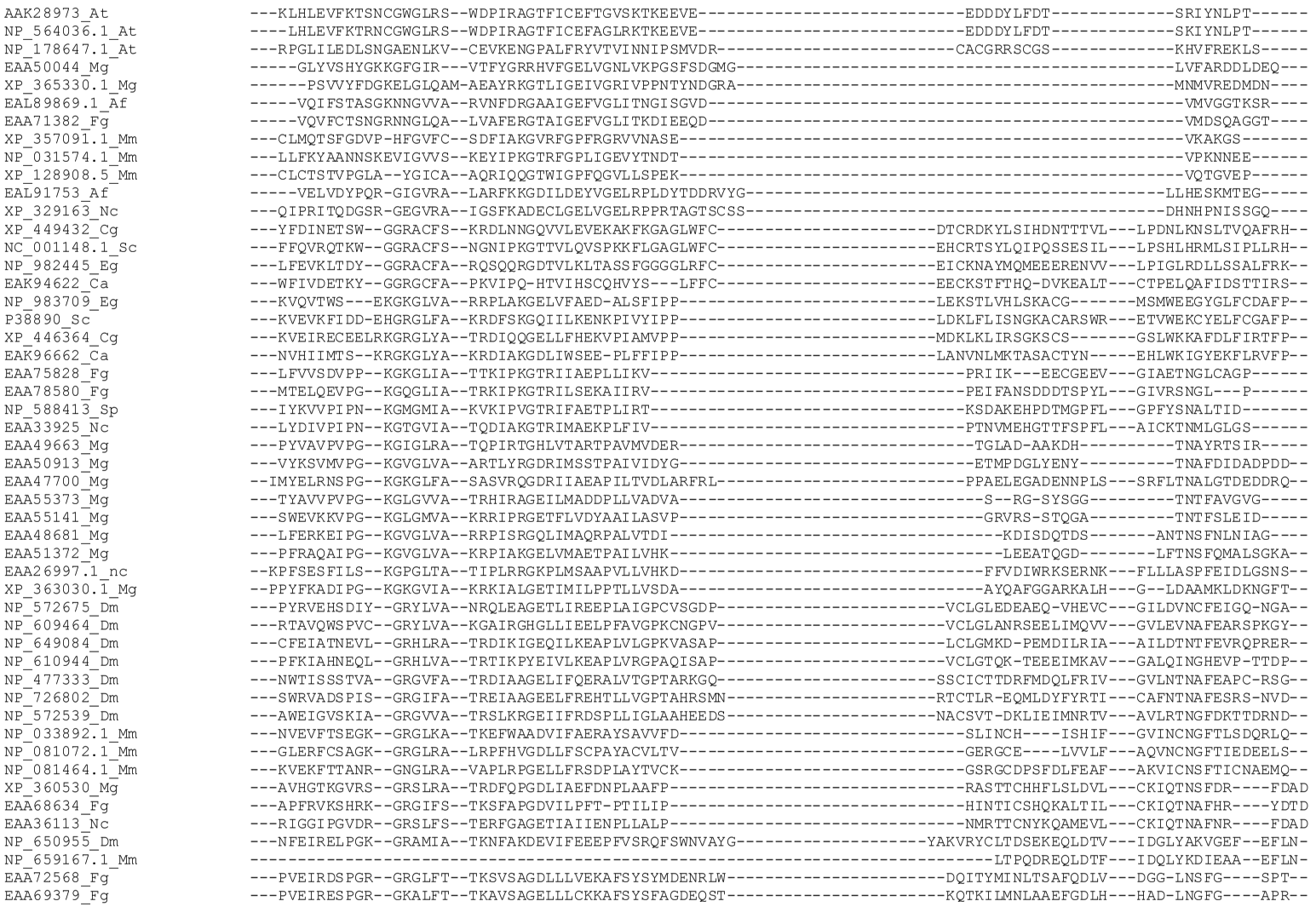
Gamma shape parameter = .790

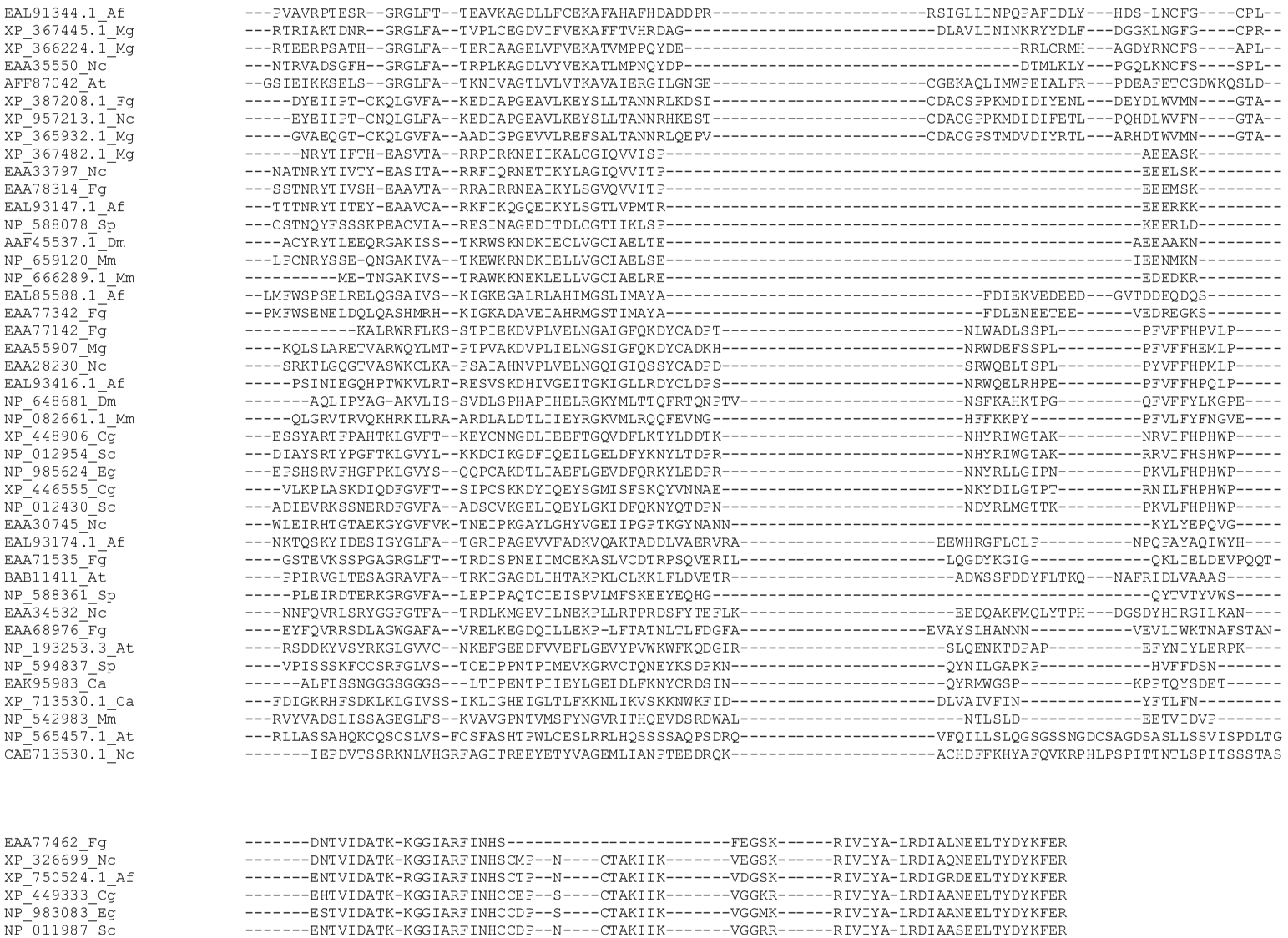
Appendix E

CLUSTAL X (1.83) multiple alignment of all 183 non-redundant SET-domain hits.

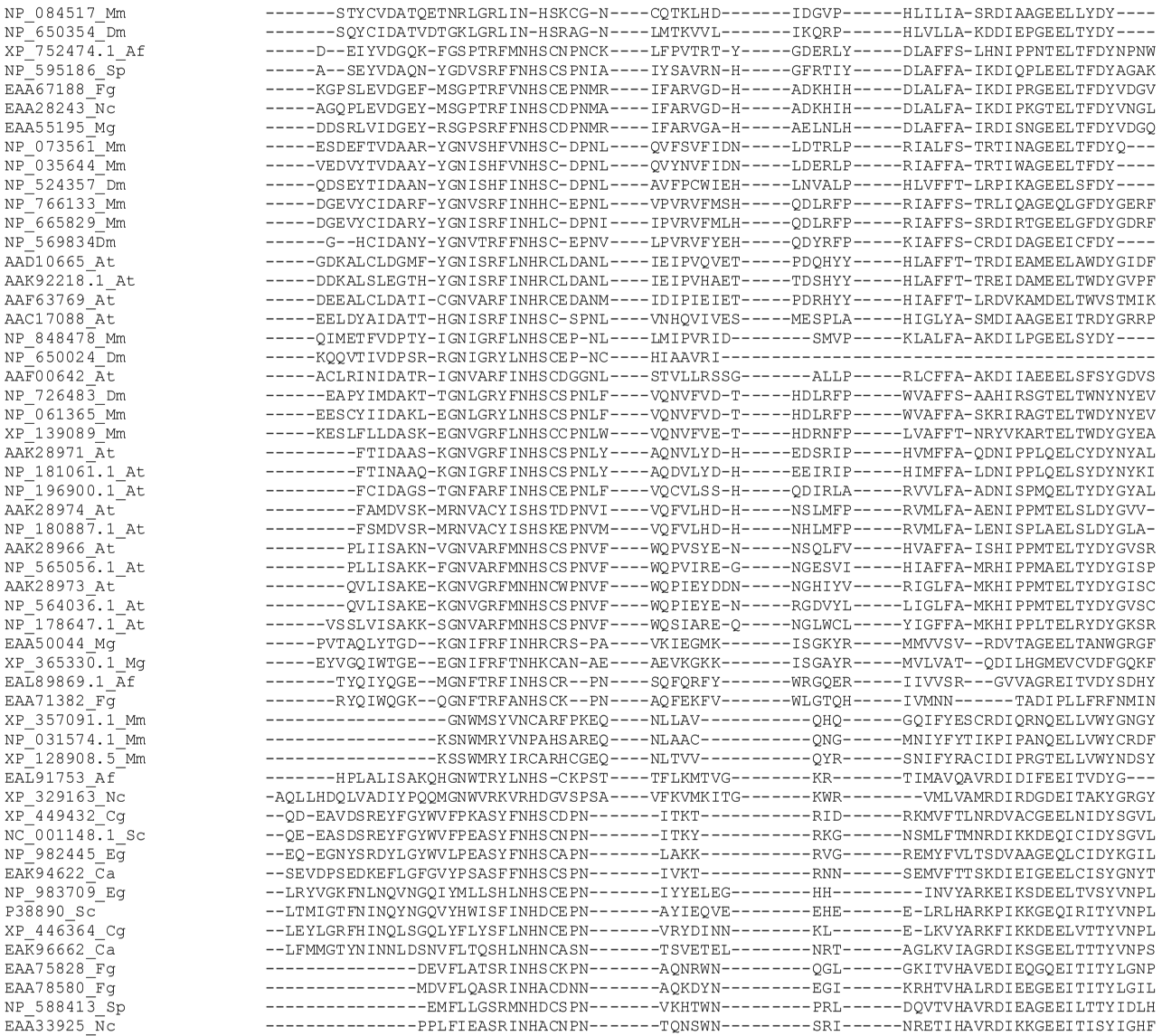
```
EAA77462_Fg          ---PVKFARSAIH--NWGLYA--MENIAKDDMIIEYVGEQVRQQISEIRENR----------------------YLKSGIGS-------------SYLFRID-------
XP_326699_Nc         ---PVKFARSAIH--NWGLYA--MENINKDDMIIEYVGEEVRQQIAELREAR----------------------YLKSGIGS-------------SYLFRID-------
XP_750524.1_Af       ---PVRFARSAIH--NWGLYA--EENISANDMIIEYVGEKVRQQVADMRERQ----------------------YLKSGIGS-------------SYLFRID-------
XP_449333_Cg         ---PVTFARSAIH--NWGLYA--LEPINAKEMVIEYVGERIRQPVAEMRERR----------------------YIKNGIGS-------------SYLFRID-------
NP_983083_Eg         ---PVTFARSAIH--NWGLYA--LEPISAKEMIIEYVGERIRQPVAEMREKR----------------------YLKSGIGS-------------SYLFRVD-------
NP_011987_Sc         ---PVMFARSAIH--NWGLYA--LDSIAAKEMIIEYVGERIRQPVAEMREKR----------------------YLKNGIGS-------------SYLFRVD-------
EAK99965_Ca          ---PVTFARSAIH--NWGLYA--MEPIAAKEMIIEYVGERIRQQVAEHREKS----------------------YLKTGIGS-------------SYLFRID-------
EAA57106_Mg          ---PVKFERSAIH--NWGLYA--MEHIPKDDMIIEYVGEEVRPSVAQVREAR----------------------YDRSGIGS-------------SYLFRID-------
NP_808249_Mm         --------KSHIH--DWGLFA--MEPIAADEMVIEYVGQNIRQVIADMREKR----------------------YEDEGIGS-------------SYMFRVD-------
NP_821172_Mm         --------RSRIH--EWGLFA--MEPIAADEMVIEYVGQNIRQMVADMREKR----------------------YVQEGIGS-------------SYLFRVD-------
NP_587812_Sp         ---QLHFGPSRIH--TLGLFA--MENIDKNDMVIEYIGEIIRQRVADNREKN----------------------YVREGIGD-------------SYLFRID-------
BAB10481_At          ---HLRFQQSKIH--DWGLVA--LEPIEAEDFVIEYVGELIRSSISEIRERQ----------------------YEKMGIGS-------------SYLFRLD-------
XP_110671_Mm         --------RSPIH--GRGLFC--KRNIDAGEMVIEYAGNVIRSIQTDKREKY----------------------YDSKGIG--------------CYMFRID-------
XP_194342_Mm         --------RSAIH--GRGLFC--KRNIDAGEMVIEYSGIVIRSVLTDKREKF----------------------YDGKGIG--------------CYMFRMD-------
NP_476769_Dm         --------RSHIH--GRGLYC--TKDIEAGEMVIEYAGELIRSTLTDKRERY----------------------YDSRGIG--------------CYMFKID-------
XP_355579_Mm         --------RSRIQ--GLGLYA--ARDIEKHTMVIEYIGTIIRNEVANRKEKL----------------------YESQNRG-------------VYMFRMD-------
XP_354930_Mm         --------RSRIQ--GLGLYA--AKDLEKHTMVIEYIGTIIRNEVANRREKI----------------------YEEQNRG-------------IYMFRIN-------
NP_726773_Dm         ----VYLARSKIQ--GLGLYA--ARDIEKHTMIIEYIGEVIRTEVSEIREKQ----------------------YESKNRG-------------IYMFRLD-------
AAK01237_At          ---RLAFGKSGIH--GFGIFA--KLPHRAGDMMIEYTGELVRPSIADKREQL----------------------IYNSM-VGAG-----------TYMFRID-------
AAF29390_At          ---RLAFGKSGIH--GFGIFA--KLPHRAGDMVIEYTGELVRPPIADKREHL----------------------IYNSM-VGAG-----------TYMFRID-------
BAA97320_At          ---RVCFGRSGIH--GWGLFA--RRNIQEGEMVLEYRGEQVRGIIADLREAR----------------------YRR--EGKD-----------CYLFKIS-------
NP_194520_At         ---RVCFGRSGIH--GWGLFA--RRNIQEGEMVLEYRGEQVRGSIADLREAR----------------------YRR--VGKD-----------CYLFKIS-------
CAA71599_At          ---RVLLGISDIS--GWGAFL--KNSVSKHEYLGEYTGELISHKE--ADKRG----------------------KIYD-RENC-------------SFLFNLN-------
AAD09108_At          ---RILLGKSDVA--GWGAFL--KNSVSKNEYLGEYTGELISHHE--ADKRG----------------------KIYD-RANS-------------SFLFDLN-------
NP_031996_Mm         ---------SDVA--GWGTFI--KESVQKNEFISEYCGELISQDE--ADRRG----------------------KVYD-KYMS-------------SFLFNLN-------
NP_031997_Mm         ---------SDVA--GWGIFI--KDPVQKNEFISEYCGEIISQDE--DDRRG----------------------KVYD-KYMC-------------SFLFNLN-------
NP_524021_Dm         ---HLLMAPSDIA--GWGIFL--KEGAQKNEFISEYCGEIISQDE--ADRRG----------------------KVYD-KYMC-------------SFLFNLN-------
AAC39446_At          ---KILIGKSDVH--GWGAFT--WDSLKKNEYLGEYTGELITHDE--ANERG----------------------RIED-RIGS-------------SYLFTLN-------
EAA67874_Fg          ---GLALGQSQLEGVGYGLFT--VEPIAQDDFIIEYVGELITHDEGVEARRG----------------------DVFDEESNI-------------SYVFTLLEN------
EAA35807_Nc          ----VLLGKSQLEGCGYGLFT--AEDISQDEFVIEYTGELITHDEGVEARRG----------------------EGFGSQGTS-------------SYLFTLLEH-----
EAA48494_Mg          ---PLCLGESGIEGCGYGLFT--AVDIAADEFIIEYVGELIQHDEGVEARRG----------------------NVFDEESNV-------------SYLFTLLED-----
EAA75102_Fg          ----ANVSVIKTEKKGFGLRA--DSDLQPNDFVFEYIGEVINEPTFRRRMIQ----------------------YDEEGI-------------KHFYFMSLN-------
XP_322355_Nc         ----ADVSVIKTEKKGFGLRA--NTDLQVNDFIFEYIGEVINEPTFRSRMVK----------------------YDKEGI-------------KHFYFMSLT-------
EAL91994.1_Af        ----ANVAVIKTEKKGFGLRA--ETDLRPHQFIFEYVGEVINEAQFRRRMRQ----------------------YDEEGI-------------KHFYFMSLS-------
EAA56010_Mg          ----ANVSVIKTENKGYGLRA--DANLEPNDFVFEYIGEVIGEELFRSRLMK----------------------YDTQRL-------------EHFYFMSLT-------
NP_594980_Sp         ----AKVDVFLTEKKGFGLRA--DANLPKDTFVYEYIGEVIPEQKFRKRMRQ----------------------YDSEGI-------------KHFYFMMLQ-------
XP_445194_Cg         ----ANIAVFKTKMKGYGVRA--ESDIEINDFIYEYKGEVIEEEEFRDRLVD----------------------YDQKKF-------------RHFYFMMLQ-------
NP_012367_Sc         ----APIAIFKTKHKGYGVRA--EQDIEANQFIYEYKGEVIEEMEFRDRLID----------------------YDQRHF-------------KHFYFMMLQ-------
NP_984733_Eg         ----ADIAVFQTEKKGYGVRA--ERDIEANEFIYEYIGEVISEADFRDRMVD----------------------YDMRGF-------------KHFYFMMLQ-------
EAK95292_Ca          ----SKVKVIQTELKGYGLIA--EQDIEENQFIYEYIGEVIDEISFRQRMIE----------------------YDLRHL-------------KHFYFMMLS-------
```


XP_135176_Mm
NP_572888_Dm
AAC34358_At
XP_132006.4_Mm
NP_032765_Mm
XP_284391_Mm
NP_733239_Dm
AAC23419_At
CAB75815_At
CAA18207_At
AAF04434_At
EAA71619_Fg
EAA_34880.1_Nc
XP_964116.1_Nc
EAL92920.1_Af
NP_524160_Dm
NP_619620_Mm
BAB09537.1_At
NP_197821.1_At
NP_081464_Mm
NP_084517_Mm
NP_650354_Dm
XP_752474.1_Af
NP_595186_Sp
EAA67188_Fg
EAA28243_Nc
EAA55195_Mg
NP_073561_Mm
NP_035644_Mm
NP_524357_Dm
NP_766133_Mm
NP_665829_Mm
NP_569834Dm
AAD10665_At
AAK92218.1_At
AAF63769_At
AAC17088_At
NP_848478_Mm
NP_650024_Dm
AAF00642_At
NP_726483_Dm
NP_061365_Mm
XP_139089_Mm
AAK28971_At
NP_181061.1_At
NP_196900.1_At
AAK28974_At
NP_180887.1_At
AAK28966_At
NP_565056.1_At


AAK28973_At
NP_564036.1_At
NP_178647.1_At
EAA50044_Mg
XP_365330.1_Mg
EAL89869.1_Af
EAA71382_Fg
XP_357091.1_Mm
NP_031574.1_Mm
XP_128908.5_Mm
EAL91753_Af
XP_329163_Nc
XP_449432_Cg
NC_001148.1_Sc
NP_982445_Eg
EAK94622_Ca
NP_983709_Eg
P38890_Sc
XP_446364_Cg
EAK96662_Ca
EAA75828_Fg
EAA78580_Fg
NP_588413_Sp
EAA33925_Nc
EAA49663_Mg
EAA50913_Mg
EAA47700_Mg
EAA55373_Mg
EAA55141_Mg
EAA48681_Mg
EAA51372_Mg
EAA26997.1_nc
XP_363030.1_Mg
NP_572675_Dm
NP_609464_Dm
NP_649084_Dm
NP_610944_Dm
NP_477333_Dm
NP_726802_Dm
NP_572539_Dm
NP_033892.1_Mm
NP_081072.1_Mm
NP_081464.1_Mm
XP_360530_Mg
EAA68634_Fg
EAA36113_Nc
NP_650955_Dm
NP_659167.1_Mm
EAA72568_Fg
EAA69379_Fg


EAL91344.1_Af
XP_367445.1_Mg
XP_366224.1_Mg
EAA35550_Nc
AFF87042_At
XP_387208.1_Fg
XP_957213.1_Nc
XP_365932.1_Mg
XP_367482.1_Mg
EAA33797_Nc
EAA78314_Fg
EAL93147.1_Af
NP_588078_Sp
AAF45537.1_Dm
NP_659120_Mm
NP_666289.1_Mm
EAL85588.1_Af
EAA77342_Fg
EAA77142_Fg
EAA55907_Mg
EAA28230_Nc
EAL93416.1_Af
NP_648681_Dm
NP_082661.1_Mm
XP_448906_Cg
NP_012954_Sc
NP_985624_Eg
XP_446555_Cg
NP_012430_Sc
EAA30745_Nc
EAL93174.1_Af
EAA71535_Fg
BAB11411_At
NP_588361_Sp
EAA34532_Nc
EAA68976_Fg
NP_193253.3_At
NP_594837_Sp
EAK95983_Ca
XP_713530.1_Ca
NP_542983_Mm
NP_565457.1_At
CAE713530.1_Nc
EAA77462_Fg
XP_326699_Nc
XP_750524.1_Af
XP_449333_Cg
NP_983083_Eg
NP_011987_Sc

```
EAK99965_Ca          -------DNTVIDATK-KGGIARFINHCCSP--S----CTAKIIK-------VEGKK------RIVIYA-LRDIEANEELTYDYKFER
EAA57106_Mg          -------EDAVIDATK-KGGIARFINHSCMP--N----CTAKIIR-------VEGTK------RIVIYA-LRDIARNEELTYDYKFEL
NP_808249_Mm         -------HDTIIDATK-CGNFARFINHSCNP--N----CYAKVIT-------VESQK------KIVIYS-KQHINVNEEITYDYKFPI
NP_821172_Mm         -------HDTIIDATK-CGNLARFINHCCTP--N----CYAKVIT-------IESQK------KIVIYS-KQPIGVDEEITYDYKFPL
NP_587812_Sp         -------EDVIVDATK-KGNIARFINHSCAP--N----CIARIIR-------VEGKR------KIVIYA-DRDIMHGEELTYDYKFPE
BAB10481_At          -------DGYVLDATK-RGGIARFINHSCEP--N----CYTKIIS-------VEGKK------KIFIYA-KRHIDAGEEISYNYKFPL
XP_110671_Mm         -------DSEVVDATM-HGNAARFINHSCEP--N----CYSRVIN-------IDGQK------HIVIFA-MRKIYRGEELTYDYKFPI
XP_194342_Mm         -------DFDVVDATM-HGNAARFINHSCEP--N----CFSRVIH-------VEGQK------HIVIFA-LRRILRGEELTYDYKFPI
NP_476769_Dm         -------DNLVVDATM-RGNAARFINHCCEP--N----CYSKVVD-------ILGHK------HIIIFA-LRRIVQGEELTYDYKFPF
XP_355579_Mm         -------NDHVIDATL-TGGPARYINHSCAP--N----CVAEVVT-------FERGH------KIIISS-NRRIQKGEELCYDYKFDF
XP_354930_Mm         -------NEHVIDATL-TGGPA-----------------AEVVT-------FDKED------KIIIIS-SRRIPKGEELTYDYQFDF
NP_726773_Dm         -------EDRVVDATL-SGGLARYINHSCNP--N----CVTEIVE-------VDRDV------RIIIFA-KRKIYRGEELSYDYKFDI
AAK01237_At          -------DERVIDATR-TGSIAHLINHSCVP--N----CYSRVIT-------VNGDE------HIIIFA-KRHIPKWEELTYDYRFFS
AAF29390_At          -------NERVIDATR-TGSIAHLINHSCEP--N----CYSRVIS-------VNGDE------HIIIFA-KRDVAKWEELTYDYRFFS
BAA97320_At          -------EEVVVDATE-KGNIARLINHSCMP--N----CYARIMS-------VGDDES-----RIVLIA-KTTVASCEELTYDYLFDP
NP_194520_At         -------EEVVVDATD-KGNIARLINHSCTP--N----CYARIMS-------VGDEES-----RIVLIA-KANVAVGEELTYDYLFDP
CAA71599_At          -------DQFVLDAYR-KGDKLKFANHSPEP--N----CYAKVIM-------VAGDH------RVGIFA-KERILAGEELFYDYRYEP
AAD09108_At          -------DQYVLDAQR-KGDKLKFANHSAKP--N----CYAKVMF-------VAGDH------RVGIFA-NERIEASEELFYDYRYGP
NP_031996_Mm         -------NDFVVDATR-KGNKIRFANHSVNP--N----CYAKVVM-------VNGDH------RIGIFA-KRAIQAGEELFFDYRYSQ
NP_031997_Mm         -------NDFVVDATR-KGNKIRFANHSVNP--N----CYAKVMM-------VNGDH------RIGIFA-KRAIQTGEELFFDYRYSQ
NP_524021_Dm         -------NDFVVDATR-KGNKIRFANHSINP--N----CYAKVMM-------VTGDH------RIGIFA-KRAIQPGEELFFDYRY--
AAC39446_At          -------DQLEIDARR-KGNEFKFLNHSARP--N----CYAKLMI-------VRGDQ------RIGLFA-ERAIEEGEELFFDYCYGP
EAA67874_Fg          -------EGIWVDAAT-YGNLSRYINHASESDKN----ITPRILY-------VNGEY------RIKFTA-MRDIAAGEELFFNYGENF
EAA35807_Nc          -------EGIWVDAAM-YGNLSRYINHASENDKN----ITPKIIY-------VNNEY------RIKFTA-LRDIKAGEELFFNYGDNF
EAA48494_Mg          -------DGIWVDAAV-YGNLSRYMNHASESDRN----VVPKIVQ-------VNGDF------RIRFTA-LRDIKAGEELFFNYGENF
EAA75102_Fg          -------KSEFVDATK-KGNYGRFCNHSCNP--N----CYVDKWV-------VGDKL------RMGIFT-SRKIQSGEELVFNYNVDR
XP_322355_Nc         -------KSEFVDATK-KGNLGRFCNHSCDP--N----CYVDKWV-------VGDKL------RMGIFA-GRAIKAGEELVFNYNVDR
EAL91994.1_Af        -------RGEFVDATK-KGNLGRFCNHSCNP--N----CYVDKWV-------VGEKL------RMGIFA-ERAIQAGEELVFNYNVDR
EAA56010_Mg          -------RTEYVDATK-KGNLGRFCNHSCNP--N----CYVDKWV-------VGDKL------RMGIFA-MRAIKAGEELCFNYNVDR
NP_594980_Sp         -------KGEYIDATK-RGSLARFCNHSCRP--N----CYVDKWM-------VGDKL------RMGIFC-KRDIIRGEELTFDYNVDR
XP_445194_Cg         -------SGEFIDATI-KGSLARFCNHSCNP--N----AYVNKWV-------VAGKL------RMGIFA-KRKILKGEEITFDYNVDR
NP_012367_Sc         -------NGEFIDATI-KGSLARFCNHSCSP--N----AYVNKWV-------VKDKL------RMGIFA-QRKILKGEEITFDYNVDR
NP_984733_Eg         -------AGEFIDATE-RGCLARFCNHSCNP--N----AYVSKWD-------VAGKL------KMGIFA-HRKILKGEEITFDYNVDR
EAK95292_Ca          -------NDSFIDATE-KGSLGRFINHSCNP--N----AFVDKWH-------VGDRL------RMGIFA-KRKISRGEEITFDYNVDR
XP_135176_Mm         -------NDEIIDATQ-KGNCSRFMNHSCEP--N----CETQKWT-------VNGQL------RVGFFT-TKLVPSGSELTFDYQF--
NP_572888_Dm         -------GEAVIDATS-KGNISRYINHSCDP--N----AETQKWT-------VNGEL------RIGFFS-VKPIQPGEEITFDYQYLR
AAC34358_At          -------GNEVIDAGA-KGNLGRFINHSCEP--N----CRTEKWM-------VNGEI------CVGIFS-MQDLKKGQELTFDYNYVR
XP_132006.4_Mm       -------KDRIIDAGP-KGNYSRFMNHSCQP--N----CETLKWT-------VNGDT------RVGLFA-VCDIPAGTELTFNYNLDC
NP_032765_Mm         -------KDRIIDAGP-KGNYARFMNHCCQP--N----CETQKWS-------VNGDT------RVGLFA-LSDIKAGTELTFNY----
XP_284391_Mm         -------KDRIIDAGP-KGNYSRFMNHSCNP--N----CETQKWT-------VNGDV------RVGLFA-LCDIPAGMELTFNY----
NP_733239_Dm         -------KDFIIDAGP-KGNLARFMNHSCEP--N----CETQKWT-------VNCIH------RVGIFA-IKDIPVNSELTFNYLWD-
AAC23419_At          -------RDMVIDATH-KGNKSRYINHSCNP--N----TQMQKWI-------IDGET------RIGIFA-TRGIKKGEHLTYDYQFVQ
CAB75815_At          -------WNMVIDATH-KGNKSRYINHSCSP--N----TEMQKWI-------IDGET------RIGIFA-TRFINKGEQLTYDYQFVQ
CAA18207_At          -------KDFTIDATF-KGNASRFLNHSCNP--N----CVLEKWQ-------VEGET------RVGVFA-ARQIEAGEPLTYDYRFVQ
AAF04434_At          -------ASEAIDATK-KGSLARFINHSCRP--N----CETRKWN-------VLGEV------RVGIFA-KESISPRTELAYDYNFEW
EAA71619_Fg          -------QNMIIDAT--TGSIARFVNHSCNP--N----CRMIKWI-------VSGQP------RMALFAGDKPIMTGDELTYDYNFDP
EAA_34880.1_Nc       -------QNMIIDAT--TGSIARFVNHSCSP--N----CRMIKWI-------VSGQP------RMALFAGDRPIQTGEELTYDYNFDP
XP_964116.1_Nc       -------QNMIIDAT--TGSIARFVNHSCSP--N----CRMIKWI-------VSGQP------RMALFAGDRPIQTGEELTYDYNFDP
EAL92920.1_Af        -------QNMIIDAT--RGSIARFVNHSCEP--N----CRMEKWT-------VAGKP------RMALFAGDHGIMTGEELTYDYNFDP
NP_524160_Dm         -------GGLVIDGQR-MGSDCRFVNHSCEP--N----CEMQKWS-------VNGLS------RMVLFA-KRAIEEGEELTYDYNF--
NP_619620_Mm         -------SGMVIDSYR-MGNEARFINHSCDP--N----CEMQKWS-------VNGVY------RIGLYA-LKDMPAGTELTYDYNF--
BAB09537.1_At        -------TVICPDKFGNISRFINGINNHNPVAKN----CKCVRYS-------INGEC------RVLLVA-TRDISKGERLYYDYNGYE
NP_197821.1_At       -------CVICPDRRSNIARFISGINNHSPEGRN----LKCVRFN-------INGEA------RVLLVA-NRDISKGERLYYDYNGYE
NP_081464_Mm         -------STYCVDATQETNRLGRLIN-HSKCG-N----CQTKLHD-------IDGVP------HLILIA-SRDIAAGEELLYDYGDRS
```

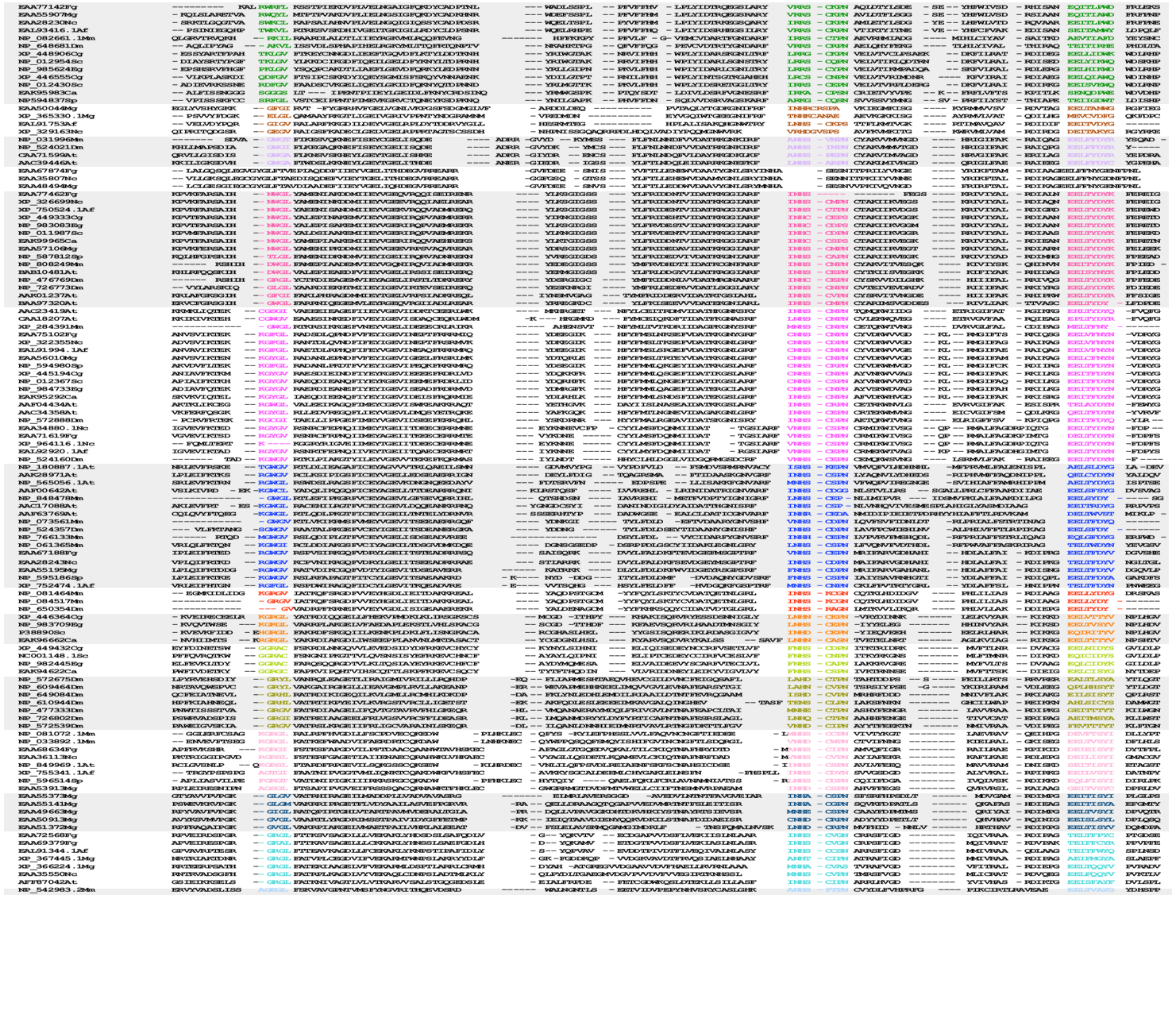

NP_084517_Mm -------STYCVDATQETNRLGRLIN-HSKCG-N----CQTKLHD-------IDGVP------HLILIA-SRDIAAGEELLYDY----
NP_650354_Dm -------SQYCIDATVDTGKLGRLIN-HSRAG-N----LMTKVVL-------IKQRP------HLVLLA-KDDIEPGEELTYDY----
XP_752474.1_Af -----D--EIYVDGQK-FGSPTRFMNHSCNPNCK----LFPVTRT-Y-----GDERLY-----DLAFFS-LHNIPPNTELTFDYNPNW
NP_595186_Sp -----A--SEYVDAQN-YGDVSRFFNHSCSPNIA----IYSAVRN-H-----GFRTIY-----DLAFFA-IKDIQPLEELTFDYAGAK
EAA67188_Fg -----KGPSLEVDGEF-MSGPTRFVNHSCEPNMR----IFARVGD-H-----ADKHIH-----DLALFA-IKDIPRGEELTFDYVDGV
EAA28243_Nc -----AGQPLEVDGEY-MSGPTRFINHSCDPNMA----IFARVGD-H-----ADKHIH-----DLALFA-IKDIPKGTELTFDYVNGL
EAA55195_Mg -----DDSRLVIDGEY-RSGPSRFFNHSCDPNMR----IFARVGA-H-----AELNLH-----DLAFFA-IRDISNGEELTFDYVDGQ
NP_073561_Mm -----ESDEFTVDAAR-YGNVSHFVNHSC-DPNL----QVFSVFIDN-----LDTRLP-----RIALFS-TRTINAGEELTFDYQ---
NP_035644_Mm -----VEDVYTVDAAY-YGNISHFVNHSC-DPNL----QVYNVFIDN-----LDERLP-----RIAFFA-TRTIWAGEELTFDY----
NP_524357_Dm -----QDSEYTIDAAN-YGNISHFINHSC-DPNL----AVFPCWIEH-----LNVALP-----HLVFFT-LRPIKAGEELSFDY----
NP_766133_Mm -----DGEVYCIDARF-YGNVSRFINHHC-EPNL----VPVRVFMSH-----QDLRFP-----RIAFFS-TRLIQAGEQLGFDYGERF
NP_665829_Mm -----DGEVYCIDARY-YGNISRFINHLC-DPNI----IPVRVFMLH-----QDLRFP-----RIAFFS-SRDIRTGEELGFDYGDRF
NP_569834Dm ------G--HCIDANY-YGNVTRFFNHSC-EPNV----LPVRVFYEH-----QDYRFP-----KIAFFS-CRDIDAGEEICFDY----
AAD10665_At -----GDKALCLDGMF-YGNISRFLNHRCLDANL----IEIPVQVET-----PDQHYY-----HLAFFT-TRDIEAMEELAWDYGIDF
AAK92218.1_At -----DDKALSLEGTH-YGNISRFINHRCLDANL----IEIPVHAET-----TDSHYY-----HLAFFT-TREIDAMEELTWDYGVPF
AAF63769_At -----DEEALCLDATI-CGNVARFINHRCEDANM----IDIPIEIET-----PDRHYY-----HIAFFT-LRDVKAMDELTWVSTMIK
AAC17088_At -----EELDYAIDATT-HGNISRFINHSC-SPNL----VNHQVIVES-----MESPLA-----HIGLYA-SMDIAAGEEITRDYGRRP
NP_848478_Mm -----QIMETFVDPTY-IGNIGRFLNHSCEP-NL----LMIPVRID--------SMVP-----KLALFA-AKDILPGEELSYDY----
NP_650024_Dm -----KQQVTIVDPSR-RGNIGRYLNHSCEP-NC----HIAAVRI--
AAF00642_At -----ACLRINIDATR-IGNVARFINHSCDGGNL----STVLLRSSG-------ALLP-----RLCFFA-AKDIIAEEELSFSYGDVS
NP_726483_Dm ------EAPYIMDAKT-TGNLGRYFNHSCSPNLF----VQNVFVD-T-----HDLRFP-----WVAFFS-AAHIRSGTELTWNYNYEV
NP_061365_Mm -----EESCYIIDAKL-EGNLGRYLNHSCSPNLF----VQNVFVD-T-----HDLRFP-----WVAFFA-SKRIRAGTELTWDYNYEV
XP_139089_Mm -----KESLFLLDASK-EGNVGRFLNHSCCPNLW----VQNVFVE-T-----HDRNFP-----LVAFFT-NRYVKARTELTWDYGYEA
AAK28971_At ---------FTIDAAS-KGNVGRFINHSCSPNLY----AQNVLYD-H-----EDSRIP-----HVMFFA-QDNIPPLQELCYDYNYAL
NP_181061.1_At ---------FTINAAQ-KGNIGRFINHSCSPNLY----AQDVLYD-H-----EEIRIP-----HIMFFA-LDNIPPLQELSYDYNYKI
NP_196900.1_At ---------FCIDAGS-TGNFARFINHSCEPNLF----VQCVLSS-H-----QDIRLA-----RVVLFA-ADNISPMQELTYDYGYAL
AAK28974_At ---------FAMDVSK-MRNVACYISHSTDPNVI----VQFVLHD-H-----NSLMFP-----RVMLFA-AENIPPMTELSLDYGVV-
NP_180887.1_At ---------FSMDVSR-MRNVACYISHSKEPNVM----VQFVLHD-H-----NHLMFP-----RVMLFA-LENISPLAELSLDYGLA-
AAK28966_At --------PLIISAKN-VGNVARFMNHSCSPNVF----WQPVSYE-N-----NSQLFV-----HVAFFA-ISHIPPMTELTYDYGVSR
NP_565056.1_At --------PLLISAKK-FGNVARFMNHSCSPNVF----WQPVIRE-G-----NGESVI-----HIAFFA-MRHIPPMAELTYDYGISP
AAK28973_At --------QVLISAKE-KGNVGRFMNHNCWPNVF----WQPIEYDDN-----NGHIYV-----RIGLFA-MKHIPPMTELTYDYGISC
NP_564036.1_At --------QVLISAKE-KGNVGRFMNHSCSPNVF----WQPIEYE-N-----RGDVYL-----LIGLFA-MKHIPPMTELTYDYGVSC
NP_178647.1_At ------VSSLVISAKK-SGNVARFMNHSCSPNVF----WQSIARE-Q-----NGLWCL-----YIGFFA-MKHIPPLTELRYDYGKSR
EAA50044_Mg -----PVTAQLYTGD--KGNIFRFINHRCRS-PA----VKIEGMK-------ISGKYR-----MMVVSV--RDVTAGEELTANWGRGF
XP_365330.1_Mg -----EYVGQIWTGE--EGNIFRFTNHKCAN-AE----AEVKGKK-------ISGAYR-----MVLVAT--QDILHGMEVCVDFGQKF
EAL89869.1_Af -------TYQIYQGE--MGNFTRFINHSCR--PN----SQFQRFY-------WRGQER-----IIVVSR---GVVAGREITVDYSDHY
EAA71382_Fg -------RYQIWQGK--QGNFTRFANHSCK--PN----AQFEKFV-------WLGTQH-----IVMNN------TADIPLLFRFNMIN
XP_357091.1_Mm ------------------GNWMSYVNCARFPKEQ----NLLAV-----------QHQ------GQIFYESCRDIQRNQELLVWYGNGY
NP_031574.1_Mm -----------------KSNWMRYVNPAHSAREQ----NLAAC-----------QNG------MNIYFYTIKPIPANQELLVWYCRDF
XP_128908.5_Mm ----------------KSSWMRYIRCARHCGEQ----NLTVV-----------QYR------SNIFYRACIDIPRGTELLVWYNDSY
EAL91753_Af -------HPLALISAKQHGNWTRYLNHS-CKPST----TFLKMTVG--------KR-------TIMAVQAVRDIDIFEEITVDYG---
XP_329163_Nc -AQLLHDQLVADIYPQQMGNWVRKVRHDGVSPSA----VFKVMKITG-------KWR--------VMLVAMRDIRDGDEITAKYGRGY
XP_449432_Cg --QD-EAVDSREYFGYWVFPKASYFNHSCDPN-------ITKT-----------RID------RKMVFTLNRDVACGEELNIDYSGVL
NC_001148.1_Sc --QE-EASDSREYFGYWVFPEASYFNHSCNPN-------ITKY-----------RKG------NSMLFTMNRDIKKDEQICIDYSGVL
NP_982445_Eg --EQ-EGNYSRDYLGYWVLPEASYFNHSCAPN-------LAKK-----------RVG------REMYFVLTSDVAAGEQLCIDYKGIL
EAK94622_Ca --SEVDPSEDKEFLGFGVYPSASFFNHSCSPN-------IVKT-----------RNN------SEMVFTTSKDIEIGEELCISYGNYT
NP_983709_Eg --LRYVGKFNLNQVNGQIYMLLSHLNHSCEPN-------IYYELEG--------HH---------INVYARKEIKSDEELTVSYVNPL
P38890_Sc --LTMIGTFNINQYNGQVYHWISFINHDCEPN-------AYIEQVE--------EHE------E-LRLHARKPIKKGEQIRITYVNPL
XP_446364_Cg --LEYLGRFHINQLSGQLYFLYSFLNHNCEPN-------VRYDINN--------KL-------E-LKVYARKFIKKDEELVTTYVNPL
EAK96662_Ca --LFMMGTYNINNLDSNVFLTQSHLNHNCASN-------TSVETEL--------NRT------AGLKVIAGRDIKSGEELTTTYVNPS
EAA75828_Fg ---------------DEVFLATSRINHSCKPN-------AQNRWN---------QGL------GKITVHAVEDIEQGQEITITYLGNP
EAA78580_Fg ---------------MDVFLQASRINHACDNN-------AQKDYN---------EGI------KRHTVHALRDIEEGEEITITYLGIL
NP_588413_Sp ----------------EMFLLGSRMNHDCSPN-------VKHTWN---------PRL------DQVTVHAVRDIEAGEEILTTYIDLH
EAA33925_Nc ---------------PPLFIEASRINHACNPN-------TQNSWN---------SRI------NRETIHAVRDIKKGEEITISYIGHF

```
EAA49663_Mg             ------------RFHSTFVEVSRMNHECSPN-------CAAYFD---------PMT------MSQRIYAIRDIMPGEELTVSYIDPV
EAA50913_Mg             ------------QFYVVFPEISRCNHDCRPN-------ADYYYD---------PET------LTQHVHAVRQINIGEEISLSYLDPL
EAA47700_Mg             --Q-----NPHRKAVPAVFRNASRLNHSCVPN-------CHFSWD---------RVT------RTCLVHAITDVSPGDELTVAYDSRG
EAA55373_Mg             -------------DYMGLYPEIARINHACSPN-------SFSRFH---------PSD------LTMDVGAMRDIMPGEEITISYIPLG
EAA55141_Mg             ------------GYMGLFPTISRINHACGPN-------SQVKFD---------PAT------LSQKAFASHDIEAGEEITISYAEFG
EAA48681_Mg             --------------HAGTFLGVSRLNHDCRPNNP-----TSLAFHV--------AKS------FVHTTHATRDIAAGEELTISYVDSY
EAA51372_Mg             -------------HYGNFPNVSKLNHDCRPN---------MVFHI--------DNN------LVHRTHAVRDIKPGEELTISYVDQM
EAA26997.1_nc           --QA--TLQTDHSKHYINYPSMFLFTHSCRPN---------IAFHI--------DGT------LALRTTVA--VAPGEELSIAYIDPL
XP_363030.1_Mg          --------YTEESQHLGVYAQAAAINHDCRPS-----------------------------INYRLNDITQIQPGEELSVSYVDLM
NP_572675_Dm            --------------ARTLYPSAFLLAHDCTPN-------TAHTDDPS-------S--------FEILLRTSRRVREREALTLSYAYTL
NP_609464_Dm            --------------LRCLFPYTGILAHNCVPN-------TSRSIYPS-----E-G--------YKIRLRAMVDLEEGQPLHHSYTYTL
NP_649084_Dm            --------------IRALYPGAAMISHDCVPN-------MRHRFDDD----------------MNIVFLAKRKIAKGEILSISYTQPL
NP_610944_Dm            --------------VAVFYT-ASFTENSCLPN-------LAKSFNKN----------------GHCILWAPREIKKNAHLSICYSDAM
NP_477333_Dm            -------------LLRGLFPLTAIMNHECTPN-------ASHYFENG-------R---------LAVVRAARDIPKGGEITTTYTKIL
NP_726802_Dm            -------------LVRALFPLAGLLNHQCTPN-------AAHHFENG-------E---------TIVVCATERIPAGAEITMSYAKLL
NP_572539_Dm            -------------NYRALYPLFGVVNHDCIPN-------AYYTFEEK------TN---------NMIVRAAVDIPEGFEVTTTYTKLF
NP_033892.1_Mm          --------------GVGIFPNLGLVNHDCWPN-------CTVIFN--------NG---------KIELRALGKISEGEELTVSYIDFL
NP_081072.1_Mm          --------------GSAIFPDVALMNHSCCPN-------VIVTYK---------GT--------LAEVRAVQEIHPGDEVFTSYIDLL
NP_081464.1_Mm          --------------GVGLYPSMSLLNHSCDPN-------CSIVFN---------GP--------HLLLRAVREIEAGEELTICYLDML
XP_360530_Mg            --------------GTFLDPVLAMANHSCVPN-------AVVLFW---------RR--------KAYLRAEMPIKQGSEISISYIDYT
EAA68634_Fg             --------------GIFLGTKLAMANHSCIPN-------AMVQFI---------GR--------RAILRAEKPIKIDDEIEISYTDYT
EAA36113_Nc             --------------GIYLHPSLSMVNHSCVPN-------AYIAFE---------KR--------KAFLKAERDLEPGDEILISYIGMA
NP_650955_Dm            --------------GSGLYLLQSKINHSCVPN-------ACSTFP--------YSN-------DIVVLKALAPIQQGEEICISYLDEC
NP_659167.1_Mm          --------------GSGLFVLQSCCNHSCVPN-------AETSFP--------ENN-------FVLHVTALEDIKPGEEICISYLDCC
EAA72568_Fg             -----SCRERPLFTTAGVWLLAARINHSCVGN-------CRRSFIGD-------IQI------VR----AARDIPAGTELTFPYCPTG
EAA69379_Fg             -----TGDEGFKYTTSGIWLLASRINHSCVGN-------CRPSFIGD-------MQI------VR----ATKDVPAKTEIFFCYRPPV
EAL91344.1_Af           -----DHDEQ-KFHSCGVWLLASYINHSCCSN-------ARRSFIGD-------MMI------VR----AAQDLAAGTEITFWYQSPL
XP_367445.1_Mg          ------VQEGGPESSTGMWLHAAYANHTCIPN-------ATRAFIGD-------MMI------VR----AARDIPAGAEIFMGYASLA
XP_366224.1_Mg          ------GLARQPPSARGLWPLAAAMNHACVAS-------TVRAFVGD-------VFI------TR----ALRDIEPGEELTQQYVPVR
EAA35550_Nc             ------TYN-SVCMAKGLWAHSSLMNHSCVPN-------TMRSFVGD-------MLI------CR----ATRDVQEGEELFQQYVPVK
AFF87042_At             -----VMGKNKEYYGVGLWTLASFINHSCIPN-------ARRLHVGD-------YVI------VH----ASRDIKTGEEISFAYFDVL
XP_387208.1_Fg          --------RDGRPDVAAVHPYWCLANHDCDPN-------VTWEWGGR-------MVL------EARKERVVGGIKKGEEILNHYCDVN
XP_957213.1_Nc          --------RDGRPEVAAVHPFWCLANHDCDPN-------VTWEWGGR-------MTL------FARETRVVGNIKAGEEILNHYCDVD
XP_365932.1_Mg          --------RDGRPDVAAVHPFWCLANHDCDPN-------VTWEWSGR-------MVL------RAKEERVEWGIKAGDEILNHYCDVN
XP_367482.1_Mg          -------------STSLFMGPARFANHDCGAN-------ARLKT----------ADQ------SIMEVQALRNIEVGEEITVTYGDNY
EAA33797_Nc             -------------STSLFMGPARFANHDCNAN-------ARLIT----------RGQ------AGIEIIACRNIEVGEEITVTYSESY
EAA78314_Fg             -------------CTSLFMGPARFANHDCDAN-------AKLMR----------TSH------AGIEIVATRPIDAGEEITVTYGDNY
EAL93147.1_Af           -------------TPSFFLGPARFANHDCNAN-------GKLVT----------RGS------EGMQVVATRDIYIGEEITVSYGDDY
NP_588078_Sp            -------------SMCLFLGPARFVNHDCNAN-------CRFNT----------SGK------R-IWLRCVRDIKPGEEITTFYSSNY
AAF45537.1_Dm           -------------CAQLWLGPAAYINHDCRAN-------CKFLA----------TGR------DTACVKVLRDIEVGEEITCFYGEDF
NP_659120_Mm            -------------CAQLWLGPAAFINHDCRPN-------CKFVS----------TGR------DTACVKALRDIEPGEEISCYYGDGF
NP_666289.1_Mm          -------------SAQLWLGPAAFINHDCKPN-------CKFVP----------SDG------NTACVKVLRDIEPGDEVTCFYGEGF
EAL85588.1_Af           --------------SKGMVPLADILNADADRNN------ARLF----------QED------DSLVMKAIKPIRVGEEIFNDYGELP
EAA77342_Fg             --------------MMGMVPMADILNADAEFN-------AHVN----------HEE------ESLTVTSLRPIKAGEEILNYYGPHP
EAA77142_Fg             --------LYIDTRQEGSLA--RYVRRSCKPN------AQLDTYLSDES---------E----YHFWIVSDRHISANEQITLP-----
EAA55907_Mg             --------LYIDTRQEGSRA--RYVRRSCKPN------AVLDTFLSGGS---------E----YHFWLVSDRSIAANEQITIA-----
EAA28230_Nc             --------IYIDTRKEGSQA--RYIRRSCKPN------AMLETYLSDGY---------E----LHFWLVTDRQVAAREEITLP-----
EAL93416.1_Af           --------IYIDSRHEGSIL--RYVRRSCRPN------VTIKTYITNEV---------E----YHFCFVAKEDISANSEITAM-----
NP_648681_Dm            --------VCVDTRTYGNDA--RFVRRSCRPN------AELQHYFEKG----------T----LHLYIVALTHIRAQTEITIR-----
NP_082661.1_Mm          --------MCVDARTFGNDA--RFIRRSCTPN------AEVRHMIADG----------M----IHLCIYAVSAITKDAEVTIA-----
XP_448906_Cg            --------LYIDARSKGNLT--RFLRRGCKPN------VELVTVCLPSAG---EEKD------IKFILRATRDIDEGEELLID-----
NP_012954_Sc            --------IYIDARLSGNST--RYLRRSCQPN------VELVTIKLQDTD---NRNDKR----IKFVLRALRDISEDEELYIK-----
NP_985624_Eg            --------IYIDARLCGNLT--RYLRRSCYPN------VELVTIKMPADD---RQAS-K----VKFVLKALRDLERGEELHIK-----
XP_446555_Cg            --------LYINTSGTKGAHNVEHLRCSCNPN------VELVTVRIMDNK---PR--------IKFVIRAIRDIAEGEELQIA-----
```

```
NP_012430_Sc            --------LYIDSRETGGLT--RYIRRSCEPN------VELVTVRPLDEK---PRGD-C----VKFVLRAIRDIRKGEEISVE-----
EAA30745_Nc             ------WDTDRIIDAGKMGNWTRFMNSSCDPN------VSESIMQIG-----------V----RVIAFFANKTLKSGDELCIYYGDDY
EAL93174.1_Af           ------MIALQPENAKVLVLGLAFVNHACIAN----SLYYTLRYPRD-----ENGEEDI----GRAVVRASRDIRPGEEITVAYFYAK
EAA71535_Fg             -------EDEDISNASTGLWVRSYINHSCIPN------AKKDLIG------------------DLILFRATRRIASGEEITHAYDEST
BAB11411_At             --------VEGEGAVGHAVYMPSFYNHDCDPN-------AHIIWLHN---------------ADARLNTLRDVEEGEELRICYIDAS
NP_588361_Sp            ------------EKQGLALGLGSMFNHDRHPN------VYWKKDNRN---------------NYISYYTLREIKTNEELCISYGDHL
EAA34532_Nc             ------SFAINPYSSDIISVYTSRLNHACRSVAN-----VLFDFDF--------QDP------ECITLTVSKPVKAGSELFISYGGSP
EAA68976_Fg             ------PSKGRCYDEAGLFPIASRFNHACEPMNS-----VKWHYN--------HADG-------NLKMTIQAKIEAGKELTITLPYFS
NP_193253.3_At          -------YDLVVVDAMHMANYASRICHSCRPN------CEAKVTAVDG--------------HYQIGIYSVRAIEYGEEITFDYNSVT
NP_594837_Sp            --------SQLVVDSRVAGSKARFARKGCQSN-----SVVSSVYMN------GSNSVP------RFILYSTTHIAPETEIIGD-----
EAK95983_Ca             --------LDIVLDSRFVGNESRFIRKACPSSAN----CRIETVYVP-----EQNKFR------FLVFTSKPITLKSENQDEELRLP-
XP_713530.1_Ca          ---------YQRQELGIALDPFALINHSCIPN------CCQITNDCN-----------------EFQIVSTLPINNGEELTVTYVSLG
NP_542983_Mm            --------EPYNHVSKYCASLGHKANHSFTPN------CVYDLFVHP-------RFG------PIKCIRTLRAVEAEEELTVAYGYDH
NP_565457.1_At          ----MEPCSVSNEKRSVRAYGTSFFNHDCLPN------ACRFDYVDS-----ASDGN------TDIIIRMIHDVPEGREVCLSYFPVN
CAE713530.1_Nc          -------ASISTSTPPTPTTNGSLINHCCTSPSAY---VARKRTERG-----PNCEWG-----LAHFVKPKHIIKEGEQLTWDYGKKE
```

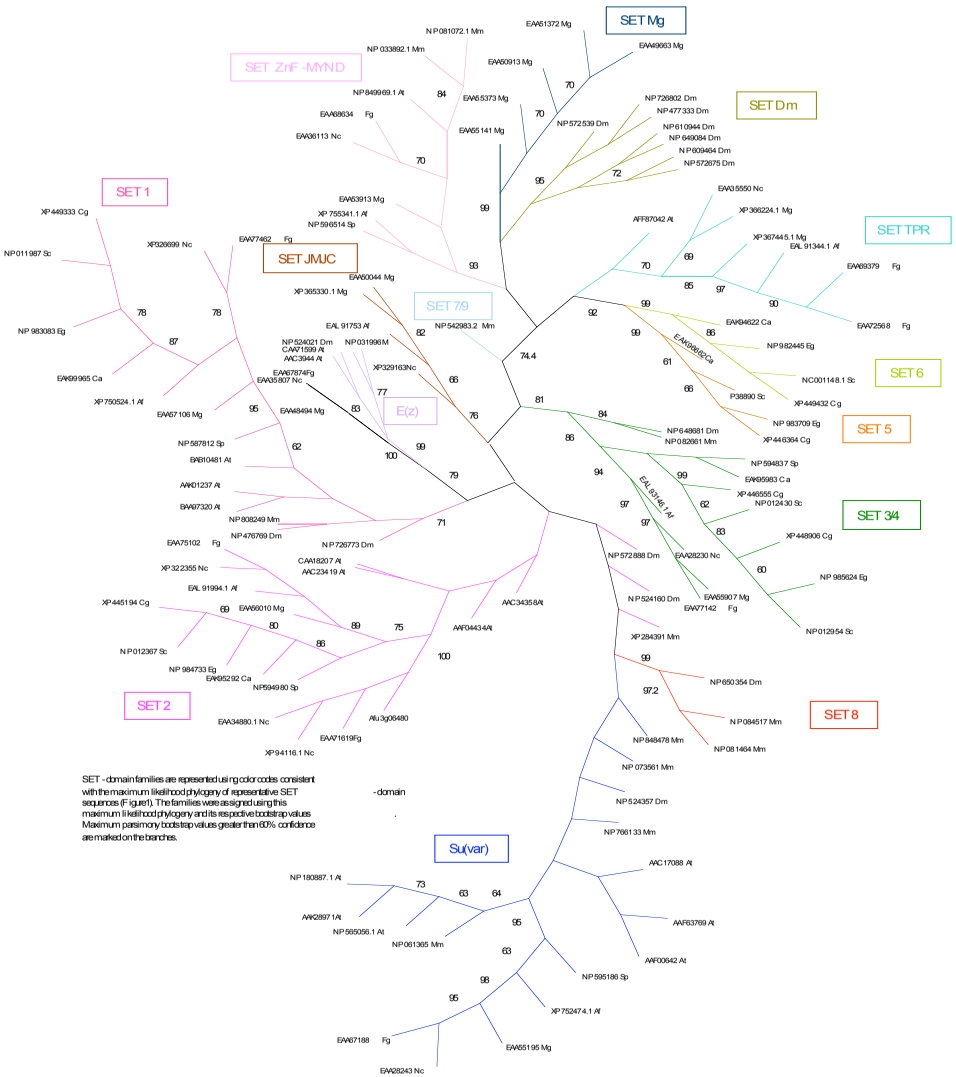
Appendix F

The CLUSTAL X (1.83) multiple alignment of 114 representative SET-domain sequences.


SET 3/4
SET JMJC
E(z)
Unknown group
SET 1
SET 2
Su(var)
SET 8
SET 5
SET 6
SET Dm
SET MYND
SET Mg
SET TRR
SET 7/9

Appendix G

Maximum parsimony phylogeny of 114 representative SET-domain sequences: SET-domain families are represented using color codes consistent with the maximum likelihood phylogeny of representative SET-domain sequences (Figure 3.1). The families were assigned using this maximum likelihood phylogeny and its respective bootstrap values. Maximum parsimony bootstrap values greater than 60% confidence are marked on the branches.


SET Mg
SET ZnF -MYND
SET Dm
SET 1
SET JMJC
SET TPR
SET 7/9
E(z)
SET 6
SET 5
SET 3/4
SET 2
SET 8
Su(var)
- domain
SET - domain families are represented using color codes consistent with the maximum likelihood phylogeny of representative SET sequences (Figure1). The families were assigned using this maximum likelihood phylogeny and its respective bootstrap values. Maximum parsimony bootstrap values greater than 60% confidence are marked on the branches.

Appendix H

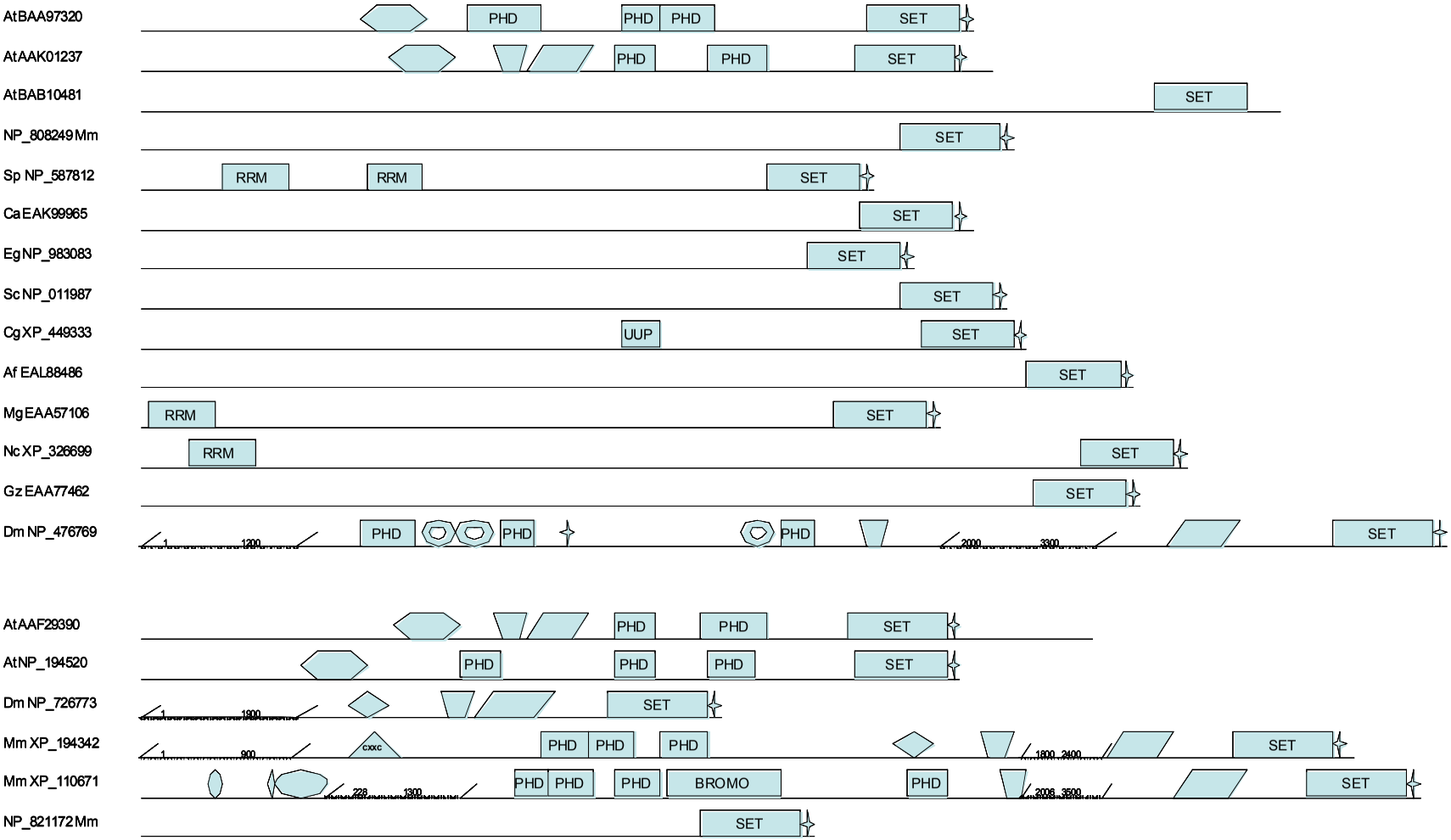
Gene architecture of SET-domain families represented in Figure 3.1.

Legend

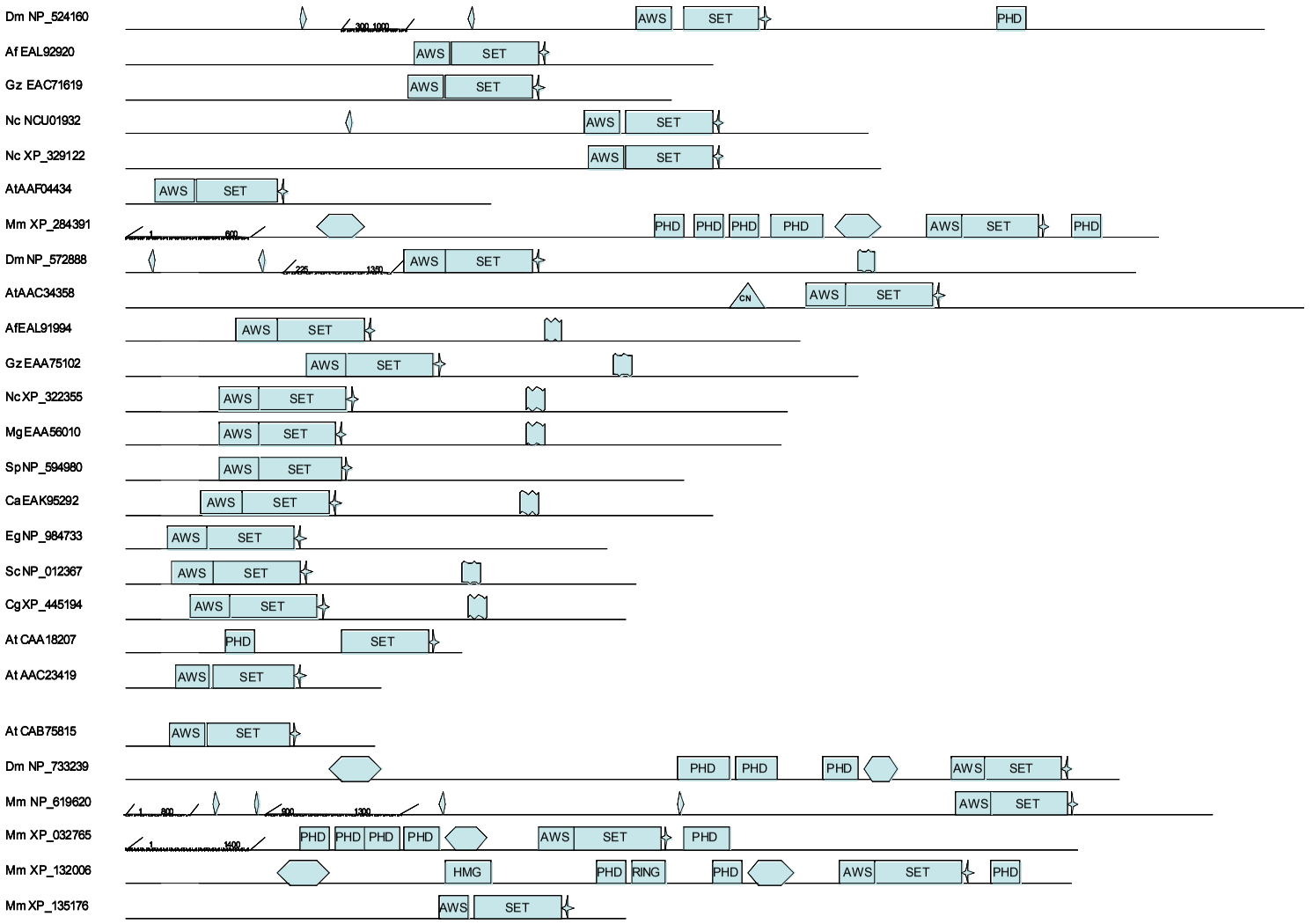
Symbol	Domain
SET	SET
SET	SET domain (conservation only in C terminal region)
SET	SET domain (conservation only in N terminal region)
	PHD ZnF
	ZnF MYND
	ZnF MYND (conservation only in C terminal region)
	ZnF MYND (conservation only in N terminal region)
COG	COG
	MORN
IPT	IPT
	ANK
TPR	TPR
TPR	TPR (conservation only in C terminal region)
DUF	DUF
	SANT
PHD	PHD
JMJC	JMJC (conservation only in C terminal region)
JMJC	JMJC
TPPN	TPPN
TPPM	TPPM
UV Endonudease	UV Endonuclease
AWS	AWS
RING	RING
HMG	HMG
PRS	PRS
PRS	PRS (conservation only in C terminal region)
PRS	PRS (conservation only in N terminal region)
HMTM	HMT MBD
SRA	SRA

Symbol	Domain
GTPe	GTP elongation factor
RRM	RRM
UUP	UUP
BROMO	BROMO
	PostSET
	WW
CN	ZnF CN
	PWWP
	AT Hook
	CHROMO
	ZnF
	ZnF (conservation only in C terminal region)
	RING
	FYRN
	FYRC
CXXC	ZnF CXXC
Start End	Region of gene between start and end position not shown

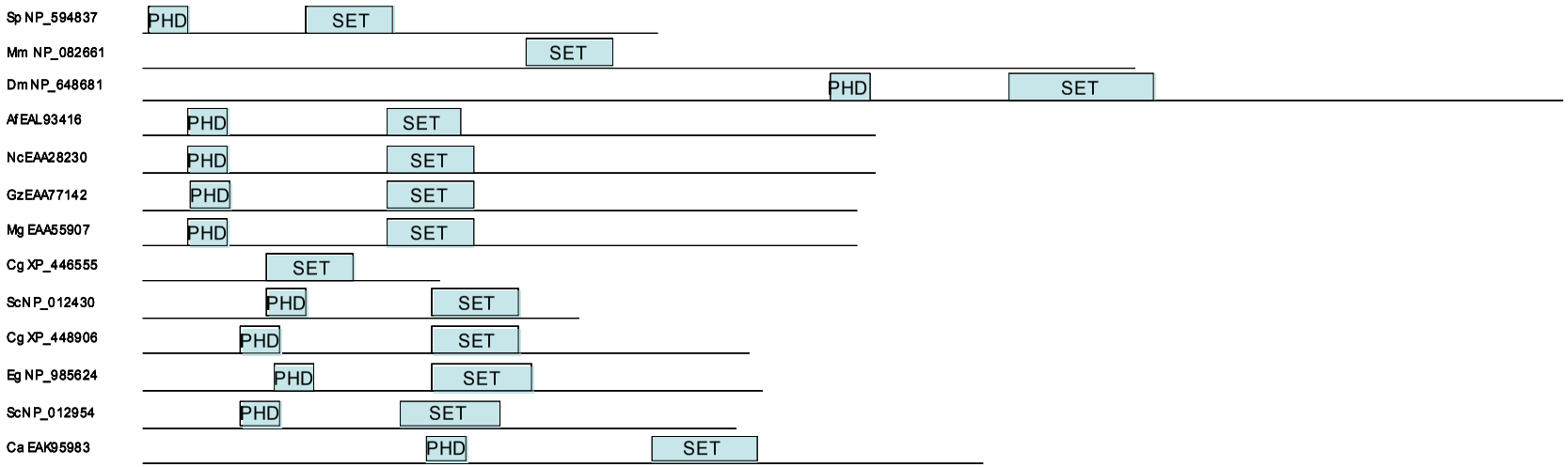
SET 1


AtBAA97320
PHD
PHD
PHD
SET
AtAAK01237
PHD
PHD
SET
AtBAB10481
SET
NP_808249Mm
SET
Sp NP_587812
RRM
RRM
SET
Ca EAK99965
SET
Eg NP_983083
SET
Sc NP_011987
SET
Cg XP_449333
UUP
SET
Af EAL88486
SET
Mg EAA57106
RRM
SET
Nc XP_326699
RRM
SET
Gz EAA77462
SET
Dm NP_476769
1
1200
PHD
PHD
PHD
2000
3300
SET
AtAAF29390
PHD
PHD
SET
AtNP_194520
PHD
PHD
PHD
SET
Dm NP_726773
1
1800
SET
Mm XP_194342
1
900
cxxc
PHD
PHD
PHD
1800
2400
SET
Mm XP_110671
228
1300
PHD
PHD
PHD
BROMO
PHD
2008
3500
SET
NP_821172Mm
SET

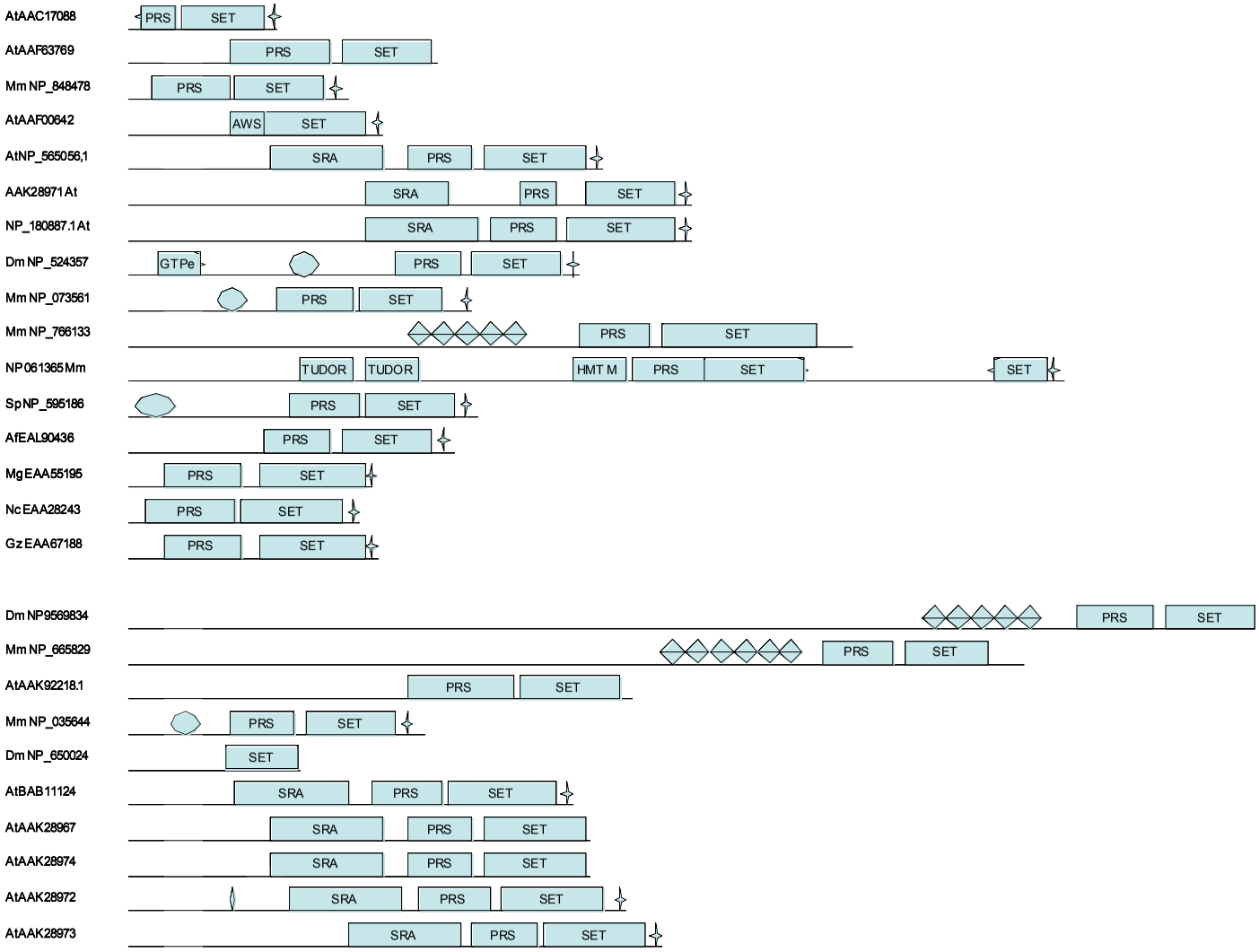
SET 2


Dm NP_524160
300
1000
AWS
SET
PHD
Af EAL92920
AWS
SET
Gz EAC71619
AWS
SET
Nc NCU01932
AWS
SET
Nc XP_329122
AWS
SET
AtAAF04434
AWS
SET
Mm XP_284391
1
600
PHD
PHD
PHD
PHD
AWS
SET
PHD
Dm NP_572888
225
1350
AWS
SET
AtAAC34358
CN
AWS
SET
AfEAL91994
AWS
SET
GzEAA75102
AWS
SET
NcXP_322355
AWS
SET
MgEAA56010
AWS
SET
SpNP_594980
AWS
SET
CaEAK95292
AWS
SET
EgNP_984733
AWS
SET
ScNP_012367
AWS
SET
CgXP_445194
AWS
SET
At CAA18207
PHD
SET
At AAC23419
AWS
SET
At CAB75815
AWS
SET
Dm NP_733239
PHD
PHD
PHD
AWS
SET
Mm NP_619620
1
800
900
1300
AWS
SET
Mm XP_032765
1
1400
PHD
PHD
PHD
PHD
AWS
SET
PHD
Mm XP_132006
HMG
PHD
RING
PHD
AWS
SET
PHD
Mm XP_135176
AWS
SET

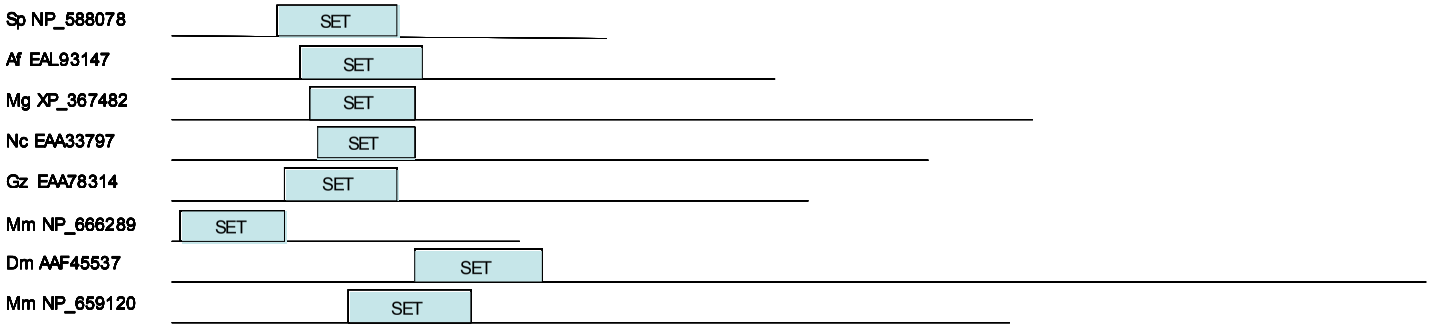
SET 3/4


Sp NP_594837
PHD
SET
Mm NP_082661
SET
Dm NP_648681
PHD
SET
Af EAL93416
PHD
SET
Nc EAA28230
PHD
SET
Gz EAA77142
PHD
SET
Mg EAA55907
PHD
SET
Cg XP_446555
SET
Sc NP_012430
PHD
SET
Cg XP_448906
PHD
SET
Eg NP_985624
PHD
SET
Sc NP_012954
PHD
SET
Ca EAK95983
PHD
SET

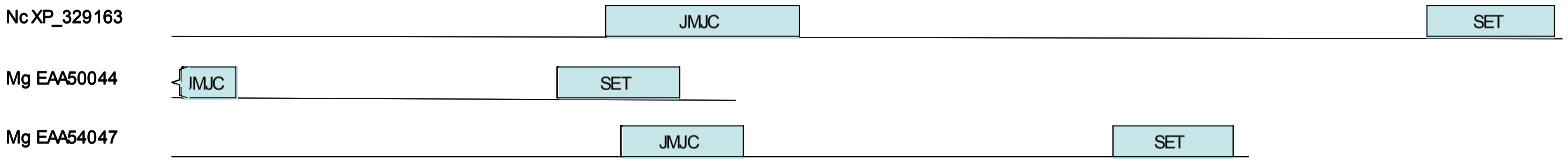
Su(var)3 -9


AtAAC17088
PRS
SET
AtAAF63769
PRS
SET
Mm NP_848478
PRS
SET
AtAAF00642
AWS
SET
AtNP_565056,1
SRA
PRS
SET
AAK28971 At
SRA
PRS
SET
NP_180887.1 At
SRA
PRS
SET
Dm NP_524357
GTPe
PRS
SET
Mm NP_073561
PRS
SET
Mm NP_766133
PRS
SET
NP 061365 Mm
TUDOR
TUDOR
HMT M
PRS
SET
SET
Sp NP_595186
PRS
SET
Af EAL90436
PRS
SET
Mg EAA55195
PRS
SET
Nc EAA28243
PRS
SET
Gz EAA67188
PRS
SET
Dm NP9569834
PRS
SET
Mm NP_665829
PRS
SET
AtAAK92218.1
PRS
SET
Mm NP_035644
PRS
SET
Dm NP_650024
SET
AtBAB11124
SRA
PRS
SET
AtAAK28967
SRA
PRS
SET
AtAAK28974
SRA
PRS
SET
AtAAK28972
SRA
PRS
SET
AtAAK28973
SRA
PRS
SET

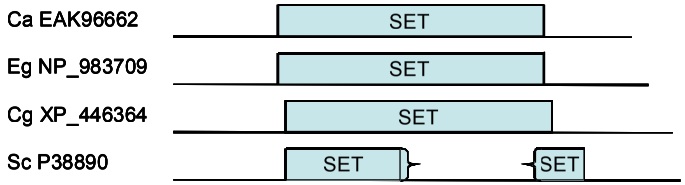
Su(var) 4-20



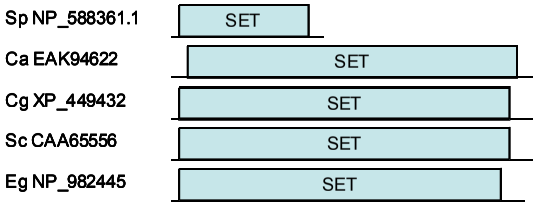
SET JmjC



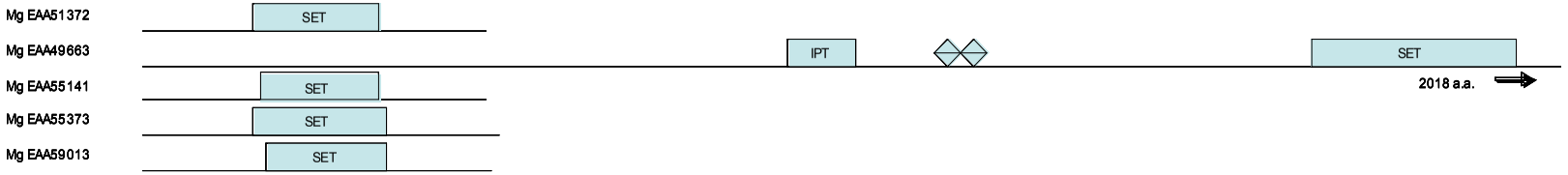
SET 5



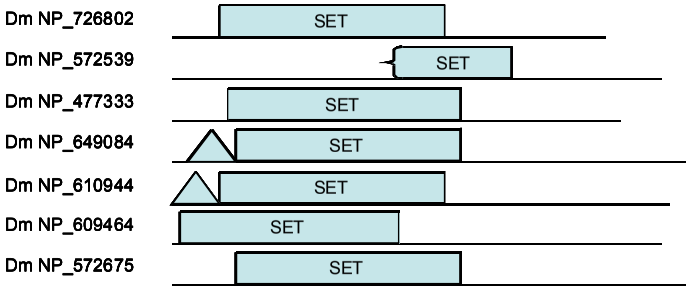
SET 6



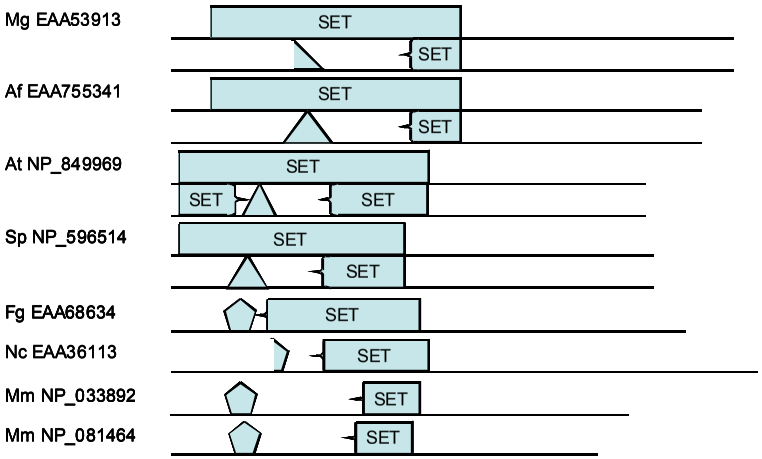
SET Mg



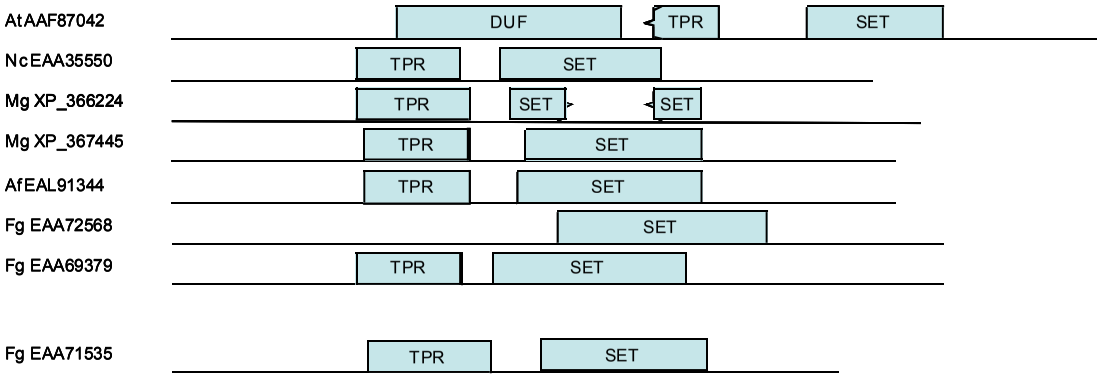
SET Dm



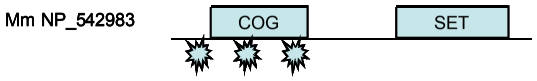
SET MYND



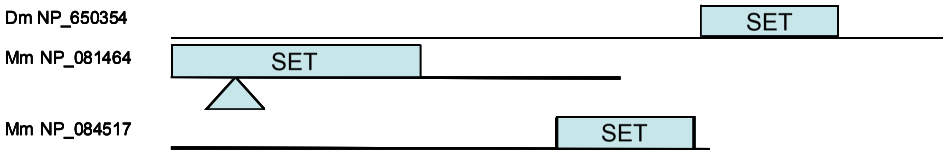
SET TPR



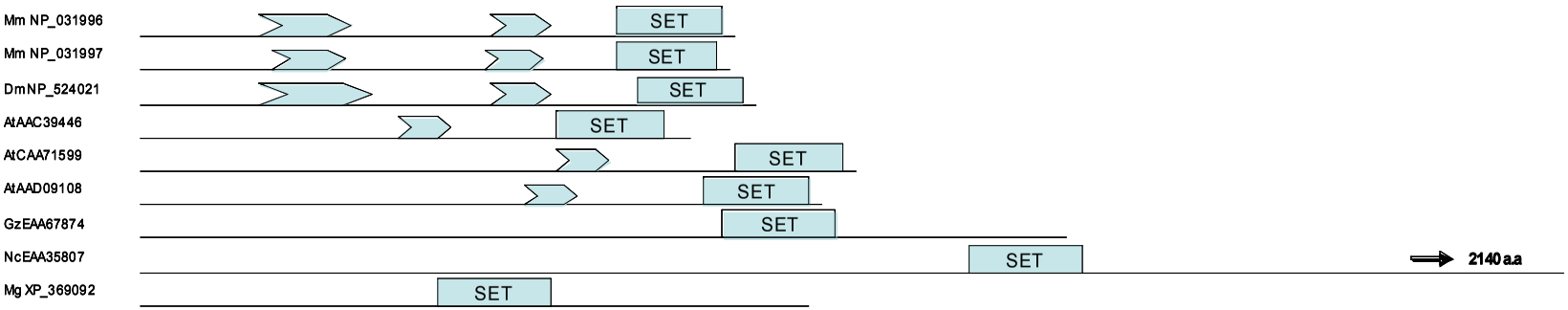
SET 7/9



SET 8



E(z)



Appendix I

Unknown SET-domain sequences.

N. crassa

CAE76340 - similarity to SET MYND
XP_957213.1 - similarity to N-terminal region of SET MYND
EAA 26997.1 - similarity to C-terminal region of SET MYND
EAA33925 - similarity to C-terminal region of SET TPR
EAA34532 - unknown
EAA30745 - similar to human EHMT2 (H3 K9 methyltransferase)

A. fumigatus

EAL85588.1 - similarity to C-terminal region of SET MYND
EAL89869.1 - Unknown

F. graminearum

EAA77342 - Unknown
EAA71535 - similar to SET TPR
XP_387208 –similarity to N-terminal region of SET MYND
EAA78580 - similarity to C terminal region of SET TPR
EAA75828 - similarity to C-terminal region of SET MYND
EAA68976 - similarity to C-terminal region of SET MYND
EAA71382 - Unknown

M. griesa

XP_365932.1 - similarity to N-terminal region of SET MYND
EAA48681 - similar to SET TPR
XP_363030 - similarity to C-terminal region of SET MYND
EAA47700 - unknown
XP_360530 - SET MYND

C. albicans

XP_713530 – similarity to SET MYND

S. pombe

NP_588361 - Unknown
NP_588413 - similarity to C terminal region of SET MYND

M. musculus

NP_659167.1 - SET MYND
XP_128908.5 – SET ZnF (SET and Zinc finger domains)
NP_031574.1 - SET MYND
XP_357091.1 – SET ZnF

D. melanogaster

NP_650995 - SET MYND
NP_610202 - SET MYND

A. thaliana

NP_197821.1 - SET1 related
BAB09537 - SET1 related
NP_565457 - ASH1 related
NP_193253 - non histone methyltransferase
BAB11411 - unknown

D. hansenii

XP_456345 - Unknown
XP_462118 - non histone methyltransferase

Y. lipolytica

XP_503477 – related to human G9a protein (H3 K9 methyltransferase)
XP_506045 – non histone methyltransferase

Appendix J

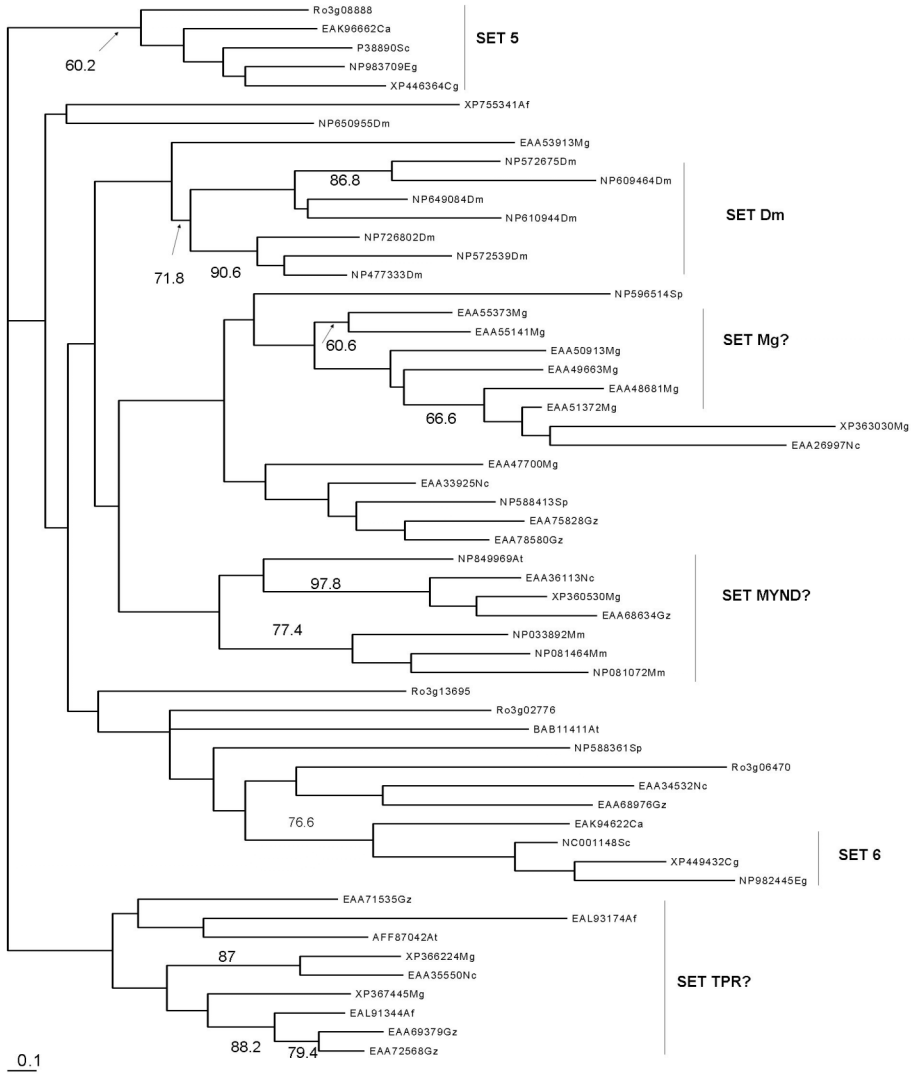
CLUSTAL X (1.83) multiple alignment of SET-domain sequences from the SET 3/4 family to reconstruct a family specific maximum likelihood phylogeny. SET-domain hits from closely related species of *Saccharomyces cerevisiae*(Sc), *S. bayanus*(Sb), *Schwanniomyces castelli*(Scas), *Candida glabrata*(Cg), and in addition, *Candida albicans*(Ca), *Kluyveromyces lactis*(Kl), *Kluyveromyces waltii*(Kw), *Yarowia lioplytica*(Yl) and *Debaryomyces hansenii*(Dh) were used to perform the multiple alignment followed by the reconstruction of the maximum likelihood phylogeny(Figure 3.2)

```
NP_012430Sc                     EVRKSSNERDFGVFAADSCVKGELIQEYLGKIDFQKNYQTDPNNDYRLMGTTKPKVLFHPHWPLYIDSRETGGLTRYIRRSCEPNVEL
AACA01000375.1(6257-6931)Sb     DIKKSSNGKDLGVFISNSCVKGQLIQN-IAKLTFKKITRQHASNNYRLMGTTQPKVLFHPHWPLFIDSREVGGLTRYIRRSCEPNV-L
XP_446555Cg                     KPLASKDIQDFGVFTSIPCSKKDYIQEYSGMISFSKQYVNNAENKYDILGTPTRNILFHPHWPLYINTSGTKGAHEHLRCSCNPNVEL
AACF01000014.1(32239-32583)Scas ------------------CLKGDLIDEFLGVVDFQKNYIMDPINQYRLWGTTKRDVIFHSQWPLLIDAREKSGRFRSLRRSCNPNVEL
NP_012954Sc                     YSRTYPGFTKLGVYLKKDCIKGDFIQEILGELDFYKNYLTDPRNHYRIWGTAKRRVIFHSHWPIYIDARLSGNSTRYLRRSCQPNVEL
AACG020000197.1(8320-7895)Sb    YSRTYPGFTKLGVYLKRDCMKGDFVQEFLGELDFRKNYLTDSRNHYRIWGTPKRRVIFHPHWPIYIDARSSGNLTRYLRRSCQPNVEL
XP_448906Cg                     YARTFPAHTKLGVFTKEYCNNGDLIEEFTGQVDFLKTYLDDTKNHYRIWGTAKNRVIFHPHWPLYIDARSKGNLTRFLRRGCKPNVEL
XP_455210Kl                     HSRLFPGYPKLAVYLKQGCDEGDYIDEFLGEIDFQRKYLEDPRNNYRVWGTAKPKVIIHPHWPIYIDARLSGNLTRYLRRSCNPNVEL
NP_985624Eg                     HSRVFHGFPKLGVYSQQPCAKDTLIAEFLGEVDFQRKYLEDPRNNYRLLGIPNPKVLFHPHWPIYIDARLCGNLTRYLRRSCYPNVEL
AADM01000081.1(81822-82199)Kw   HSRSFTGLSKMGLFVKQACPVGTYIEEVLGEVDFSKKYYADPRNNYRIFGTTKPKVFLHPHWPIYIDCRLSGNITRFMRRSCYPNVEL
AACF01000109.1(21368-21757)Scas YSRTFPGFPKLGTFLPEGCNESALIQEFLGELNFKEDYLDDPRNMYRIWGTVKSKVVFHPNWPLCIDARSCGNLARYIRRCCNPNVGL
XP_504932Yl                     PKSKFCGFSKHGLVANQPIQKDRFIIEYVGHVSHKDQYKADPINQYRIHPVPKSSVLFHPTLPLVIDGRLVGNDARFMRRSCNPNCRV
EAK95983Ca                      -ALFISSNGGGSGGGSLTIPENTPIIEYLGEIDLFKNYCRDSINQYRMWGSPKPKVLKTTILDIVLDSRFVGNESRFIRKACSANCRI
NC_006044Dh                     --------KDLKYEKEIIVPQETPIIEYLGEIDLFDNYVNDQVNQYSAFGTPKPKVLKVDLLEIVSDSRFVGNESRFIRKACTSNCKI

NP_012430Sc                     VTVRPLDEKPRGDND-CRVKFVLRAIRDIRKGEEISVEWQWDLRNP
AACA01000375.1(6257-6931)Sb     ITVRPLNEKQRGESD-CRVTFVLMATRDIKVGEEISLKWQWDLRNP
XP_446555Cg                     VTVRIMDNKPR-------IKFVIRAIRDIAEGEELQIAWQWDINHP
AACF01000014.1(32239-32583)Scas VTVKLGQDQDKHWGD-STVKFMIRAKRTIQAGEELHINWQWDLRHP
NP_012954Sc                     VTIKLQDTDNRNDKSSSRIKFVLRALRDISEDEELYIKWQWDSKHP
AACG020000197.1(8320-7895)Sb    VTIRLQDLDHENVNSVSKIKFVLRALRDISEDEELYIKWQWDLKQP
XP_448906Cg                     VTVCLPSAGEEK-----DIKFILRATRDIDEGEELLIDWKWDLRHP
XP_455210Kl                     VTIKLTNSNNQD-----EIKFVLRALRFIDKGEELHIKWDWDLRHP
NP_985624Eg                     VTIKMPADDRQASDVNKSVKFVLKALRDLERGEELHIKWDWDLRHP
AADM01000081.1(81822-82199)Kw   VSMRMNDE--------SNVKFVLRALRDIDDDEELQIGWQWDLSHP
AACF01000109.1(21368-21757)Scas STVKIKETN--------EIKFVLKALRDINPGEELHLSWHWDKKHP
XP_504932Yl                     ATVVVNNTD---------IIFVVFATEPIKPGTELTLAWEWDTHHP
EAK95983Ca                      ETVYVPEQNKFR---FLVFTSKPITLKSENQDEELRLPWEWDVDHP
NC_006044Dh                     KPIYIPETNTFR---LLVVTSKPIILSSNTNEEELRLDWEWDRLHP
```

Appendix K

ML phylgeny from SET 5/6 'Super' group
All representatives from the group containing SET 5, SET 6, SET Mg, SET Dm, SET MYND, SET TPR from the draft phylogeny (see branch with arrow in Appendix D) were chosen. XP_365932, XP_957213.1 and XP_387208 were removed because of high sequence divergence. SET-domain families reliably classified in Figure 3.1 but lacking support in this phylogeny are indicated using a '?' symbol.500 replicates for bootstrap. Confidence level greater than 60% shown.


Ro3g08888
EAK96662Ca
P38890Sc
NP983709Eg
XP446364Cg
SET 5
60.2
XP755341Af
NP650955Dm
EAA53913Mg
NP572675Dm
NP609464Dm
86.8
NP649084Dm
NP610944Dm
SET Dm
NP726802Dm
NP572539Dm
71.8
90.6
NP477333Dm
NP596514Sp
EAA55373Mg
EAA55141Mg
60.6
EAA50913Mg
EAA49663Mg
EAA48681Mg
SET Mg?
EAA51372Mg
66.6
XP363030Mg
EAA26997Nc
EAA47700Mg
EAA33925Nc
NP588413Sp
EAA75828Gz
EAA78580Gz
NP849969At
EAA36113Nc
97.8
XP360530Mg
EAA68634Gz
SET MYND?
77.4
NP033892Mm
NP081464Mm
NP081072Mm
Ro3g13695
Ro3g02776
BAB11411At
NP588361Sp
Ro3g06470
EAA34532Nc
EAA68976Gz
76.6
EAK94622Ca
NC001148Sc
XP449432Cg
SET 6
NP982445Eg
EAA71535Gz
EAL93174Af
AFF87042At
87
XP366224Mg
EAA35550Nc
SET TPR?
XP367445Mg
EAL91344Af
EAA69379Gz
88.2
79.4
EAA72568Gz
0.1