Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes

Ritu Pandey[1,2], Andreas Müller[1], Carolyn A. Napoli[1], David A. Selinger[1,3], Craig S. Pikaard[4], Eric J. Richards[4], Judith Bender[5], David W. Mount[2] and Richard A. Jorgensen[1,*]

[1]Department of Plant Sciences, University of Arizona, Tucson, AZ 85721-0036, USA, [2]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721-0106, USA, [3]Pioneer Hi-bred International, DuPont Agriculture and Nutrition, Johnston, IA 50131, USA, [4]Department of Biology, Washington University, St Louis, MO 63130, USA and [5]Department of Biochemistry and Molecular Biology, Johns Hopkins University, Baltimore, MD 21205, USA

ABSTRACT

Sequence similarity and profile searching tools were used to analyze the genome sequences of *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans* and *Drosophila melanogaster* for genes encoding three families of histone deacetylase (HDAC) proteins and three families of histone acetyltransferase (HAT) proteins. Plants, animals and fungi were found to have a single member of each of three subfamilies of the GNAT family of HATs, suggesting conservation of these functions. However, major differences were found with respect to sizes of gene families and multi-domain protein structures within other families of HATs and HDACs, indicating substantial evolutionary diversification. Phylogenetic analysis identified a new class of HDACs within the RPD3/HDA1 family that is represented only in plants and animals. A similar analysis of the plant-specific HD2 family of HDACs suggests a duplication event early in dicot evolution, followed by further diversification in the lineage leading to *Arabidopsis*. Of three major classes of SIR2-type HDACs that are found in animals, fungi have representatives only in one class, whereas plants have representatives only in the other two. Plants possess five CREB-binding protein (CBP)-type HATs compared with one to two in animals and none in fungi. Domain and phylogenetic analyses of the CBP family proteins showed that this family has evolved three distinct types of CBPs in plants. The domain architecture of CBP and TAF$_{II}$250 families of HATs show significant differences between plants and animals, most notably with respect to bromo-domain occurrence and their number. Bromodomain-containing proteins in *Arabidopsis* differ strikingly from animal bromodomain proteins with respect to the numbers of bromodomains and the other types of domains that are present. The substantial diversification of HATs and HDACs that has occurred since the divergence of plants, animals and fungi suggests a surprising degree of evolutionary plasticity and functional diversification in these core chromatin components.

INTRODUCTION

Gene expression in eukaryotes involves a complex interplay among transcription factors and chromatin proteins that pack chromosomal DNA into the confined space of the nucleus while poising genes for activation or repression (1). The basic unit of chromatin is the nucleosome core particle, a structure in which ~146 bp of DNA is wrapped around a protein octamer made up of two subunits each of the core histones H2A, H2B, H3 and H4 (2). Core histones can exist in multiple alternative states of acetylation, methylation, phosphorylation, ubiquitination or ADP-ribosylation (3). The regulatory significance of these modifications for processes including gene repression, gene activation and replication is increasingly clear (4–6).

Lysines at the N-terminal ends of the core histones are the predominant sites of acetylation and methylation and a

regulatory role for these modifications was proposed as early as 1964 (7). However, decades passed before it was demonstrated that active genes are preferentially associated with highly acetylated histones whereas inactive genes are associated with hypoacetylated histones (8). The N-termini of histones H3 and H4 were subsequently shown to be essential for repression of the silent mating type loci in *Saccharomyces cerevisiae* (9,10). Enhancer-dependent activation of other *S.cerevisiae* genes also required these N-terminal sequences (11–13). Collectively, these studies suggested that histones are integral to both gene activation and gene repression mechanisms. A breakthrough was the finding that a *Tetrahymena thermophila* protein with histone acetyltransferase (HAT) activity shared substantial similarity with *S.cerevisiae* Gcn5p (14), the catalytic subunit of several multi-protein complexes required to activate a diverse set of genes. A complementary breakthrough was the finding that a purified mammalian histone deacetylase (HDAC) was similar to Rpd3p (15), a protein which helps repress numerous genes in *S.cerevisiae* (16), also as part of a larger protein complex (17–19). Histone acetylation and deacetylation are thought to exert their regulatory effects on gene expression by altering the accessibility of nucleosomal DNA to DNA-binding transcriptional activators, other chromatin-modifying enzymes or multi-subunit chromatin remodeling complexes capable of displacing nucleosomes (20,21).

Sequence characterization reveals at least four distinct families of HATs and three families of HDACs (3,22,23). HATs include: (i) the GNAT (GCN5-related N-terminal acetyltransferases)-MYST family (24,25) whose members have sequence motifs shared with enzymes that acetylate non-histone proteins and small molecules; (ii) the p300/CREB-binding protein (CBP) co-activator family in animals implicated in regulating genes required for cell cycle control, differentiation and apoptosis (26,27); and (iii) the family related to mammalian TAF$_{II}$250, the largest of the TATA binding protein-associated factors (TAFs) within the transcription factor complex TFIID (28). These three families are widespread in eukaryotic genomes, and homologous proteins are also involved in non-HAT reactions in prokaryotes and Archaea. Mammals have a fourth HAT family that includes nuclear receptor coactivators such as steroid receptor coactivator (SRC-1) and ACTR, a thyroid hormone and retinoic acid coactivator that is not represented in plants, fungi or lower animals (22,29,30).

Major groups of HDACs include the RPD3/HDA1 superfamily, the Silent Information Regulator 2 (SIR2) family (31) and the HD2 family. RPD3/HDA1-like HDACs are found in all eukaryotic genomes. Interestingly, homologous proteins that have acetate utilization and acetylpolyamine amino-hydrolase activities are also present in bacteria and Archaea, organisms that lack histones (17,32). The SIR2 family of HDACs is distinctive in that it has no structural similarity to other HDACs and requires NAD as a cofactor (33). In *S.cerevisiae*, *SIR2* is known to play roles in repression of silent mating type loci (34), repression of rRNA gene recombination (35), and repression of protein-coding genes inserted near telomeres (34) or within rRNA gene arrays (36). Mutations in *SIR2* also affect aging and longevity in *S.cerevisiae* (37,38). SIR2-related proteins form a large family with members present in all kingdoms of life, including bacteria (39). The

third family, the HD2-type HDACs, were first identified in maize and appear to be present only in plants (40,41). HD2-type HDACs are homologous to a class of *cis–trans* prolyl isomerases present in other eukaryotes (42).

Limited information is available concerning the roles of most proteins in the four HAT homology groups and the three HDAC homology groups in control of gene expression in multicellular eukaryotes, especially in plants (43,44). Here, we present phylogenetic and domain analyses of HAT- and HDAC-related proteins identified in searches of the essentially complete *Arabidopsis thaliana* genome sequence. To test and correct open-reading frames (ORFs) predicted by exon-modeling algorithms, cDNA sequences were determined for most of these proteins. Alternative splicing was demonstrated for 3 of 16 genes encoding HDACs. Together, these data provide a foundation for the functional analysis of these important chromatin-modifying activities in *Arabidopsis*, as well as in other plants and model organisms.

MATERIALS AND METHODS

Database similarity searches of the *Arabidopsis* genome and other plant sequences

Known HDAC and HAT protein sequences available from a variety of eukaryotic organisms (Table S1) were used as queries to search the complete *Arabidopsis* genome sequence (45) using the TBLASTN and TFASTX programs (46,47). To assure that all homologous genes in these families had been identified, three additional searches were performed. First, all *Arabidopsis* protein sequences in GenBank including those predicted by genome annotation were searched with the query sequences using BLASTP, FASTA and SSEARCH. Secondly, these protein sequences were searched for protein family (Pfam) domains known to be present in previously characterized HDAC and HAT proteins using the program HMMER (http://hmmer.wustl.edu/). Thirdly, predicted *Arabidopsis* HDAC and HAT proteins were used as queries to search for additional paralogous genes in the *Arabidopsis* genome sequence using TBLASTN and TFASTX. Sequences having an *E*-value of 0.01 or less were investigated further. However, this third approach did not find any proteins in addition to those that had already been identified by the initial TBLASTN or TFASTX searches.

Gene nomenclature

The genes identified in this study are listed in Table 2. To designate newly identified genes, we used three-letter symbols that specify the homology group to which a gene belongs as follows: *HAG* for HATs of the GNAT/MYST superfamily, *HAC* for HATs of the CBP family, *HAF* for HATs of the TAF$_{II}$250 family, *HDA* for HDACs of the RPD3/HDA1 superfamily, *SRT* for HDACs of the SIR2 family (sirtuins), and *HDT* for HDACs of the HD2 family ('HD-tuins'). To designate individual genes within a homology group, the three-letter symbol is followed by a numeral that does not imply orthology because in many cases it was not possible to determine orthology. To ensure that orthology is not inferred from numerals, a different series of numerals was assigned to different species: *A.thaliana* genes are indicated by the numerals 1–99, *Zea mays* by 101–199, *S.cerevisiae* by

201–299, *Caenorhabditis elegans* by 301–399, *Drosophila melanogaster* by 401–499 and *Schizosaccharomyces pombe* by 601–699 (for other organisms, see Table 2). Names of genes previously assigned in the literature or in GenBank were retained, except in *Arabidopsis*, for which we propose that the designations defined here should be used. To avoid possible confusion with HDA1 of *S.cerevisiae*, the *Arabidopsis* HDA series begins with HDA2.

Gene annotation

High quality plant protein sequences for phylogenetic analysis were obtained in several steps. First, the gene prediction programs GeneMark (48), GenScan (49) and NetPlantGene (50) were used to produce gene models for these sequences. From these separate models, a single consensus model was derived. To verify the gene models, RNA gel blots were used to determine the length of the mRNA from each expressed gene. The positions of exons in the consensus model were then tested by analysis of available *Arabidopsis* EST sequences using the gene prediction tool, GeneSeqer (51). For genes that were not completely represented by EST sequences, EST clones were obtained from the ABRC and Kazusa stock centers and sequenced. Remaining gaps between known cDNA sequences were filled by sequencing RT–PCR amplification products obtained using total RNA as the template and using primers that annealed to predicted coding sequences. Although the actual start codon for each protein has not been identified with certainty, none of the predicted proteins lack known conserved N-terminal or C-terminal domains, suggesting that the modified gene models are reasonably accurate.

cDNA sequence was not determined for *HDA10* and *HDA17* because these genes are truncated in their HDAC domains, *HAC2* because it could not be amplified by RT–PCR, and *HAC12* and *HAF2* because they are highly similar to *HAC1* and *HAF1*, respectively. *HAC12* and *HAF2* were annotated according to the splicing models of *HAC1* and *HAF1*. In the case of *HAC4*, the only transcript we detected carries a premature nonsense codon that would eliminate conserved regions of the protein, although the transcript extends beyond this and contains these conserved regions. Thus, for purposes of the phylogenetic and domain analyses presented here, we have used an algorithm-derived splicing model that predicts the conserved CBP-type HAT domain and cDNA sequence-derived splice junctions in the remainder of *HAC4*. Alternative splicing products were observed for three genes: *HDA2*, *HDA15* and *SRT2*. For purposes of the phylogenetic analyses presented here, we used the predicted protein sequence that possessed intact conserved domains (HDA2alt1, HDA15alt1 and SRT2alt1).

The cDNA sequence data for the HAT and HDAC genes have been submitted to the GenBank data library under the following accession numbers: *HAC4* (AF512559, AF512560, AH011643), *HAC5* (AF512557, AF512558, AH011642), *HAG2* (AF512724), *HAF1* (AF510669), *HDA2alt1* (AF510671), *HDA2alt2* (AF510165), *HDA7* (AF510166), *HDA9* (AF512725), *HDA15alt2* (AF510169), *HDA15alt3* (AF510170), *HDA18* (AF510670), *SRT2alt2* (AF510171), *SRT2alt3* (AF510172), *SRT2alt4* (AF510173), *SRT2alt5* (AF510174), *SRT2alt6* (AF510175). For rest of the genes, cDNA sequences submitted by other groups were found in GenBank and were identical to the sequence data generated by

Plant Chromatin Consortium. Their accession numbers are as follows: *HAC1* (AF323954), *HAG1* (AF338768), *HAG3* (AY056323), *HAG4* (AY099684), *HAG5* (NM_121011), *HDA5* (AY090936), *HDA6* (AF195548), *HDA8* (AY097371), *HDA14* (AY052234), *HDA15alt1* (NM_112737), *HDA19* (AY093153), *HDT1* (AF195545), *HDT2* (AF044914), *HDT3* (AF372889), *HDT4* (AF255713), *SRT1* (AF283757), *SRT2alt1* (AY045873).

Similarity searches of non-plant genomes for HDAC and HAT genes

The genomes of a diverse group of organisms were searched with the query sequences (Table S1), as well as with any *Arabidopsis* HDAC and HAT sequences showing similarity to these query sequences. First, BLASTP searches of the individual proteomes of baker's yeast (*S.cerevisiae*), nematodes (*C.elegans*), fruit flies (*D.melanogaster*), and several species of bacteria and Archaea were conducted. Secondly, genomic sequences of humans (*Homo sapiens*), fission yeast (*S.pombe*; http://www.sanger.ac.uk/Projects/S_pombe/), and leishmania (*Leishmania major*; http://www.sanger.ac.uk/Projects/L_major/) were individually searched for homologous sequences using TBLASTN. Thirdly, the public GenBank nr (non-repeating) databases were searched to identify homologs in additional species using BLAST and PSI-BLAST. Fourthly, a large number of plant EST collections (including *Z.mays*, *Oryza sativa*, *Lycopersicon esculentum*, *Medicago truncatula*, *Glycine max*, *Triticum aestivum*, *Sorghum bicolor*, *Gossypium arboreum*, *Solanum tuberosum*, *Hordeum vulgare*, *Lotus japonicus* and *Mesembryanthemum crystallinum*) were searched using TBLASTN. Plant ESTs were assembled into contigs using the FAKtory DNA sequence assembly system (http://bcf.arl.arizona.edu/faktory/) and the contigs were translated into amino acid sequences for further analysis.

Analysis of protein families

Phylogenetic analysis. Protein sequences and domains were aligned using Clustal W (52), edited with Genedoc (http://www.psc.edu/biomed/genedoc/), and an unrooted phylogenetic tree was constructed by the distance method using the neighbor-joining algorithm implemented in the program Neighbor in the PHYLIP (3.5) package (53). The Dayhoff PAM model of protein evolution was used to compute the distances between the sequences (54) using the PROTDIST program. This analysis allowed the identification of the most similar protein sequences in the same or different organisms based upon protein sequence similarity in the multiple sequence alignment. These alignments are available in Figures S1–S3. Identification of a paralogous family of sequences was revealed by the presence of a cluster of similar sequences from one organism or group of organisms that appeared to have arisen by gene duplication. Assignments of likely orthology were based upon the observation of a high level of sequence similarity among unique sets of sequences present in diverse organisms. In order to assess how well the multiple sequence alignment supported the branch patterns in the predicted phylogenetic tree of the sequences, a bootstrap analysis was performed using PHYLIP. This method resampled columns in the multiple sequence alignment to generate 500–1000 new alignments, each of which was used to

produce a new tree. The number of alignments that support each branch pattern in the tree was then assessed and is reported in the appropriate figure. When a clear majority of bootstrap trees (>70%) were in agreement, support was considered to be good. In many cases, bootstrap support was excellent, in the 95–100% range.

For phylogenetic analysis of the HD2 family (Fig. 5), mRNA and EST sequences encoding the HD2-type HDAC domains were aligned by CLUSTALW. These alignments are available in Figure S4. Following some minor editing to match the codons to the protein sequence alignment, an unrooted tree was produced using the maximum likelihood method as implemented in the DNAML program using the default transition/transversion ratio of 2:1 in the PHYLIP suite.

Motif analysis of the RPD3/HDA1-related HDACs. To identify common motifs in the HDAC domain, a multiple sequence alignment of representative proteins in each of the three HDAC classes was generated by CLUSTALW. Each multiple sequence alignment was then searched for regions with strongly conserved patterns having high information content (55). Information content was determined by producing a sequence logo using WebLogo (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi). A logo is a graph that displays the amount of information at each column in the alignment and is measured in bits (reduction in uncertainty above background amino acid frequencies). The logo also shows the contribution of each amino acid to this information.

Domain analysis of HAT proteins. HAT protein sequences identified in *Arabidopsis* and in the proteomes of other organisms were analyzed for the presence of any domain present in the Pfam (Protein Family) database. The collection of Pfam hidden Markov model (HMM) profiles for domain families (version 6.5) was downloaded from the Pfam web site. Sequence profile searches were performed using the software HMMER (http://hmmer.wustl.edu/). For certain domains such as the CBP-type HAT domain, for which a Pfam model is not available, a multiple sequence alignment generated using CLUSTALW was examined for the presence of the biochemically-defined HAT domain in human CBP protein (26) and a profile HMM was constructed using programs in the HMMER package. The Predict Protein resource based on neural networks (PHD at http://maple.bioc.columbia.edu/predictprotein/) and the Discrimination of protein secondary structure server (DSC at http://bioweb.pasteur.fr/seqanal/interfaces/dsc-simple.html) was used for predicting the secondary structure for proteins. The KIX domains in CBP-type HAT proteins were further searched against a database of position-specific-scoring-matrices representing conserved structural domains (3D-pssm at http://www.sbg.bio.ic.ac.uk/~3dpssm/) to find similarity with the known KIX domain structure.

RESULTS

Identification of HDAC and HAT proteins and alternative transcripts encoded by the *Arabidopsis* genome

The *Arabidopsis* genome sequence was searched for homologs of known HDAC and HAT proteins as described in the

Materials and Methods. A total of 16 *Arabidopsis* HDAC genes and 12 HAT genes were identified (Table 1). Of the 16 HDACs, 10 belong to the RPD3/HDA1 superfamily and were named with the symbol *HDA*, four belong to the HD2 family and were given the name *HDT* ('HD-tuins'), and two belong to the SIR2 family and were named with the symbol *SRT*. Two additional members of the HDA family were found that have partial HDAC domains. Of the 12 HATs, five belong to the GNAT/MYST superfamily and were named with the symbol *HAG*, five belong to the CBP family and were named with the symbol *HAC*, and two belong to the TAF$_{II}$250 family and were named with the symbol *HAF*.

Consensus gene splicing models were first developed by comparison of several computationally determined models. Because computational methods do not predict all splice sites correctly, cDNA sequences were generated from EST clones and RT–PCR products (see Materials and Methods). In addition to revising splicing models, the cDNA sequence analysis detected multiple splicing products for three HDAC genes (*HDA2*, *HDA15* and *SRT2*) (Fig. 1). Revised coding sequences, predicted proteins and alternative splicing products are available at the Plant Chromatin Database, ChromDB (http://www.chromdb.org).

The phylogenetic and domain analyses presented here are based on alternative products designated 'alt1' (Fig. 1), each of which is predicted to encode intact, conserved HDAC domains. The HDAC domain is disrupted in alternative transcripts produced by *HDA2* and *HDA15*. *SRT2* produced six alternative transcripts via different combinations of the same splice sites, affecting a putative nuclear localization signal and the SIR2 domain. The alternative splice site in the SIR2 domain appears to be evolutionarily conserved because it also occurs in a putative ortholog in tomato. Alternative splicing in the 5′-untranslated region (5′-UTR) of SRT2alt2 and alt5 could affect translation efficiency or mRNA stability. Details are presented in Figure 1.

The RPD3/HDA1 superfamily of HDACs

A total of 10 representatives possessing the complete HDAC domain (Pfam designation PF00850) that defines the RPD3/HDA1 superfamily were identified in *Arabidopsis* (Table 1). Two additional predicted proteins, HDA10 and HDA17, were found that possess only the 30 and 40 C-terminal amino acids, respectively, of the HDAC domain.

Sequence similarity searches of a variety of eukaryotic and prokaryotic genomes, as well as other sequences available in public databases (including ESTs), led to the identification of a total of 72 RPD3/HDA1 superfamily protein sequences (including 10 in *Arabidopsis*) that possess an intact HDAC domain. For 80% of these sequences, the 300 amino acid HDAC domain constitutes more than half of the protein. For the remaining 20% of the sequences, additional sequences were present. Searching these larger proteins using Pfam (version 6.5) did not reveal any additional domains, although there is a possibility of the presence of additional domains that have not yet been identified.

Figure 2 shows an unrooted phylogenetic tree illustrating the relationships among the 72 RPD3/HDA1 superfamily proteins (listed in Table 2), produced by aligning their HDAC domains (for double-domain proteins each domain was analyzed separately). The analysis in Figure 2 is based on a

Table 1. Genes encoding HAT and HDAC homologs in *Arabidopsis*

HDAC and HAT gene family	*Arabidopsis* gene name (synonym)	MIPS accession no.	BAC clone and genomic locus	Chromosome
RPD3/HDA1	*HDA2*	At5g26040	T1N24.9	V
	HDA5	At5g61060	MAF19.7	V
	HDA6 (AtRPD3B)	At5g63110	MDC12.7	V
	HDA7	At5g35600	K2K18.5	V
	HDA8	At1g08460	T27G7.7	I
	HDA9	At3g44680	T18B22.80	III
	HDA10[a]	At3g44660	T18B22.60	III
	HDA14	At4g33470	F17M5.230	IV
	HDA15	At3g18520	MYF24.23	III
	HDA17[a]	At3g44490	F14L2	III
	HDA18	At5g61070	MAF19.8	V
	HDA19 (AtRPD3A)	At4g38130	F20D10.250	IV
HD2	*HDT1 (AtHD2A)*	At3g44750	T32N15.8	III
	HDT2 (AtHD2B)	At5g22650	MDJ22.7	V
	HDT3	At5g03740	F17C15.160/MED24.1	V
	HDT4	At2g27840	F15K20.6	II
SIR2	*SRT1*	At5g55760	MDF20.20	V
	SRT2	At5g09230	T5E8.30	V
GNAT	*HAG1 (atGCN5)*	At3g54610	T14E10.180	III
	HAG2	At5g56740	MIK19.19	V
	HAG3	At5g50320	MXI22.3	V
MYST	*HAG4*	At5g64610	MUB3.13	V
	HAG5	At5g09740	F17I14.70/MTH16.20	V
CBP	*HAC1*	At1g79000	YUP8H12R.3	I
	HAC2	At1g67220	8 F1N21.4	I
	HAC4	At1g55970	F14J16.27/T6H22.23	I
	HAC5	At3g12980	MGH6.9	III
	HAC12	At1g16710	F17F16.21	I
TAF$_{II}$250	*HAF1*	At1g32750	F6N18.20	I
	HAF2	At3g19040	K13E13.16	III

[a]Genes with partial HDAC domains.

mixture of both predicted and experimentally determined protein sequences. In order to confirm these results, the analysis was also performed using only experimentally derived sequences (i.e. those confirmed by cDNA sequences). The clustering patterns and the bootstrap support for these patterns were similar to those shown in Figure 2 (data not shown). The RPD3/HDA1 superfamily, represented by these 76 domain sequences, is divided into two major clades based on a strongly supportive bootstrap value (85%). These clades, shown in Figure 2 as two lightly shaded ovals, include three classes of eukaryotic proteins: Classes I and II, both of which have been reported previously based on a smaller number of sequences (56,57), and a new class of proteins, Class III. These Class III proteins include a recently cloned and characterized human HDAC11 (58).

The two major clades include proteins from both prokaryotes and eukaryotes. The rightmost clade includes acetylpolyamine aminohydrolase proteins from multiple species of Archaea and bacteria, suggesting that HDAC proteins in this clade could be derived from these prokaryotic proteins. The leftmost clade includes acetoin-utilizing proteins from bacteria (but no Archaea sequences), suggesting that the HDAC proteins in this clade could have originated from these bacterial proteins. Proteins from other lower eukaryotic organisms, including *Plasmodium falciparum*, *T.thermophila* and *L.major*, were present in only the leftmost clade of Figure 2. This evolutionary link between the prokaryotic proteins and the HDACs is also evident at the level of

enzymatic activity. HDACs and acetylpolyamine aminohydrolases catalyze the removal of an acetyl group from acetylated aminoalkyls by cleaving an amide bond and reconstituting the positive charge on the substrate; acetoin utilization proteins catalyze deacetylation of acetoin (32).

Class I proteins. The total number of Class I proteins found in the *Arabidopsis* genome is similar to the numbers found in other sequenced genomes (Table 4). The four *Arabidopsis* proteins lie within a cluster comprised of *S.cerevisiae* RPD3p and several animal Class I proteins with good bootstrap support (70%) (Fig. 2). Three of the *Arabidopsis* proteins, along with other plant proteins, group into two branches forming clusters A and B, each with excellent bootstrap support (100%). The proteins in cluster A (including *Arabidopsis* HDA19) are 73–80% identical at the amino acid level and may comprise an orthologous group. The proteins in cluster B (which includes *Arabidopsis* HDA6 and HDA7) are somewhat more divergent than the proteins in A (58–74% identical at the amino acid level) cluster. The strongly supported separation of clusters A and B suggests the possibility of functional diversification. Because both clusters contain dicot and monocot proteins, they would seem to have originated by gene duplication predating divergence of the monocot and dicot lineages. Immunological data indicates zmRPD3 (cluster A) and zmHD1b-II (cluster B) to be associated with human Rbap46/48 like proteins (59) found in the NuRD and SIN3 HDAC complex (60).
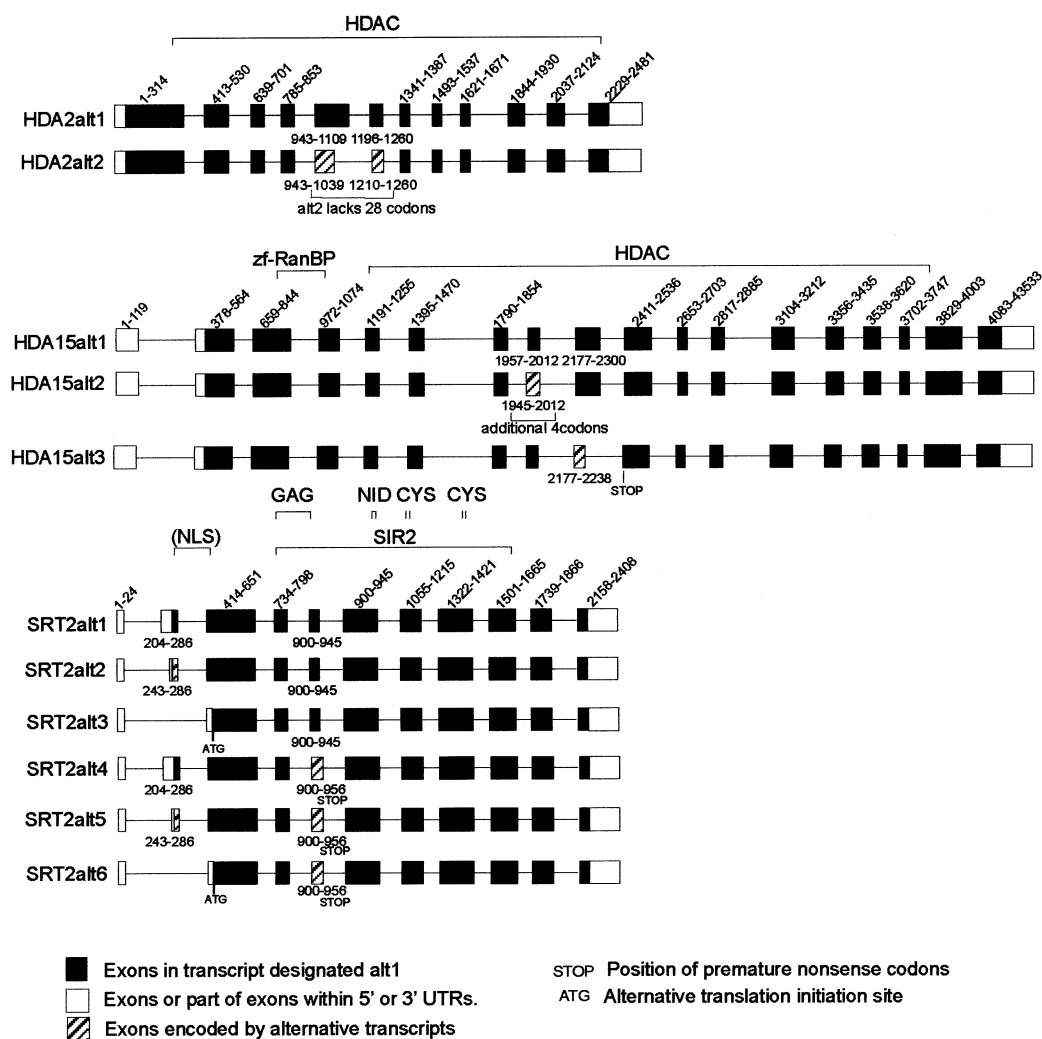
Figure 1. Alternative splicing of *HDA2*, *HDA15* and *SRT2*. Sequence coordinates indicate the position of exons within the unspliced transcripts relative to the start of the 'alt1' RT–PCR product sequences. The approximate location of predicted protein domains and conserved amino acid motifs is marked by brackets, their Pfam accessions are listed here. HDAC, the histone deacetylase domain (PF00850); SIR2, the multidomain core [including the conserved GAG, NID and CYS motifs (89) (PF02146)]; zf-RanBP, Ran-binding protein zinc finger (PF00641); (NLS), sequence similar to the bipartite nuclear localization sequence. SRT2alt3 and alt6 completely lack exon 2, which contains the predicted translation initiation codon of SRT2alt1. The nearest downstream ATG codon for translation initiation from SRT2alt3 and alt6 is located at position 446 of the unprocessed transcript. A consequence of translation initiation at this position would be a protein lacking a putative nuclear localization signal. Alternative splicing of exon 2 in SRT2alt2 and alt5 removes 39 nt of the 5′-UTR. Alternative splicing of exon 5 in SRT2alt4, alt5, and alt6 introduces a premature nonsense codon within the conserved multi-domain SIR2 core; alternative splicing at this position is conserved between *SRT2* and a putative ortholog in tomato represented by ESTs 12635152 and 12625887.

One of the *Arabidopsis* class I proteins, HDA9, is highly similar at the nucleotide level to *HDA10* and *HDA17*, both of which possess an incomplete HDAC domain. *HDA10* lies ~11 kb from *HDA9*, and the interval between these two genes contains an ORF annotated in GenBank as encoding a 'disease-resistance-like' gene. *HDA17* lies on a neighboring BAC clone, adjacent to a second copy of this 'disease-resistance-like' gene, suggesting that *HDA10* and *HDA17* were derived from *HDA9* by sequence rearrangements that duplicated part of *HDA9* and its flanking sequences. These events appear to be relatively recent in evolution, considering that the homologous regions of these three genes are 97% identical at the nucleotide level. Genetic and biochemical analyses will be required to determine whether *HDA10* and

HDA17 possess some function, perhaps related to that of *HDA9*, or are non-functional pseudogenes.

Class II proteins. The *Arabidopsis* genome possesses three Class II proteins (designated HDA5, HDA15 and HDA18), a total similar to that found in other sequenced genomes (Table 4). A subset of Class II proteins found in humans, mice, *C.elegans* and *D.melanogaster* are 'double-domain' proteins, i.e. they possess two tandem HDAC domains separated by a small, but variable, spacer region. In human and mouse proteins, each domain has been found to be an independently functional catalytic domain (57). Double-domain proteins have not been found in either *S.cerevisiae* or *S.pombe*, each of which has a single Class II protein with a
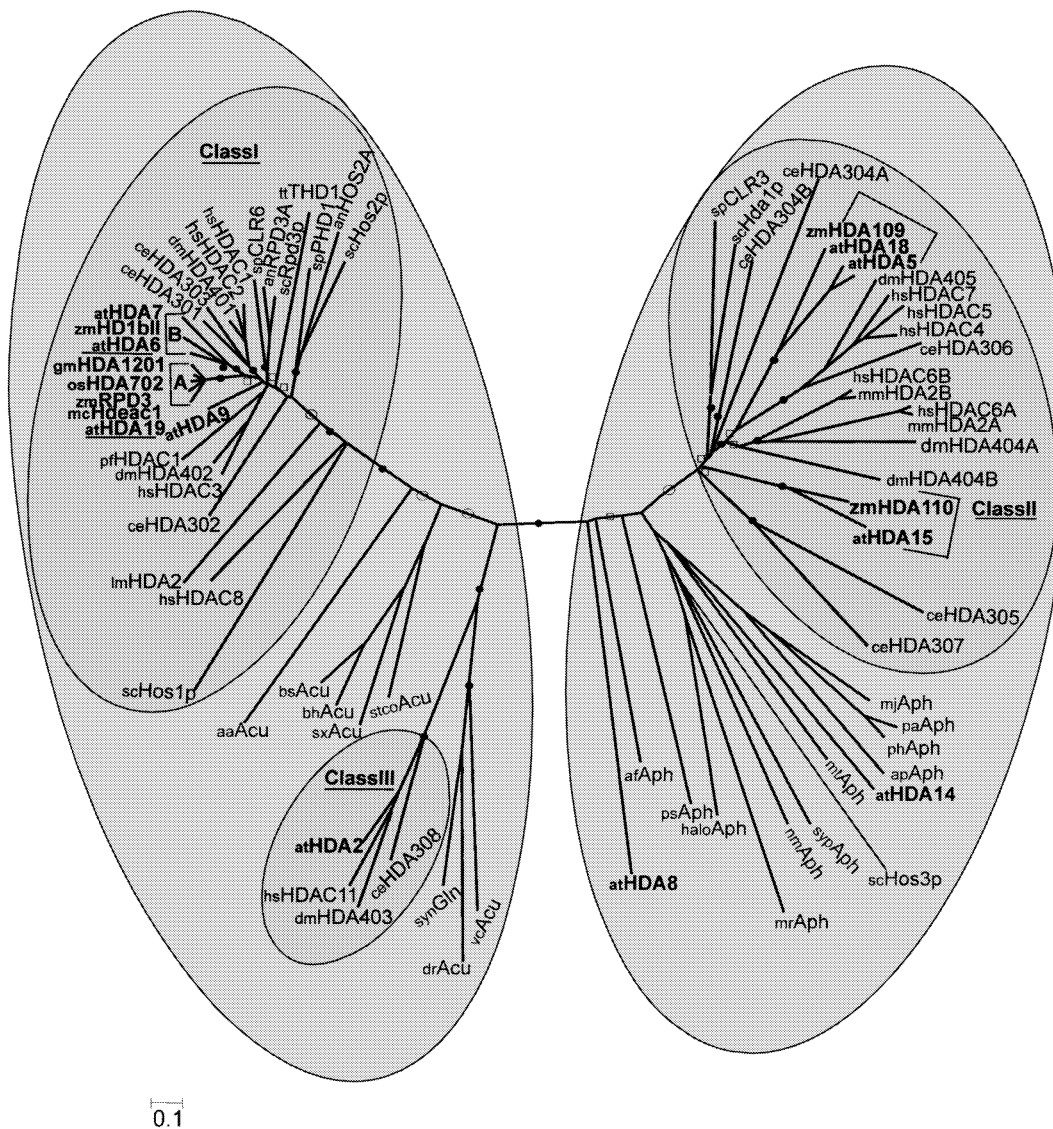
Figure 2. Phylogenetic analysis of the RPD3/HDA1 HDAC superfamily. Unrooted neighbor-joining tree of 76 RPD3/HDA1 superfamily sequences includes four double-domain sequences with each domain being analyzed separately. Confidence levels of the branching patterns are: filled circle, excellent support (>99% of bootstrap replicas); empty square, good or >70%; empty circle, majority support or >50%. Eukaryotic gene names and sequence accession numbers are listed in Table 2. The plant proteins are highlighted in bold and the three eukaryotic classes are represented in gray shaded ovals. Prokaryotic genes are represented by Acu (acetoin utilization proteins) or by Aph (acetylpolyamine aminohydrolase proteins). All the proteins have abbreviated species names as prefix. The proteins and their accession numbers are identified in Table 2. Abbreviations for species are: *Aeropyrum pernix* (ap), *Arabidopsis thaliana* (at), *Archaeoglobus fulgidus* (af), *Aquifex aeolicus* (aa), *Aspergillus nidulans* (an), *Bacillus halodurans* (bh), *Bacillus subtilis* (bs), *Caenorhabditis elegans* (ce), *Deinococcus radiodurans* (dr), *Drosophila melanogaster* (dm), *Glycine max* (gm), *Halobacterium* sp. NRC-1 (halo), *Homo sapiens* (hs), *Leishmania major* (lm), *Mesembryanthemum crystallinum* (mc), *Methanobacterium thermoautotrophicum* (mt), *Methanococcus jannaschii* (mj), *Mus musculus* (mm), *Mycoplana ramose* (mr), *Neisseria meningitides* (nm), *Oryza sativa* (os), *Plasmodium falciparum* (pf), *Pseudomonas aeruginosa* (ps), *Pyrococcus abyssi* (pa), *Pyrococcus horikoshii* (ph), *Saccharomyces cerevisiae* (sc), *Schizosaccharomyces pombe* (sp), *Staphylococcus xylosus* (sx), *Streptomyces coelicolor* (stco), *Synechococcus* PCC7002 (syp), *Synechocystis* PCC6803 (syn), *Tetrahymena thermophila* (tt), *Vibrio cholerae* (vc), *Zea mays* (zm).

single domain. Likewise, the *Arabidopsis* genome does not contain any double-domain Class II proteins. Recently, HDAC6, a human double-domain protein, has been shown to be a cytoplasmic tubulin deacetylase, not an HDAC (61).

Class II proteins are more divergent in sequence than are Class I proteins, resulting in longer, more poorly supported branches (Fig. 2), and making it impossible to definitively classify orthologous and paralogous groups. Two clusters of plant Class II proteins (indicated by brackets in Fig. 2) can be

identified by phylogenetic analysis. HDA5 and HDA18 appear to be more closely related to the double-domain proteins from animals than to HDA15, and so may act on proteins other than histones. Sequence analysis revealed the presence of putative nuclear export signals in HDA5 and HDA18. Similar nuclear export signals in human and mouse class II proteins are known to be involved in shuttling these proteins between an active state in the nucleus and an inactive, phosphorylated state in the cytoplasm (62,63). Interestingly, HDA15 contains a RanBP

Table 2. Sequence accession numbers for HAT and HDAC genes analyzed

Genes (synonyms)	Organism	GenBank accession no.	
		Protein	ESTs
HDACs			
(1) RPD3/HDA1 family			
Eukaryotes			
Class I			
atHDA6 (AtRPD3B)	*Arabidopsis thaliana*	BAB10553	
atHDA7	*Arabidopsis thaliana*	BAB09994	
atHDA9	*Arabidopsis thaliana*	CAB72470	
atHDA19 (AtRPD3A)	*Arabidopsis thaliana*	AAB66486	
zmRPD3	*Zea mays*	AAC50038	
zmHd1b-II	*Zea mays*	AAD10139	
scRPD3	*Saccharomyces cerevisiae*	P32561	
scHOS2	*Saccharomyces cerevisiae*	P53096	
scHOS1	*Saccharomyces cerevisiae*	Q12214	
ceHDA301	*Caenorhabditis elegans*	CAB03984	
ceHDA302	*Caenorhabditis elegans*	Q09440	
ceHDA303	*Caenorhabditis elegans*	CAB03224	
dmHDA401	*Drosophila melanogaster*	AAC61494	
dmHDA402	*Drosophila melanogaster*	AAC83649	
hsHDAC1	*Homo sapiens*	Q13547	
hsHDAC2	*Homo sapiens*	NP_001518	
hsHDAC3	*Homo sapiens*	NP_003874	
hsHDAC8	*Homo sapiens*	NP_060956	
spCLR6	*Schizosaccharomyces pombe*	CAA19053	
spPHD1	*Schizosaccharomyces pombe*	O13298	
osHDA702	*Oryza sativa*	AAK01712.1	
gmHDA1201	*Glycine max*		BF066371, BF324960, AW396525, AW308961 AI736699, AW460060, AW598663
mc*Hdeac1*	*Mesembryanthemum crystallinum*	AAF82385	
anRPD3A	*Aspergillus nidulans*	AAF80489	
anHOS2A	*Aspergillus nidulans*	AAF80490	
ttTHD1	*Tetrahymena thermophila*	AAG00980	
pfHDA1	*Plasmodium falciparum*	AAD22407	
lmHDA2	*Leishmania major*	CAC14522	
Class II			
atHDA5	*Arabidopsis thaliana*	NP_200914	
atHDA15	*Arabidopsis thaliana*	BAB01118	
atHDA18	*Arabidopsis thaliana*	NP_200915	
zmHDA109	*Zea mays*		AW216192, AW256098, AW258010, BE238824
zmHDA110	*Zea mays*		AW231694, BE510782
scHDA1	*Saccharomyces cerevisiae*	P53973	
ceHDA304	*Caenorhabditis elegans*	Q20296	
ceHDA305	*Caenorhabditis elegans*	CAA90401	
ceHDA306	*Caenorhabditis elegans*	AAB71243	
ceHDA307	*Caenorhabditis elegans*	CAA21669	
dmHDA404	*Drosophila melanogaster*	AAD21090	
dmHDA405	*Drosophila melanogaster*	AAF48245	
hsHDAC4	*Homo sapiens*	NP_006028	
hsHDAC5	*Homo sapiens*	NP_005465	
hsHDAC6	*Homo sapiens*	NP_006035	
hsHDAC7	*Homo sapiens*	AAF63491	
spCLR3	*Schizosaccharomyces pombe*	CAC01518	
mmHDA2	*Mus musculus*	T13964	
Class III			
atHDA2	*Arabidopsis thaliana*	AAD40129	
ceHDA308	*Caenorhabditis elegans*	CAA94910	Z71185
dmHDA403	*Drosophila melanogaster*	AAF56350	
hsHDAC11	*Homo sapiens*		BE613615, HSM802049, AA346002
Unclassified			
atHDA8	*Arabidopsis thaliana*	AAF22892	
atHDA14	*Arabidopsis thaliana*	CAB38805	
scHOS3	*Saccharomyces cerevisiae*	Q02959	
Prokaryotes			
aaAcu	*Aquifex aeolicus*	D70388	
afAph	*Archaeoglobus fulgidus*	B69266	
apAph	*Aeropyrum pernix*	BAA79001	
bhAph	*Bacillus halodurans*	BAB06956	
bsAcu	*Bacillus subtilis*	S39643	
drAcu	*Deinococcus radiodurans*	AAF10411	
haloAcu	*Halobacterium sp. NRC-1*	AAG18756	

Table 2. *Continued*

Genes (synonyms)	Organism	GenBank accession no. Protein	ESTs
mjAph	*Methanococcus jannaschii*	AAB98526	
mrAph	*Mycoplana ramose*	Q48935	
mtAph	*Methanobacterium thermoautotrophicum*	C69026	
nmAph	*Neisseria meningitides*	AAF41032	
paAph	*Pyrococcus abyssi*	B75095	
phAph	*Pyrococcus horikoshii*	H71071	
psAph	*Pseudomonas aeruginosa*	D83174	
stcoAcu	*Streptomyces coelicolor*	T36278	
sxAcu	*Staphylococcus xylosus*	Q56195	
synGln	*Synechococcus PCC7002*	CAA78367	
sypAph	*Synechocystis PCC6803*	S74557	
vcAcu	*Vibrio cholerae*	AAF95190	
(2) HD2 family[a]			
HDT1 (AtHD2A)	*Arabidopsis thaliana (Arabidopsis)*	AAB70032	
HDT2 (AtHD2B)	*Arabidopsis thaliana (Arabidopsis)*	AAC02539	
HDT3	*Arabidopsis thaliana (Arabidopsis)*	AAF70197	
HDT4	*Arabidopsis thaliana (Arabidopsis)*	AAF70198	
Hd2a	*Zea mays (maize)*	AAC61674	
Hd2b	*Zea mays (maize)*	AAF68624	
Hd2c	*Zea mays (maize)*	AAF68625	
HDT101	*Zea mays (maize)*	AAK67143	
HDT701	*Oryza sativa (rice)*		AU082425, AU68818, D22916, C72062, D15380, AU067990, D39074
HDT801	*Triticum aestivum (wheat)*		BE586047, BE404638, BE415745, BE400260
HDT802	*Triticum aestivum (wheat)*		BE518130, BE517665, BE402732, BE415484
HDT901	*Sorghum bicolor (wheat)*		AW565004, BE359652, BE356690, AW564222, BE594360, BE366313
HDT1001	*Hordeum vulgare (barley)*		BE194667
HDT1101	*Lycopersicon esculentum (tomato)*		AW218751, AI488715, AW625291, BE435034, AW092756, AW093039, AW623008, AW97980
HDT1102	*Lycopersicon esculentum (tomato)*		AW929159, AW429116, BE449829, AI776222, AW615938, AW039158
HDT1103	*Lycopersicon esculentum (tomato)*		AW037620, BE463291
HDT1202	*Glycine max (soybean)*		AW707012, BE022138
HDT1301	*Medicago truncatula (barrel medic)*		AA660660, AW688123, BE322886, AL375160,
HDT1302	*Medicago truncatula (barrel medic)*		BE318805, BE32160
HDT1401	*Solanum tuberosum (potato)*		BE343097, BE342419, BE340133, BE342346, BE922565
HDT1402	*Solanum tuberosum (potato)*		BE341470, BE343018, BE924296, BE921506, BE343072, BE473230, AW906030, AW906356
HDT1501	*Gossypium arboretum (cotton)*		AW730862
HDT1502	*Gossypium arboretum (cotton)*		AW729511
HDT1601	*Mesembryanthemum crystallinum (ice plant)*		BE035409
HDT1602	*Mesembryanthemum crystallinum (ice plant)*		BE034206
HDT1701	*Lotus japonicus*		AW720250, AV409362, AV422554
(3) SIR2 family			
atSRT1	*Arabidopsis thaliana*	BAB09243	
atSRT2	*Arabidopsis thaliana*	CAC05449	
zmSRT101	*Zea mays*		AI734474, AI770366, AI861441, AW000351, AW000357, AW202474
scHST1	*Saccharomyces cerevisiae*	P53685	
scHST2	*Saccharomyces cerevisiae*	P53686	
scHST3	*Saccharomyces cerevisiae*	P53687	
scHST4	*Saccharomyces cerevisiae*	P53688	
scSIR2	*Saccharomyces cerevisiae*	NP_010242	
ceSRT309	*Caenorhabditis elegans*	T24172	
ceSRT310	*Caenorhabditis elegans*	T22325	
ceSRT311	*Caenorhabditis elegans*	T22324	
ceSRT312	*Caenorhabditis elegans*	T25520	
dmSRT406	*Drosophila melanogaster*	AAC79684	
dmSRT407	*Drosophila melanogaster*	AAF46055	
dmSRT408	*Drosophila melanogaster*	AAF54513	
dmSRT409	*Drosophila melanogaster*	AAF56851	
hsSIRT1	*Homo sapiens*	NP_036370	
hsSIRT2	*Homo sapiens*	AAD40850	
hsSIRT3	*Homo sapiens*	AAD40851	
hsSIRT4	*Homo sapiens*	NP_036372	
hsSIRT5	*Homo sapiens*	NP_036373	
hsSIRT6	*Homo sapiens*	NP_057623	
hsSIRT7	*Homo sapiens*	NP_057622	
spSIR2	*Schizosaccharomyces pombe*	T39571	
spHST2	*Schizosaccharomyces pombe*	T40929	

Table 2. *Continued*

Genes (synonyms)	Organism	GenBank accession no.	
		Protein	ESTs
spHST4	*Schizosaccharomyces pombe*	AAD53752	
osSRT701	*Oryza sativa*	AAD42226	
taSRT803	*Triticum aestivum*		BE498844
leSRT1104	*Lycopersicon esculentum*		BG125699, BG134964, BE354229, BE450731
leSRT1105	*Lycopersicon esculentum*		BF050336, AW441312, BE354229
mtSRT1303	*Medicago truncatula*		AW686356, BM814514.1, BF647706.1

<u>**HATs**</u>
(1) GNAT-MYST family
<u>*GCN5*</u>

atHAG1 (atGCN5)	*Arabidopsis thaliana*	AAB92257	
scGCN5	*Saccharomyces cerevisiae*	NP_011768.1	
ceHAG303	*Caenorhabditis elegans*	AAF60658.1	
dmHAG401	*Drosophila melanogaster*	AAC39102.1	
hsPCAF	*Homo sapiens*	NP_003875.2	
spHAG601	*Schizosaccharomyces pombe*	T37933	

<u>*ELP3*</u>

atHAG3	*Arabidopsis thaliana*	BAB09451	
scELP3	*Saccharomyces cerevisiae*	NP_015239.1	
ceHAG304	*Caenorhabditis elegans*	CAB01454.1	
dmHAG408	*Drosophila melanogaster*	AAF51012.1	
hsHAG510	*Homo sapiens*	BAB14138.1	
spHAG602	*Schizosaccharomyces pombe*	CAB10146	

<u>*HAT1*</u>

atHAG2	*Arabidopsis thaliana*	BAB09892	
scHAT1	*Saccharomyces cerevisiae*	U33335.1	
ceHAG302	*Caenorhabditis elegans*	CAA88954.1	
dmHAG409	*Drosophila melanogaster*	AAF51953.1	
hsHAG509	*Homo sapiens*	XM_002242	
spHAG603	*Schizosaccharomyces pombe*	CAB16872.1	

<u>*MYST*</u>

atHAG4	*Arabidopsis thaliana*	BAB11428	
atHAG5	*Arabidopsis thaliana*	CAB89356	
scESA1	*Saccharomyces cerevisiae*	NP_014887.1	
scSAS2	*Saccharomyces cerevisiae*	CAA88552.1	
scSAS3	*Saccharomyces cerevisiae*	NP_009501.1	
ceHAG305	*Caenorhabditis elegans*	T19693	
ceHAG306	*Caenorhabditis elegans*	CAA96668.1	
ceHAG307	*Caenorhabditis elegans*	CAB04552.2	
ceHAG308	*Caenorhabditis elegans*	AAC78211.1	
dmMOF	*Drosophila melanogaster*	AAC47507.1	
dmHAG404	*Drosophila melanogaster*	AAF44628.1	
dmHAG405	*Drosophila melanogaster*	CAA21829.1	
dmHAG406	*Drosophila melanogaster*	AAF47164.1	
dmHAG407	*Drosophila melanogaster*	AAF56792.1	
hsTIP60	*Homo sapiens*	NP_006379	
hsMOZ	*Homo sapiens*	NP_006757.1	
hsMORF	*Homo sapiens*	AAF00095.1	
hsHAG511	*Homo sapiens*	XP_050187	
hsHBOA	*Homo sapiens*	NP_008998	
hsHAG512	*Homo sapiens*	XP_018398	
spHAG604	*Schizosaccharomyces pombe*	CAA93696.1	
spHAG605	*Schizosaccharomyces pombe*	CAA22591.1	

<u>*HPA2*</u>

scHPA2	*Saccharomyces cerevisiae*	NP_015519	
scHPA3	*Saccharomyces cerevisiae*	NP_010848	
spATS-1	*Schizosaccharomyces pombe*	NP_593494	

(2) CBP family

atHAC1	*Arabidopsis thaliana*	AAC17068	
atHAC2	*Arabidopsis thaliana*	AAB95246	
atHAC4	*Arabidopsis thaliana*	AAF79331	
atHAC5	*Arabidopsis thaliana*	AAF35947	
atHAC12	*Arabidopsis thaliana*	AAG09087	
ceCBP-1	*Caenorhabditis elegans*	P34545	
dmCBP	*Drosophila melanogaster*	T13828	
hsCBP	*Homo sapiens*	S39162	
hsp300	*Homo sapiens*	A54277	
mmCBP	*Mus musculus*	S39161	

Table 2. *Continued*

Genes (synonyms)	Organism	GenBank accession no.	
		Protein	ESTs
(3) TAF$_{II}$250 family			
atHAF1	*Arabidopsis thaliana*	AAF25977	
atHAF2	*Arabidopsis thaliana*	BAB01700	
scTAF$_{II}$145	*Saccharomyces cerevisiae*	AAA79178	
ceTAF$_{II}$250	*Caenorhabditis elegans*	CAB04907	
dmTAF$_{II}$230	*Drosophila melanogaster*	P51123	
hsTAF$_{II}$250	*Homo sapiens*	NP_004597	
spTAF$_{II}$145	*Schizosaccharomyces pombe*	CAA91179	

[a]The common names as presented in Figure 5 are included in parentheses next to the Latin name.

zinc-finger domain. Such domains have been implicated in nucleocytoplasmic transport and nuclear envelope localization (64).

HDA5 and *HDA18* occur immediately adjacent to each other on chromosome V, consistent with a gene duplication event. Their encoded proteins share 84% identity, mostly in the HDAC domain. The coding sequences of these genes share the same splice site positions throughout the HDAC domain which lies toward the 5′ end of the transcript, whereas their C-terminal regions are unrelated to each other. The C-terminal region of HDA5 does not possess any known protein domains, whereas that of HDA18 is predicted to encode a predominantly α-helical domain. This putative domain carries a leucine zipper motif and is similar to structural domains found in filamentous proteins, including coiled-coil dimers and two *S.pombe* proteins (cut3 and cut14) that are required for chromosome condensation and segregation (65).

A third gene (*At5g61050*), with partial homology to *HDA5* and *HDA18* outside the HDAC domain, was also found immediately downstream of *HDA5* (Fig. 3). *HDA5*, *HDA18* and *At5g61050* are located within a 10 kb segment on chromosome V. The five exons of *At5g61050* share similarity with some exons of HDA5 and HDA18, however, the region encoding the HDAC domain is missing in *At5g61050*, so it is not classified as an HDAC protein. The high degree of sequence identity in homologous regions of the three genes suggests two recent duplications of *HDA5* to produce the progenitors of *HDA18* and *At5g61050*. The duplication was apparently followed (or accompanied by) an internal deletion in one gene copy to form *At5g61050* and acquisition of repeated sequences elements encoding an α-helical domain in the other gene copy to form *HDA18*. This gene duplication event is not shared by all the angiosperms, and appears to be unique to a lineage within the dicots including *Arabidopsis*. Whether this event resulted in diversification of function of Class II proteins remains to be determined.

Class III: a new class of proteins in the RPD3/HDA1 superfamily. A major finding of our analysis is a new class of HDAC proteins, which we designate Class III, represented in *Arabidopsis* by HDA2. Class III includes predicted proteins HDA403 from *D.melanogaster*, HDA308 from *C.elegans* and HDAC11, an EST contig from humans (Fig. 2) that has been recently identified (58). These proteins are conserved at the amino acid level, being 45% or more identical in pairwise sequence alignments. Additional members of this class were found in the EST database, but were not included in our analysis because their HDAC domains were incomplete. Class III proteins are a part of a cluster that includes three bacterial sequences encoding acetoin utilization proteins (vcAcu and drAcu) and a cyanobacteria glutamine synthetase protein (synGln) (Fig. 2), with good bootstrap support (99%). The presence of a well supported cluster of diverse proteins is consistent with a novel function for class III HDAC proteins in higher eukaryotes, possibly of bacterial origin. No class III proteins were detected in fungal genomes.

Multiple sequence alignments of Classes I, II and III proteins identified conserved motifs within the HDAC domain, with some amino acids common to all HDAC classes and others unique to a particular HDAC class (Fig. 4). A conserved but distinct pattern of amino acids for Class III proteins is evident, providing additional support for a novel biological function for these proteins.

Unclassified proteins. The *Arabidopsis* genome encodes two additional HDAC proteins in the RPD3/HDA1 superfamily, HDA8 and HDA14. Although these proteins fall into the same major clade as Class II proteins, they do not cluster with them (Fig. 2). Instead, they are present in a poorly supported group of highly diverse proteins that includes acetylpolyamine aminohydrolases from the Archaea, as well as *S.cerevisiae* HDAC protein Hos3p. The low sequence similarity between *S.cerevisiae* Hos3p and *Arabidopsis* HDA8 and HDA14 and the poor bootstrap support for this grouping indicates that these proteins are not closely related. Searches of existing genome and EST databases, including plant sequences, using Hos3p, HDA8 and HDA14 as query sequences did not identify any additional proteins in this group.

To determine whether the sequences from Archaea and bacteria influence the classification of these eukaryotic proteins, the tree was regenerated without these sequences. In the resulting tree, *S.cerevisiae* Hos3p and *Arabidopsis* HDA14 protein moved into the class II cluster, but HDA8 did not. This test revealed that *S.cerevisiae* Hos3p and *Arabidopsis* HDA14 can not be assigned to any definitive cluster, but appear to be relatives of Class II proteins. *Arabidopsis* HDA8 seems to be more closely related to prokaryotic acetylpolyamine aminohydrolase proteins than to Class II; it is possible that this protein might have acetylpolyamine deacetylating activity or other deacetylating activity rather than histone deacetylation activity. In the motif analysis of all three HDAC classes shown in Figure 4, Hos3p,
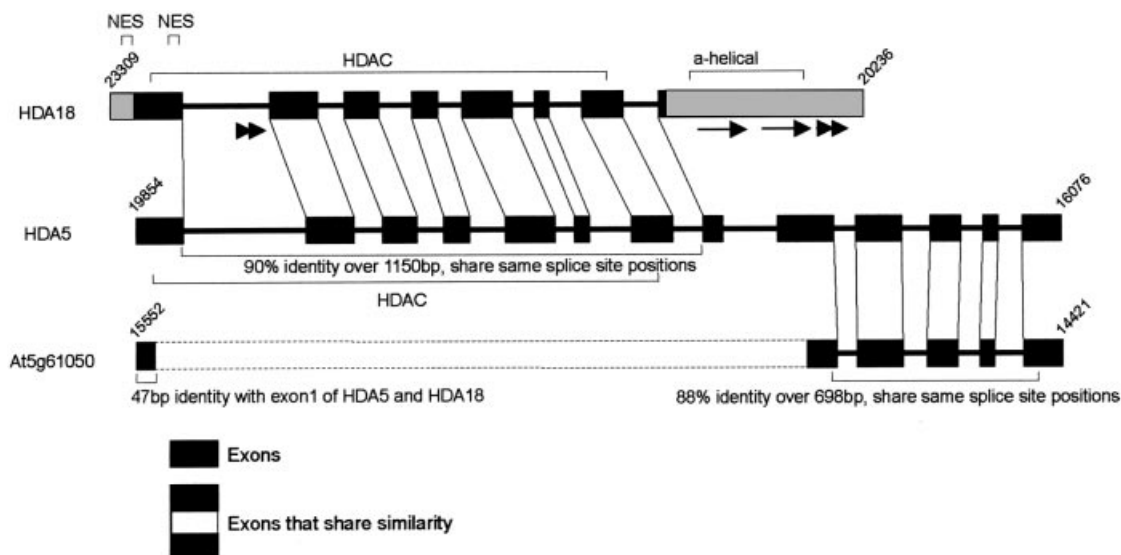
Figure 3. Schematic representation of the exon–intron and domain organization of the *HDA18-HDA5-At5g61050* gene cluster on chromosome V. Coordinates indicate the position of the start and stop codons of the three genes in the P1 clone MAF19 (accession no. AB006696). The approximate location of predicted protein domains is marked by brackets. The dotted line indicates the missing HDAC domain in *At5g61050*. NES, nuclear export signal. Arrows indicate nucleic acid sequence repeats in *HDA18*.

HDA8 and HDA14 share the conserved amino positions of Class II proteins, corresponding with their location in the same major clade as Class II proteins.

The HD2 family: unique to plants

Plants possess a family of HDAC proteins, the HD2 family, which is not found in animals or fungi (40) and is distantly related to *cis–trans* isomerases found in insects, *S.cerevisiae* and parasitic apicomplexans (42). Using maize HD2 as a query, four candidate proteins, HDT1, HDT2, HDT3 and HDT4, were identified in the *Arabidopsis* proteome (Table 1). The conserved N-terminus of these proteins contains the HD2-type HDAC domain of approximately 100 amino acids. The proteins are comprised of a conserved N-terminal domain, a central acidic domain and variant C-terminal domain. Two of these proteins, HDT1 and HDT2, have been analyzed in a recent paper showing that antisense silencing of *HDT1* results in aborted seed development (41). A sequence comparison of *Arabidopsis* and maize HD2-type proteins has been made by Dangl *et al.* (66).

Plant EST sequence databases were searched to find HD2-type HDAC proteins in other plant species (listed in Table 2 and Fig. 5). Comparison of the HDAC domains of these proteins revealed a series of highly conserved motifs within the HDAC domain. A phylogenetic analysis of the nucleotide sequences encoding these conserved motifs in the HDAC domains was performed, producing the tree shown in Figure 5. A similar analysis using protein sequences produced a tree with similar topology and the same major features although with varying but somewhat lower bootstrap support than the DNA tree. This analysis permits two general observations to be made concerning the evolution of the HD2 gene family in plants. First, dicot and monocot sequences are separated into two distinct clades strongly supported by bootstrap analysis (98%), indicating that a single HD2 gene in the ancestor of

monocots and dicots gave rise to all HD2 proteins in these groups. Secondly, the clustering pattern in dicots is consistent with a gene duplication event occurring before the diversification in dicot evolution that produced the families Solanaceae (tomato and potato), Malvaceae (cotton) and Aizoaceae (ice plant), although this conclusion is only weakly supported by bootstrap analysis (<50%). More recent duplications that are strongly supported by bootstrap analysis are also evident in several species [e.g. *Arabidopsis* HDT1 and HDT2 (100%), barrel medic HDT1301 and HDT1302 (90%), and maize HD2a, HD2b and HD2c (100%)]. It will be interesting to determine whether the considerable amount of genetic diversification of the HD2 family has been accompanied by functional diversification.

The SIR2 family of HDACs

Plants possess representatives of the SIR2 family of NAD-dependent HDAC proteins, known as sirtuins. Sirtuins occur across a wide range of organisms, including prokaryotes, fungi, plants and animals and are defined by a 175 amino acid domain (Pfam designation PF02146) comprised of a series of conserved motifs. Based on variation in this domain, the eukaryotic proteins fall into four main classes (31). A fifth class is present in some prokaryotes, but most prokaryotic sirtuins fall into Classes II and III (31). A search of the *Arabidopsis* genome identified two SIR2 family proteins, SRT1 and SRT2, fewer than are found in fungi and animals (Table 4).

In order to identify additional plant sequences for use in a phylogenetic analysis, *Arabidopsis* SRT1 and SRT2 proteins were used as queries of plant EST collections, revealing six related proteins (Table 2 and Fig. 6). Phylogenetic analysis of all plant SIR2 homologs and homologs from representative species in the Frye (31) classification of SIR2-like proteins is shown in Figure 6. Of the four classes of SIR2 proteins, plant
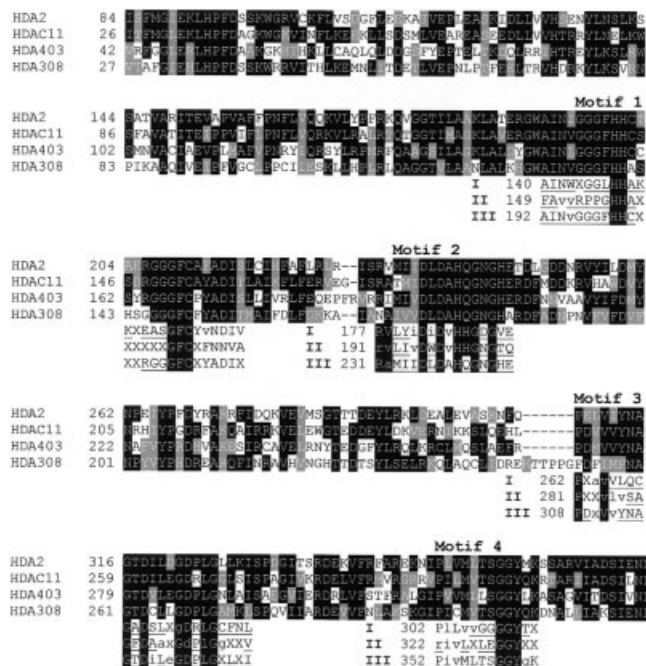
Figure 4. Class III proteins in the RPD3/HDA1 protein superfamily have distinct motifs in the HDAC domain. Alignment of the HDAC domain of *Arabidopsis* HDA2 protein with human HDAC11, *D.melanogaster* HDA403 and *C.elegans* HDA308. These proteins and their accession numbers are identified in Table 2. Shading was done based on degree of identity or conservation using the Genedoc program. Also shown below the multiple sequence alignment is a second alignment of consensus motifs found in the proteins in all the three classes of HDACs identified in Figure 2. These motifs represent the most highly conserved sequence positions in the HDAC domain. The consensus motif for each class was identified by generating a logo sequence. Each class of proteins is indicated by a consensus of the sequences in that class: black boxes, positions conserved across all three classes; underlined, positions highly conserved within a class; upper case letters, 98% conserved within a class; lower case letters, 60% conserved within a class; X, variable positions. The amino acid positions in each sequence class refer to the location of these motifs in *Arabidopsis* HDA19 (Class I), HDA5 (Class II) and HDA2 (Class III) proteins.

Figure 5. Maximum likelihood analysis of the plant HD2 family nucleic acid sequences. This analysis is based upon a codon-by-codon alignment of the first 273 positions of the maize HD2 cDNA sequence, corresponding to the HDAC domain, with other plant cDNA and EST sequences listed in Table 2. The common name for each species is listed in parentheses and where common names are not available, the Latin name is included. The gene names and their accession numbers are identified in Table 2. Confidence levels for the tree branches that are best supported by bootstrap analysis are shown as percentages.

proteins are only found within divergent plant lineages in Classes II or IV. Both classes contain plant and animal proteins but no fungal proteins (Table 4). Class IV includes two divergent animal lineages represented in flies and humans. All plant Class IV proteins cluster in a single, less divergent lineage associated with one of these animal lineages. Both plants and animals have a single lineage of Class II proteins. No plant proteins cluster with proteins of Class I, which includes all five *S.cerevisiae* sirtuins, as well as homologs in animals and *S.pombe*.

Representation of HATs in the GNAT/MYST superfamily in the *Arabidopsis* genome

In the GNAT/MYST superfamily of HAT proteins, GNAT proteins are defined by the presence of a HAT domain (Pfam designation PF00583) which is comprised of four motifs, A–D, whereas MYST proteins possess only the A motif of the HAT domain (22).

The GNAT family is generally considered to be comprised of four subfamilies designated GCN5, ELP3, HAT1 and HPA2. The HPA2 subfamily has *in vitro* histone acetylation
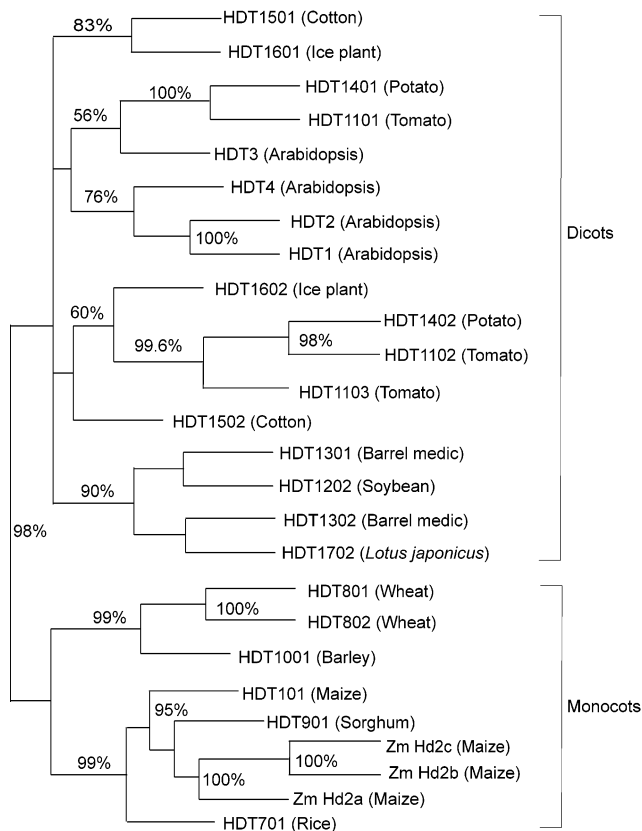
activity (67), but it is not yet known whether these proteins play any role in the control of gene expression. In the *Arabidopsis* genome, we identified a single homolog of each of the GCN5, ELP3 and HAT1 subfamilies (HAG1, HAG3 and HAG2, respectively) and no homolog of the HPA2 subfamily. HAG1 (atGCN5) and its associated adaptor proteins [similar to yeast SAGA complex (22)] in *Arabidopsis* have been known for their involvement in cold regulated gene expression (68). Searches of the *S.cerevisiae*, *S.pombe*, *D.melanogaster* and *C.elegans* genomes, as well as the nearly complete human genome, also identified a single representative of the GCN5, ELP3 and HAT1 subfamilies in each; only fungi were found to possess the HPA2 subfamily (Table 4). Thus, *Arabidopsis* appears to have the same representation of GNAT family HATs as do animals, suggesting that the plant proteins may form complexes similar to those formed in yeast and animals (69).

The *Arabidopsis* genome was found to encode two MYST family proteins, HAG4 and HAG5. Fungal genomes were found to have two to three, and animal genomes four to six, MYST family proteins. Thus, the number of plant MYST family representatives is within the range found in other
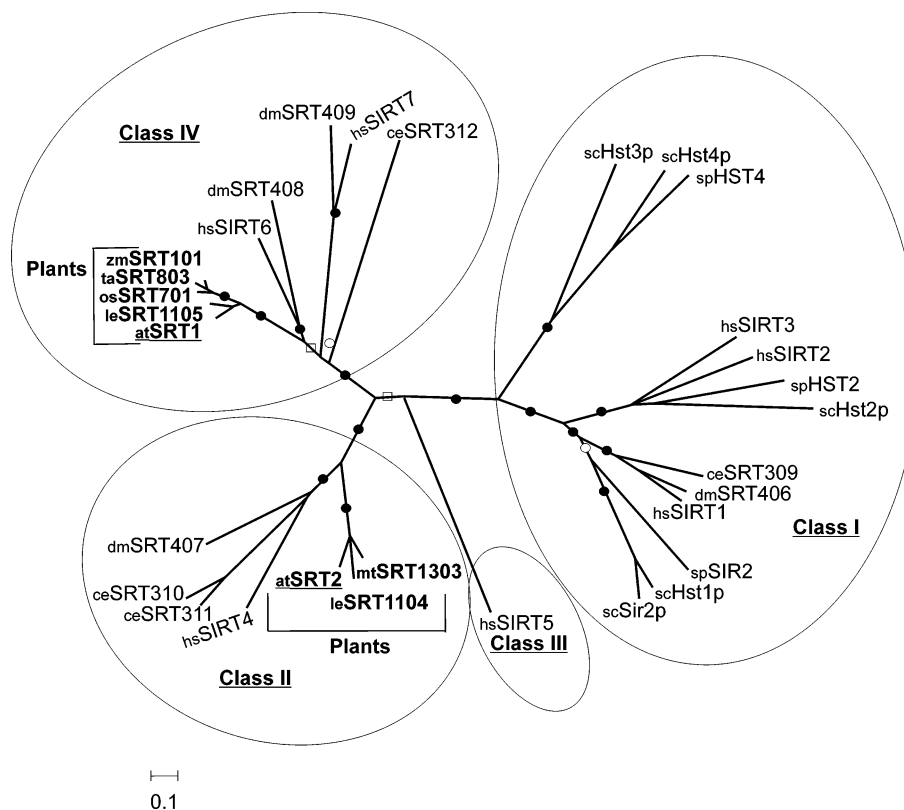
Figure 6. Phylogenetic analysis of plant SIR2 proteins. Unrooted neighbor-joining tree of 31 SIR2-related proteins shows the four previously identified classes of SIR2 proteins. The two plant protein clusters are highlighted in bold. Confidence levels of the branching patterns are: filled circle, excellent support (>99% of bootstrap replicas); empty square, good or >70%; empty circle, majority support or >50%. The genes and their accession numbers are identified in Table 2. Abbreviations for species are: *Arabidopsis thaliana* (at), *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Homo sapiens* (hs), *Lycopersicon esculentum* (le), *Medicago truncatula* (mt), *Oryza sativa* (os), *Saccharomyces cerevisiae* (sc), *Schizosaccharomyces pombe* (sp), *Triticum aestivum* (ta), *Zea mays* (zm).

eukaryotic organisms, though at the lower end of this range, and below the numbers found in animals (Table 4).

The CREB-binding protein (CBP) family of HATs

The CBP family of HAT proteins is comprised of large, multi-domain proteins (Fig. 7A) which, until recently, had been reported only in animals. The histone acetylation domain of the CBP family is unrelated to that of the GNAT/MYST superfamily; we refer to this as the CBP-type HAT domain. The *Arabidopsis* genome encodes five CBP-type HAT domain proteins (HAC1, HAC2, HAC4, HAC5 and HAC12), whereas the number of CBP proteins predicted in animals is only one to two (Table 4). The absence of the CBP family in fungi suggests that this type of protein was lost during the evolution of fungi.

Phylogenetic analysis of the plant and animal CBP-type HAT domains indicates an early divergence of HAC2 from the lineage leading to the other four *Arabidopsis* HAC proteins (Fig. 7B). Consistent with this divergence, *in vitro* assays of HAC2 did not detect any HAT activity, whereas it was readily detected for HAC1 (70). Similarly, HAC4 has diverged significantly from HAC1, HAC12 and HAC5. Interestingly, the HAT domains of human and mouse CBP proteins are 96% identical, whereas the two closest *Arabidopsis* CBP paralogs (HAC1 and HAC12) are only 90% identical in the HAT domain.

The domain architecture of CBP-type HAT proteins differs between plants and animals (Fig. 7A) in four major respects. (i) Bromodomains. As was noted also by Bordoli *et al.* (70), plant CBP-type HATs lack a bromodomain. The role of the bromodomain in the animal proteins is to bind acetylated histones (71). The lack of a bromodomain in the plant proteins suggests that these proteins utilize a different domain to perform this function or that another bromodomain protein acts as a bridge between acetylated histones and CBP-type HATs. (ii) KIX domains. All animal CBP-type HAT proteins possess a KIX domain by which they bind the nuclear factor CREB (72). Bordoli *et al.* (70) reported that the *Arabidopsis* proteins lack KIX domains. However, we found a weakly defined KIX-like domain in four of the five *Arabidopsis* proteins (Fig. 7A). The KIX domain is known to be comprised of three α-helices joined by connecting loops (73). The plant KIX-like domains from HAC1, HAC5 and HAC12 have three α-helices with about the same spacing as in the animal KIX domain, whereas HAC4 has two α-helices. A search of all four plant KIX-like sequences against a database of position-specific-scoring-matrices representing conserved structural domains (3D-pssm) produced a match with the matrix representing the KIX domain. Interestingly, the location of the KIX domain relative to the TAZ-type zinc finger domain in the animal proteins differs from the location of the KIX-like
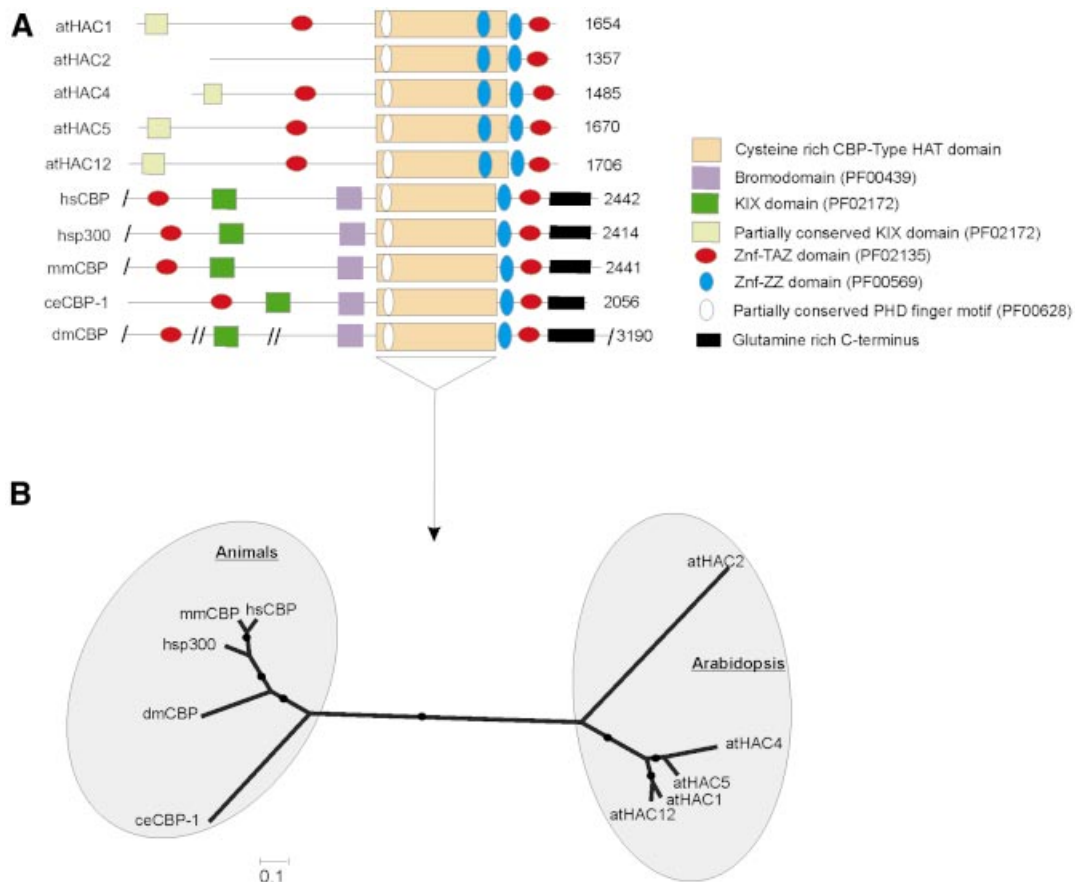
Figure 7. Domain architecture of the CBP-type HAT family and phylogenetic analysis of their HAT domains. (**A**) Schematic representation of the domain organization of *Arabidopsis* and animal CBP proteins. Different domains are identified by different symbols and colors, and are shown at their approximate relative location in the protein sequence. The protein lengths are listed on the right. //, indicates position of extra sequence; /, indicates more sequence at the N- and C-terminus. The CBP-type HAT domain is conserved throughout its length between plants and animals, however, in plants a ZZ-type zinc finger domain is inserted near the C-terminus of the HAT domain. The Pfam accession number for the domain profiles is indicated in parentheses. (**B**) Unrooted neighbor-joining tree of 10 CBP-type HAT proteins based on the HAT domain. Distinct animals and *Arabidopsis* clusters are shown by two shaded ovals. Confidence levels of the branching patterns are: filled circle, excellent support (>99% of bootstrap replicas). The genes and their accession numbers are identified in Table 2. Abbreviations for species are as follows: *Arabidopsis thaliana* (at), *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Homo sapiens* (hs), *Mus musculus* (mm).

domain relative to this domain in the plant proteins (Fig. 7A). (iii) Zinc finger domains. ZZ and TAZ types of zinc finger domains are found only in CBP-type proteins and are known to mediate protein–protein interactions with transcription factors (74). Animal CBP-type proteins have one ZZ-type zinc-finger domain located near the C-terminal end of the CBP-type HAT domain, whereas all the plant proteins have two such domains, one of which lies within the HAT domain. Both plant and animal proteins possess two TAZ-type zinc fingers, one on each side of the HAT domain. The N-terminal TAZ-type domain is located at a greater distance from the HAT domain in the animal proteins than in the plant proteins. (iv) Glutamine-rich regions. Animal CBP-type HATs possess an extensive glutamine-rich region near the C-terminus, which harbors the binding site for the unrelated mammal-specific HATs, SRC-1 and ACTR (75,76). Plant proteins lack such a C-terminus (70) (Fig. 7A), which is not particularly surprising given that plants lack this family of HATs (22), which we have confirmed by searching the *Arabidopsis* genome.

The TAF$_{II}$250 family of HAT proteins

The human TAF$_{II}$250 protein is a subunit of transcription factor IID (TFIID) (77) and has a HAT domain unrelated to the GNAT/MYST and CBP-type HAT domains. Using animal protein sequences as queries, two *Arabidopsis* TAF$_{II}$250 homologs were identified and designated HAF1 and HAF2 (Table 1). These long predicted proteins are 72% identical to each other at the amino acid level. A similar search against the complete *C.elegans*, *D.melanogaster*, *S.pombe* and *S.cerevisiae* genomes, and the nearly complete human genome, identified only one homolog in each organism. Hence, *Arabidopsis* is unusual in encoding two predicted TAF$_{II}$250 HAT proteins.

The human and *D.melanogaster* proteins have a 260 amino acid long TAF$_{II}$250-type HAT domain (28). A multiple sequence alignment revealed the presence of a domain in the *Arabidopsis* and *C.elegans* proteins that is similar in length to the human and *D.melanogaster* TAF$_{II}$250 HAT domains. This
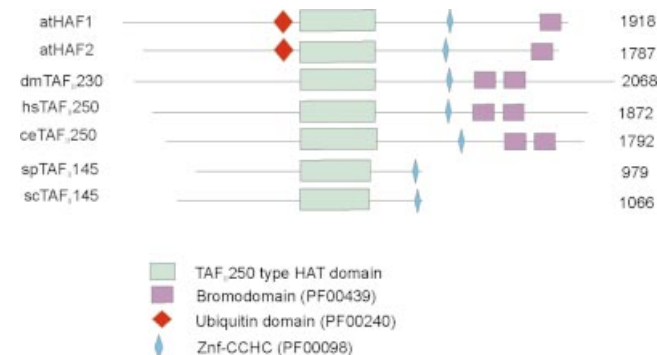
Figure 8. Domain architecture of the $TAF_{II}250$ proteins. A schematic representation is shown of the domain organization of *Arabidopsis* and animal $TAF_{II}250$ proteins aligned by the N-terminus of the HAT domain. Different domains are identified by different symbols and colors, and are shown at their approximate relative locations in the protein sequences. The protein lengths are listed on the right. Pfam accession numbers for the domain profiles are indicated in parentheses underneath the alignment. The sequences and their accession numbers are identified in Table 2. Abbreviations for species are: *Arabidopsis thaliana* (at), *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Homo sapiens* (hs), *Saccharomyces cerevisiae* (sc), *Schizosaccharomyces pombe* (sp).

domain is 45–75% identical among this group of organisms. A similar type of HAT domain in *S.cerevisiae* is shorter in length, lacking amino acids at the C-terminus of the domain, but still has HAT activity (28). Thus, the plant proteins are more similar to the animal proteins in this respect than to the fungal proteins.

The overall domain architecture of $TAF_{II}250$-type proteins in plants, animals and fungi is presented in Figure 8 and shows three interesting features. (i) In addition to the $TAF_{II}250$-type HAT domain, the human and *D.melanogaster* proteins have two bromodomain copies on the C-terminal side of the HAT domain, whereas the *Arabidopsis* proteins possess only a single bromodomain in this region. (ii) A zinc-finger-type C_2HC domain is located at an approximately equal distance downstream of the HAT domain in each of the seven sequences, presumably with a role in DNA binding or protein–protein interactions. (iii) A conserved ubiquitin signature at the N-terminal side of the HAT domain was found in each *Arabidopsis* protein, but not in the animal or fungal proteins. No other Pfam ubiquitin-associated domain was found in the animal or fungal proteins. In *D.melanogaster*, the region of $TAF_{II}230$ responsible for ubiquitin-conjugating activity for histone H1 overlaps the $TAF_{II}250$-type HAT domain (78), and these regions are presumably present in the highly conserved $TAF_{II}250$-type HAT domains in the *Arabidopsis* proteins.

Arabidopsis bromodomain proteins

Because of the disparity in number and occurrence of bromodomain between plant and animal HAT proteins, we performed a preliminary search for all bromodomain-containing proteins in *Arabidopsis* using the bromodomain HMM profile from Pfam. Twenty-nine *Arabidopsis* bromodomain proteins were found (Table 3), all of which had only a single bromodomain. Although the majority of bromodomain proteins in fungi and animals also possess a single

bromodomain, many have from two to five bromodomains (79). Thus, plants lack multi-bromodomain proteins.

Bromodomain proteins exist in diverse classes defined according to the presence of other domains in those proteins (80). We performed a domain analysis of the 29 *Arabidopsis* proteins for other Pfam domains. Unlike fungi or animal bromodomain proteins that commonly possess zinc fingers (81), none of the *Arabidopsis* bromodomain proteins possess any type of zinc finger, such as a PHD domain, with the exception of the C_2HC zinc knuckle observed in HAF1 and HAF2. As noted previously, bromodomains are often associated with certain other domain classes in other organisms, whereas the same associations are not observed in *Arabidopsis* proteins. In the case of CBP-type HATs, animal proteins contain both a bromodomain and multiple zinc fingers, whereas *Arabidopsis* CBP-type HATs contain only zinc fingers (Fig. 7A). Another interesting difference is that an animal homolog of fly Trithorax-related proteins (mouse protein AAK26242) has a bromodomain associated with a SET domain, whereas no bromodomain protein in *Arabidopsis* contains a SET domain. Thus, the utilization of bromodomains differs not only in HATs, but also in other types of chromatin protein, in plants as compared to animals and fungi.

DISCUSSION

The *Arabidopsis* genome is predicted to encode 16 HDAC and 12 HAT proteins, which is somewhat more than the number of such genes found in other sequenced eukaryotic genomes (Table 4). The distribution among different homology groups of HDACs and HATs in *Arabidopsis* differs from that in fungi and animals in several respects, as summarized in Table 4. Phylogenetic and domain analyses of these proteins predict that some have functionally diversified during plant evolution, whereas others appear to have conserved the functions of their ancestral homologs. In addition, the observed alternative mRNA splicing of three HDAC genes suggests the possibility of further functional diversification of these protein families and a complex relationship between gene number and the actual number of gene products encoded within plant genomes, as also appears to be the case for the human genome (82).

The most obvious indication of diversification of histone acetylation/deacetylation functions in plants as compared to animals and fungi is that plants possess a unique family of HDACs, the HD2 gene family (66). Because no homologs of HD2 are found in any animal or fungal genome, these proteins could serve a novel plant function or could provide a function similar to one carried out by a different type of HDAC in animals and fungi. Our phylogenetic analysis is consistent with a greater degree of functional diversification in the HD2 family in dicots than monocots. This analysis suggests that a gene duplication event may have occurred early in dicot evolution and that further diversification has occurred in the lineage leading to *Arabidopsis*, suggesting functional diversification of the HD2 subfamily.

We found that the SIR2 family is under-represented in plants as compared to fungi and animals. It is possible that the HD2 family has taken over some of the function(s) of sirtuins. Another possibility is that alternative splicing has provided added diversity of sirtuin functions. Plants possess two classes

Table 3. *Arabidopsis* bromodomain proteins and associated domains within these proteins

Proteins	Number of bromodomains	Associated domains	GenBank protein accession no.	MIPS accession no.
BRD1	1	AT hook	AAF80635.1	At1g20670
BRD2	1	–	AAF16663.1	At1g76380
BRD3	1	Myb-DNA-binding domain	AAC16089.1	At2g44430
BRD4	1	Myb-DNA-binding domain	AAB71473.1	At1g52110
BRD5	1	–	AAG50696.1	At1g58025
BRD6	1	Myb-DNA-binding domain	CAB75926.1	At3g60110
BRD7	1	–	CAB67626.1	At3g57980
BRD8	1	AAAAtpase domain	AAF29398.1	At1g05910
BRD9	1	Myb-DNA-binding domain	AAB88641.1	At2g42150
BRD10	1	–	AAD03360.1	At2g15030
BRD11	1	4 WD40 repeats	BAB09913.1	At5g49430
BRD12	1	5 WD40 repeats	AAC62845.2	At2g47410
BRD13	1	–	BAB10578.1	At5g55040
CHR2	1	ATP binding and helicase domain	AAC62900	At2g46020
HAF1	1	$TAF_{II}250$ HAT domain, C_2HC zinc finger	AAF25977	At1g32750
HAF2	1	$TAF_{II}250$ HAT domain, C_2HC zinc finger	BAB01700	At3g19040
GTE1	1	ET domain	AAC12830	At2g34900
GTE2	1	ET domain	CAB89388	At5g10550
GTE3	1	ET domain	AAF18720	At1g73150
GTE4	1	ET domain	AAF80220	At1g06230
GTE5	1	ET domain	AAF97259	At1g17790
GTE6	1	ET domain	CAC07919.1	At3g52280
GTE7	1	ET domain	BAA98182.1	At5g65630
GTE8	1	ET domain	BAB02121.1	At3g27260
GTE9	1	ET domain	CAB87766.1	At5g14270
GTE10	1	ET domain	BAB10737.1	At5g63330
GTE11	1	ET domain	AAF01563.1	At3g01770
GTE12	1	ET domain	BAA97526.1	At5g46550
HAG1	1	HAT domain	AAB92257	At3g54610

Table 4. Summary of HDAC and HAT homologs found in plants, fungi and animals

Homology groups	Plants	Fungi		Animals	
	At	Sc	Sp	Dm	Ce
HDAC homology groups					
RPD3/HDA1 family (*HDA* genes)					
Class I	4	3	2	2	3
Class II	3	1	1	2	4
Class III	1	0	0	1	1
Unclassified	2	1	0	0	0
HD2 family (*HDT* genes)	4	0	0	0	0
SIR2 family (*SRT* genes)					
Class I	0	5	3	1	1
Class II	1	0	0	1	2
Class IV	1	0	0	2	1
Total HDAC homologs	16	10	6	9	12
HAT homology groups					
GNAT-MYST superfamily (*HAG* genes)					
GNAT family					
GCN5	1	1	1	1	1
ELP3	1	1	1	1	1
HAT1	1	1	1	1	1
HPA2	0	2	1	0	0
MYST family	2	3	2	5	4
CBP family (*HAC* genes)	5	0	0	1	1
$TAF_{II}250$ family (*HAF* genes)	2	1	1	1	1
Total HAT homologs	12	9	7	10	9

The genes and their accession numbers are identified in Table 2. Organism abbreviations as follows: *Arabidopsis thaliana* (At), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm).

of sirtuins that are also represented in animals, but not in fungi. The SIR2 family has major biological significance including determining the life span of *S.cerevisiae* cells and aging in animals (37,38), but its function in plants remains unknown.

Phylogenetic analysis of the RPD3/HDA1 superfamily revealed another similarity between plants and animals, but not fungi, in that both possess representatives of Class III proteins, whereas fungi have none. It is possible that these unclassified proteins have an activity other than histone deacetylation.

The degree of evolutionary change differs significantly among HAT gene families (Table 4). At one extreme, gene number in three subfamilies of the GNAT family is completely conserved. The fourth GNAT subfamily (HPA2) is specific to fungi. At the other extreme, the CBP family has been amplified in plants to five genes as compared to a single representative in most animals, and none in fungi. There are two TAF$_{II}$250-type proteins in plants as compared to one in fungi and animals. The size of the MYST family ranges from two in *Arabidopsis* and *S.pombe* to five in *D.melanogaster*. Domain and phylogenetic analyses of the CBP-type proteins revealed three classes of these proteins in plants, as compared to a single class in animals, as well as major differences in domain architecture between plant and animal proteins. In addition, HAC2 appears to have diverged early in plant evolution. Its HAT domain appears to have evolved more rapidly than the lineage from which it diverged, and its N-terminal region lacks domains present in other plant CBPs, consistent with *in vitro* experiments that suggest it does not have HAT activity (70). HAC4 also appears to have evolved more rapidly than the lineage from which it diverged and has distinct features in its N-terminal region.

The *Arabidopsis* genome encodes proteins homologous to factors in yeast and mammals that associate with HAT complexes SAGA and ADA (GCN5, ADA2 homologs) and HDAC complexes NuRD and SIN3 (RPD3-like, Mi-2, MBD, RbAP46/48 homologs) (see http://www.chromdb.org), suggesting that the *Arabidopsis* GNAT family HATs and RPD3 family HDACs form complexes similar to those in other organisms (68). In contrast, an analysis of the domain structure of *Arabidopsis* CBP and TAF$_{II}$250 proteins suggests that these proteins may form complexes different from their animal relatives. Plant CBP proteins lack a bromodomain, whereas animal CBPs have one, and plant TAF$_{II}$250 proteins have a single bromodomain, compared to the two bromodomains found in their animal homologs. The plant proteins may utilize a different domain that serves the function of the second animal bromodomain in these proteins or may interact with a different bromodomain protein. A precedent for this possibility can be seen in TAF$_{II}$145 proteins in *S.cerevisiae* which do not have a bromodomain, but that interact with Bdf1p. Bdf1p contains two bromodomains and may substitute for the missing C-terminal sequences in the *S.cerevisiae* Taf$_{II}$145p protein (83). Although we identified a number of bromodomain-containing proteins in *Arabidopsis*, none of these have enough sequence similarity to Bdf1p to suggest a homologous function. However, the *Arabidopsis* genome encodes two proteins (SGA1 and SGA2; www.chromdb.org) that are similar to yeast Asf1p. Asf1p interacts with Bdf1p, and its counterpart in humans, CIA/ASF1, interacts with the two bromodomains of human TAF$_{II}$250 (84). Thus, the possibility exists that one of the many bromodomain proteins in *Arabidopsis* plays the role of Bdf1p and interacts with an Asf1p homolog. Interestingly, the *Arabidopsis* genome encodes two TAF$_{II}$250 proteins and two ASF1 homologs, whereas yeast and animals encode only one of each.

In addition, our analysis of the *Arabidopsis* genome sequence revealed that all *Arabidopsis* bromodomain-containing proteins have only a single bromodomain, in contrast to some animal and *S.cerevisiae* bromodomain proteins that have multiple copies, ranging from two to five bromodomains. Many bromodomain-containing transcription factors also possess a conserved PHD finger (85–87). Our finding of the absence of such a conserved feature in *Arabidopsis* bromodomain proteins suggests that the manner in which bromodomains are deployed and utilized differs between plants and animals.

Alternative splicing of two RPD3/HDA1 family genes and one SIR2 family gene could indicate alternative regulatory functions of the RNAs or the predicted protein products, different enzymatic or structural functions for the proteins, or no function at all. Alternative splicing that is conserved in *Arabidopsis* and tomato SIR2 homologs is suggestive evidence for function of an alternative splicing product, but it is also possible that this is a non-functional splicing product, merely an incidental consequence of a conserved RNA sequence.

These evolutionary differences in fundamental chromatin components among plants, animals and fungi suggest that there may be more evolutionary plasticity and more functional diversification in core chromatin components than might have been anticipated just a few years ago. This diversity is likely to reflect important differences in the manner in which chromatin controls gene expression in these three major kingdoms of eukaryotes, and supports the suggestion that plants have developed mechanisms of global gene regulation related to their unique developmental pathways and environmental responses (88).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

REFERENCES

1. Kadonaga,J.T. (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, **92**, 307–313.
2. Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.

3. Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

4. Grunstein,M. (1997) Histone acetylation in chromatin structure and transcription. *Nature*, **389**, 349–352.

5. Ng,H.H. and Bird,A. (2000) Histone deacetylases: silencers for hire. *Trends Biochem. Sci.*, **25**, 121–126.

6. Struhl,K., Kadosh,D., Keaveney,M., Kuras,L. and Moqtaderi,Z. (1998) Activation and repression mechanisms in yeast. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 413–421.

7. Allfrey,V.G., Faulkner,R. and Mirsky,A.E. (1964) Acetylation and methylation of histones and their possible role in regulation of RNA synthesis. *Proc. Natl Acad. Sci. USA*, **51**, 786.

8. Hebbes,T.R., Thorne,A.W. and Crane-Robinson,C. (1988) A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.*, **7**, 1395–1402.

9. Kayne,P.S., Kim,U.J., Han,M., Mullen,J.R., Yoshizaki,F. and Grunstein,M. (1988) Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell*, **55**, 27–39.

10. Thompson,J.S., Ling,X. and Grunstein,M. (1994) Histone H3 amino terminus is required for telomeric and silent mating locus repression in yeast. *Nature*, **369**, 245–247.

11. Durrin,L.K., Mann,R.K., Kayne,P.S. and Grunstein,M. (1991) Yeast histone H4 N-terminal sequence is required for promoter activation *in vivo*. *Cell*, **65**, 1023–1031.

12. Mann,R.K. and Grunstein,M. (1992) Histone H3 N-terminal mutations allow hyperactivation of the yeast GAL1 gene *in vivo*. *EMBO J.*, **11**, 3297–3306.

13. Grunstein,M. (1992) Histones as regulators of genes. *Sci. Am.*, **267**, 68B–74B.

14. Brownell,J.E., Zhou,J., Ranalli,T., Kobayashi,R., Edmondson,D.G., Roth,S.Y. and Allis,C.D. (1996) *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, **84**, 843–851.

15. Taunton,J., Hassig,C.A. and Schreiber,S.L. (1996) A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science*, **272**, 408–411.

16. Suka,N., Carmen,A.A., Rundlett,S.E. and Grunstein,M. (1998) The regulation of gene activity by histones and the histone deacetylase RPD3. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 391–399.

17. Kadosh,D. and Struhl,K. (1998) Histone deacetylase activity of Rpd3 is important for transcriptional repression *in vivo*. *Genes Dev.*, **12**, 797–805.

18. Kadosh,D. and Struhl,K. (1998) Targeted recruitment of the Sin3–Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin *in vivo*. *Mol. Cell. Biol.*, **18**, 5121–5127.

19. Kadosh,D. and Struhl,K. (1997) Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. *Cell*, **89**, 365–371.

20. Guschin,D., Wade,P.A., Kikyo,N. and Wolffe,A.P. (2000) ATP-dependent histone octamer mobilization and histone deacetylation mediated by the Mi-2 chromatin remodeling complex. *Biochemistry*, **39**, 5238–5245.

21. Fuks,F., Burgers,W.A., Brehm,A., Hughes-Davies,L. and Kouzarides,T. (2000) DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nature Genet.*, **24**, 88–91.

22. Sterner,D.E. and Berger,S.L. (2000) Acetylation of histones and transcription-related factors. *Microbiol. Mol. Biol. Rev.*, **64**, 435–459.

23. Imhof,A., Yang,X.J., Ogryzko,V.V., Nakatani,Y., Wolffe,A.P. and Ge,H. (1997) Acetylation of general transcription factors by histone acetyltransferases. *Curr. Biol.*, **7**, 689–692.

24. Neuwald,A.F. and Landsman,D. (1997) GCN5-related histone *N*-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. *Trends Biochem. Sci.*, **22**, 154–155.

25. Candau,R., Moore,P.A., Wang,L., Barlev,N., Ying,C.Y., Rosen,C.A. and Berger,S.L. (1996) Identification of human proteins functionally conserved with the yeast putative adaptors ADA2 and GCN5. *Mol. Cell. Biol.*, **16**, 593–602.

26. Bannister,A.J. and Kouzarides,T. (1996) The CBP co-activator is a histone acetyltransferase. *Nature*, **384**, 641–643.

27. Giles,R.H., Peters,D.J. and Breuning,M.H. (1998) Conjunction dysfunction: CBP/p300 in human disease. *Trends Genet.*, **14**, 178–183.

28. Mizzen,C.A., Yang,X.J., Kokubo,T., Brownell,J.E., Bannister,A.J., Owen-Hughes,T., Workman,J., Wang,L., Berger,S.L., Kouzarides,T. et al. (1996) The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell*, **87**, 1261–1270.

29. Leo,C. and Chen,J.D. (2000) The SRC family of nuclear receptor coactivators. *Gene*, **245**, 1–11.

30. Xu,L., Glass,C.K. and Rosenfeld,M.G. (1999) Coactivator and corepressor complexes in nuclear receptor function. *Curr. Opin. Genet. Dev.*, **9**, 140–147.

31. Frye,R.A. (2000) Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins. *Biochem. Biophys. Res. Commun.*, **273**, 793–798.

32. Leipe,D.D. and Landsman,D. (1997) Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res.*, **25**, 3693–3697.

33. Imai,S., Armstrong,C.M., Kaeberlein,M. and Guarente,L. (2000) Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature*, **403**, 795–800.

34. Aparicio,O.M., Billington,B.L. and Gottschling,D.E. (1991) Modifiers of position effect are shared between telomeric and silent mating-type loci in *S. cerevisiae*. *Cell*, **66**, 1279–1287.

35. Gottlieb,S. and Esposito,R.E. (1989) A new role for a yeast transcriptional silencer gene, SIR2, in regulation of recombination in ribosomal DNA. *Cell*, **56**, 771–776.

36. Smith,J.S. and Boeke,J.D. (1997) An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes Dev.*, **11**, 241–254.

37. Guarente,L. (2000) Sir2 links chromatin silencing, metabolism and aging. *Genes Dev.*, **14**, 1021–1026.

38. Guarente,L. and Kenyon,C. (2000) Genetic pathways that regulate ageing in model organisms. *Nature*, **408**, 255–262.

39. Brachmann,C.B., Sherman,J.M., Devine,S.E., Cameron,E.E., Pillus,L. and Boeke,J.D. (1995) The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression and chromosome stability. *Genes Dev.*, **9**, 2888–2902.

40. Lusser,A., Brosch,G., Loidl,A., Haas,H. and Loidl,P. (1997) Identification of maize histone deacetylase HD2 as an acidic nucleolar phosphoprotein. *Science*, **277**, 88–91.

41. Wu,K., Tian,L., Malik,K., Brown,D. and Miki,B. (2000) Functional analysis of HD2 histone deacetylase homologues in *Arabidopsis thaliana*. *Plant J.*, **22**, 19–27.

42. Aravind,L., Koonin,E.V., Dangl,M., Lusser,A., Brosch,G., Loidl,A., Haas,H. and Loidl,P. (1998) Second family of histone deacetylases. *Science*, **280**, 1167a.

43. Lusser,A., Kolle,D. and Loidl,P. (2001) Histone acetylation: lessons from the plant kingdom. *Trends Plant Sci.*, **6**, 59–65.

44. Graessle,S., Loidl,P. and Brosch,G. (2001) Histone acetylation: plants and fungi as model systems for the investigation of histone deacetylases. *Cell. Mol. Life Sci.*, **58**, 704–720.

45. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

46. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

47. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.

48. Borodovsky,M. and McIninch,J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems*, **30**, 161–171.

49. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

50. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.

51. Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

52. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

53. Felsenstein,J. (1989) PHYLIP—Phylogeny inference package version (3.2). *Cladistics*, **5**, 164–166.

54. Dayhoff,M.O., Schwartz,R.M. and Orcult,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl 3, 345–352.

55. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos—a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

56. Rundlett,S.E., Carmen,A.A., Kobayashi,R., Bavykin,S., Turner,B.M. and Grunstein,M. (1996) HDA1 and RPD3 are members of distinct yeast histone deacetylase complexes that regulate silencing and transcription. *Proc. Natl Acad. Sci. USA*, **93**, 14503–14508.

57. Grozinger,C.M., Hassig,C.A. and Schreiber,S.L. (1999) Three proteins define a class of human histone deacetylases related to yeast Hda1p. *Proc. Natl Acad. Sci. USA*, **96**, 4868–4873.

58. Gao,L., Cueto,M.A., Asselbergs,F. and Atadja,P. (2002) Cloning and functional characterization of HDAC11, a novel member of the human histone deacetylase family. *J. Biol. Chem.*, **277**, 25748–25755.

59. Lechner,T., Lusser,A., Pipal,A., Brosch,G., Loidl,A., Goralik-Schramel,M., Sendra,R., Wegener,S., Walton,J.D. and Loidl,P. (2000) RPD3-type histone deacetylases in maize embryos. *Biochemistry*, **39**, 1683–1692.

60. Ahringer,J. (2000) NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet.*, **16**, 351–356.

61. Hubbert,C., Guardiola,A., Shao,R., Kawaguchi,Y., Ito,A., Nixon,A., Yoshida,M., Wang,X.F. and Yao,T.P. (2002) HDAC6 is a microtubule-associated deacetylase. *Nature*, **417**, 455–458.

62. Grozinger,C.M. and Schreiber,S.L. (2000) Regulation of histone deacetylase 4 and 5 and transcriptional activity by 14-3-3-dependent cellular localization. *Proc. Natl Acad. Sci. USA*, **97**, 7835–7840.

63. Verdel,A., Curtet,S., Brocard,M.P., Rousseaux,S., Lemercier,C., Yoshida,M. and Khochbin,S. (2000) Active maintenance of mHDA2/mHDAC6 histone-deacetylase in the cytoplasm. *Curr. Biol.*, **10**, 747–749.

64. Vetter,I.R., Nowak,C., Nishimoto,T., Kuhlmann,J. and Wittinghofer,A. (1999) Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature*, **398**, 39–46.

65. Saka,Y., Sutani,T., Yamashita,Y., Saitoh,S., Takeuchi,M., Nakaseko,Y. and Yanagida,M. (1994) Fission yeast cut3 and cut14, members of a ubiquitous protein family, are required for chromosome condensation and segregation in mitosis. *EMBO J.*, **13**, 4938–4952.

66. Dangl,M., Brosch,G., Haas,H., Loidl,P. and Lusser,A. (2001) Comparative analysis of HD2 type histone deacetylases in higher plants. *Planta*, **213**, 280–285.

67. Angus-Hill,M.L., Dutnall,R.N., Tafrov,S.T., Sternglanz,R. and Ramakrishnan,V. (1999) Crystal structure of the histone acetyltransferase Hpa2: a tetrameric member of the Gcn5-related *N*-acetyltransferase superfamily. *J. Mol. Biol.*, **294**, 1311–1325.

68. Stockinger,E.J., Mao,Y., Regier,M.K., Triezenberg,S.J. and Thomashow,M.F. (2001) Transcriptional adaptor and histone acetyltransferase proteins in *Arabidopsis* and their interactions with CBF1, a transcriptional activator involved in cold-regulated gene expression. *Nucleic Acids Res.*, **29**, 1524–1533.

69. Ogryzko,V.V. (2001) Mammalian histone acetyltransferases and their complexes. *Cell. Mol. Life Sci.*, **58**, 683–692.

70. Bordoli,L., Netsch,M., Luthi,U., Lutz,W. and Eckner,R. (2001) Plant orthologs of p300/CBP: conservation of a core domain in metazoan p300/CBP acetyltransferase-related proteins. *Nucleic Acids Res.*, **29**, 589–597.

71. Dhalluin,C., Carlson,J.E., Zeng,L., He,C., Aggarwal,A.K. and Zhou,M.M. (1999) Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, **399**, 491–496.

72. Parker,D., Ferreri,K., Nakajima,T., LaMorte,V.J., Evans,R., Koerber,S.C., Hoeger,C. and Montminy,M.R. (1996) Phosphorylation of CREB at Ser-133 induces complex formation with CREB-binding protein via a direct mechanism. *Mol. Cell. Biol.*, **16**, 694–703.

73. Radhakrishnan,I., Perez-Alvarado,G.C., Parker,D., Dyson,H.J., Montminy,M.R. and Wright,P.E. (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell*, **91**, 741–752.

74. Ponting,C.P., Blake,D.J., Davies,K.E., Kendrick-Jones,J. and Winder,S.J. (1996) ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins. *Trends Biochem. Sci.*, **21**, 11–13.

75. Yao,T.P., Ku,G., Zhou,N., Scully,R. and Livingston,D.M. (1996) The nuclear hormone receptor coactivator SRC-1 is a specific target of p300. *Proc. Natl Acad. Sci. USA*, **93**, 10626–10631.

76. Kamei,Y., Xu,L., Heinzel,T., Torchia,J., Kurokawa,R., Gloss,B., Lin,S.C., Heyman,R.A., Rose,D.W., Glass,C.K. *et al.* (1996) A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell*, **85**, 403–414.

77. Ruppert,S., Wang,E.H. and Tjian,R. (1993) Cloning and expression of human TAFII250: a TBP-associated factor implicated in cell-cycle regulation. *Nature*, **362**, 175–179.

78. Pham,A.D. and Sauer,F. (2000) Ubiquitin-activating/conjugating activity of TAFII250, a mediator of activation of gene expression in *Drosophila*. *Science*, **289**, 2357–2360.

79. Dyson,M.H., Rose,S. and Mahadevan,L.C. (2001) Acetyllysine-binding and function of bromodomain-containing proteins in chromatin. *Front. Biosci.*, **6**, D853–D865.

80. Jeanmougin,F., Wurtz,J.M., Le Douarin,B., Chambon,P. and Losson,R. (1997) The bromodomain revisited. *Trends Biochem. Sci.*, **22**, 151–153.

81. Jones,M.H., Hamana,N., Nezu,J. and Shimane,M. (2000) A novel family of bromodomain genes. *Genomics*, **63**, 40–45.

82. Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.

83. Matangkasombut,O., Buratowski,R.M., Swilling,N.W. and Buratowski,S. (2000) Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. *Genes Dev.*, **14**, 951–962.

84. Chimura,T., Kuzuhara,T. and Horikoshi,M. (2002) Identification and characterization of CIA/ASF1 as an interactor of bromodomains associated with TFIID. *Proc. Natl Acad. Sci. USA*, **99**, 9334–9339.

85. Venturini,L., You,J., Stadler,M., Galien,R., Lallemand,V., Koken,M.H., Mattei,M.G., Ganser,A., Chambon,P., Losson,R. *et al.* (1999) TIF1gamma, a novel member of the transcriptional intermediary factor 1 family. *Oncogene*, **18**, 1209–1217.

86. Schultz,D.C., Friedman,J.R. and Rauscher,F.J.,3rd (2001) Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. *Genes Dev.*, **15**, 428–443.

87. Bochar,D.A., Savard,J., Wang,W., Lafleur,D.W., Moore,P., Cote,J. and Shiekhattar,R. (2000) A family of chromatin remodeling factors related to Williams syndrome transcription factor. *Proc. Natl Acad. Sci. USA*, **97**, 1038–1043.

88. Meyerowitz,E.M. (2002) Plants compared to animals: the broadest comparative study of development. *Science*, **295**, 1482–1485.

89. Sherman,J.M., Stone,E.M., Freeman-Cook,L.L., Brachmann,C.B., Boeke,J.D. and Pillus,L. (1999) The conserved core of a human SIR2 homologue functions in yeast silencing. *Mol. Biol. Cell*, **10**, 3045–3059.