Phylogeny-based systematization of Arabidopsis proteins with histone H1 globular domain

Maciej Kotlinski¹, Lukasz Knizewski², Anna Muszewska³, Kinga Rutowicz³,⁴, Maciej Lirski³, Anja Schmidt⁵, Célia Baroux⁴*, Krzysztof Ginalska² and Andrzej Jerzmanowski¹,³ *

¹Laboratory of Systems Biology, Faculty of Biology, University of Warsaw, Pawińskiego 5A, 02-106 Warsaw, Poland;
²Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Zwirki i Wigury 93, 02-089 Warsaw, Poland;
³Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawińskiego 5A, 02-106 Warsaw, Poland; e-mail: andyj@ibb.waw.pl
⁴Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland; e-mail: cbaroux@botinst.uzh.ch
⁵Heidelberg University - Centre for Organismal Studies, Im Neuenheimer Feld 345, 69120 Heidelberg, Germany

*Corresponding authors

A.J. and C.B. conceived the project; M.K., C.B. and A.J. designed the study; M.K. performed proteomic analyses; K.R and M.L. performed transcriptomic analyses; L.K., A.M., M.K. and K.G. performed structural and phylogenetic in silico analyses and database screens; A.S. contributed transcriptomic atlas of different tissues; M.K., C.B. and A.J. wrote the paper.

One sentence summary:

We propose a unified nomenclature of an important group of plant chromatin proteins, based on evolutionary relationships of their common nucleosome–recognition element – a linker histone-type globular domain (GH1).

ABSTRACT

H1 (or linker) histones are basic nuclear proteins that possess an evolutionarily–conserved nucleosome-binding globular domain – GH1. They perform critical functions in determining the accessibility of chromatin DNA to trans–acting factors. In most metazoan species so far studied, linker histones are highly heterogeneous, with numerous non–allelic variants co–occurring in the same cells. The phylogenetic relationships among these variants as well as their structural and functional properties have been relatively well established. This contrasts markedly with the rather
limited knowledge concerning the phylogeny and structural and functional roles of an unusually
diverse group of GH1-containing proteins in plants. The dearth of information and lack of a
coherent phylogeny-based nomenclature of these proteins can lead to misunderstandings as to their
identity and possible relationships, thereby hampering plant chromatin research. Based on published
data and our in silico and high-throughput analyses, we propose a systematization and coherent
nomenclature of GH1–containing proteins of Arabidopsis that will be useful for both identification
and structural and functional characterization of homologous proteins from other plant species.

H1s, also known as linker histones, are universal and ubiquitous component of chromatin fibers, in
which they occur at an average frequency of one molecule per nucleosome (Woodcock et al., 2006).
They are small basic proteins with a highly conserved central globular domain (GH1), and two less
conserved and mostly unstructured tail fragments: a short (~20 amino acids) N–terminal domain
(NTD) and a considerably longer (~100 amino acids) and highly positively charged C–terminal
domain (CTD). The GH1 consists of ~80 amino acids and belongs to the ‘winged helix’ family of
DNA–binding proteins. It contains a characteristic mixed α/β fold consisting of three α–helices (I–
III) and two β–strands (S2-S3). The compact bundle comprised of the three helices forms the core
of this domain. The ‘wing’ structure (from which the name of this family of DNA–binding proteins
is derived) lies within the region located C–terminally to helix III and is an extended loop joining
β–strands S2 and S3. GH1 associates with the nucleosome outside the core particle and contacts
DNA via at least two different binding sites (Zhou et al., 1998; Brown et al., 2006; Syed et al.,
2010; Zhou et al., 2013). In addition to GH1, the overall functional properties of H1 are strongly
influenced by the CTD, which binds to internucleosomal linker DNA. The CTD has an intrinsically
disordered structure capable of adopting different conformations depending on the geometry of the
target surfaces, which may be linker DNA or interacting proteins (Hansen et al., 2006). The prime
determinant of this property is the amino acid composition rather than the CTD sequence, with
charge neutralization upon DNA binding by its many lysine residues playing an important role
(Hendzel et al., 2004). According to current models, simultaneous and synergistic binding of both
GH1 and the CTD are prerequisites for correct H1 placement, and determine its role in chromatin
compaction (Stasevich et al., 2010). It is generally agreed that H1, by restricting nucleosome
mobility and impeding the access of trans–acting factors to their target sequences, exerts strong
effects on DNA–dependent activities, such as transcription and replication, and probably also
recombination and repair (Izzo et al., 2008). Recent evidence suggests an even more complex
pattern of H1 functions in the cell, in which its role as a universal architectural protein affecting
chromatin dynamics is complemented by a parallel function as a local and gene–specific regulator
(McBryant et al., 2010). Linker histones are a more divergent group of proteins than core histones.
In animals, numerous non-allelic variants, including cell type- and stage-specific isoforms, have been described (Jerzmanowski, 2004; Sancho et al., 2008). In addition, and similarly to core histones, major animal H1 variants undergo extensive posttranslational modifications of different types (Wisniewski et al., 2007), the importance of most of which is unknown.

Plant H1s exhibit the universal features of the H1 family, including the occurrence of different non-allelic variants and extensive posttranslational modifications see Tables 1 and S1 (Prymakowska-Bosak et al., 1996; Jerzmanowski et al., 2000; Jerzmanowski, 2004; Kotliński et al., 2016). Interest in their functional roles has grown considerably in recent years, since they are frequently found in high-throughput screens aimed at identifying regulators involved in processes related to development, physiology and adaptation to stresses (Wierzbicki and Jerzmanowski, 2005; She et al., 2013; Zemach et al., 2013; Over and Michaels, 2014; Rutowicz et al., 2015), see also Table S2. However, because of the exceptional diversity of plant GH1-containing proteins, a fact not realized by most researchers, the relevant reference information about members of this group available in databases is highly imprecise, lacks coherence and systematization, and is often misleading, particularly for those unfamiliar with the classification of chromatin proteins. For example, as illustrated in Table 1 and Table S1, plant linker histones, like HMGA (High Mobility Group A) and certain other proteins, are described by the general term ‘winged-helix DNA-binding transcription factor’ in several databases. Numerous plant GH1-containing proteins are listed as ‘putative’ or lack any description. Moreover, the annotation of the same proteins is inconsistent between databases.

Here, we summarize currently available information, including both published data and the findings of our in silico and high–throughput analyses, and propose a coherent system of phylogeny and structure–based nomenclature and annotation of H1s and other GH1–containing proteins of Arabidopsis. This system will be useful as a basic reference tool for the identification and characterization of homologous proteins from different plant species. In addition, we highlight some interesting trends in the evolution of chromatin–based regulation that may be specific for plants.

RESULTS AND DISCUSSION

The Arabidopsis genome encodes 15 proteins containing a genuine GH1 domain. A scheme linking GH1–based phylogenetic relationships with protein domain architectures within this group is shown in Fig. 1. Phylogenetic analysis supports an early separation into three sub–groups, which we rename here as 1) H1s, 2) GH1-HMGA-/GH1-HMGA-related, and 3) GH1-Myb-/GH1-Myb-related. The above pattern is generally conserved in angiosperm plants, as shown by a maximum–
The split into typical H1s and GH1-HMGA-/GH1-HMGA-related preceded the separation of the GH1-Myb-/GH1-Myb-related sub–group. Rapid diversification of the latter compared to the H1s, suggests that it was not initially subjected to strong purifying selection, but might have been important for ongoing adaptive evolution of plants. Perhaps this could be the reason that genes encoding Arabidopsis GH1–containing proteins other than H1s show differential expression patterns in different tissues and developmental stages (Fig. S3). Below, we discuss the properties of the three sub–groups in more detail.

H1s

We have argued previously that the formal criteria that define a typical linker histone, i.e. a protein with a GH1 domain flanked by two unstructured and highly basic tails, are fulfilled by the products of only three Arabidopsis genes, designated \textit{H1.1}, \textit{H1.2} and \textit{H1.3} (Wierzbicki and Jerzmanowski, 2005). As shown in Fig. 1, the sub–group of Arabidopsis H1s consists exclusively of this trio of H1s, none of which has any recognizable domain except GH1. Consistent with earlier analyses of phylogenetic relationships among known plant linker histones (Jerzmanowski et al., 2000; Rutowicz et al., 2015), this sub–group contains a representative (H1.3) of a distinct branch of ‘stress inducible’ H1 variants (Ascenzi and Gantt, 1997; Ascenzi and Gantt, 1999; Scippa et al., 2000; Przewloka et al., 2002; Scippa et al., 2004; Jerzmanowski, 2007). Previously, we demonstrated that this branch separated from the main H1 variants roughly 140 MYA, which coincided with the appearance of angiosperm plants on Earth (Rutowicz et al., 2015). There are no orthologues of stress inducible H1 variants in sequenced species representing green algae, bryophytes, lycophytes and conifers (gonysosperms), analyzed in Fig. S1. Importantly, only members of the H1 sub–group...
possess the characteristic regions of strong positive charge in all C–terminal and most N–terminal domains (Fig. S2), beginning immediately adjacent to GH1. It should be noted that the regions of the NTDs of H1.1 and H1.2 most distant from GH1 contain a negatively charged fragment that is targeted by posttranslational modification by phosphorylation, which further increases its negative charge (Kotliński et al., 2016). Thus, among Arabidopsis GH1–containing proteins, the pattern of charge distribution in the N– and C–terminal domains of H1s appears to be as distinctive a feature as the phylogenetic position of their GH1s.

**GH1-HMGA-/GH1-HMGA-related versus putative true Arabidopsis HMGA proteins**

In animals, High Mobility Group A (HMGA) proteins are distinguished by multiple AT–hook DNA–binding motifs: conserved 9–amino acid peptides capable of strong binding to 6 bp or longer AT–rich stretches of DNA, via the minor groove. Except for an acidic carboxy–terminal region, these proteins do not have any other recognized domains. In contrast, proteins currently defined in the literature as plant HMGA members contain a typical GH1 domain in addition to AT–hook motifs. This arrangement is restricted to angiosperm plants (Fig. S1), suggesting a relatively late occurrence of GH1-AT-hook fusion in the evolution of plants. Arabidopsis has three such proteins (GH1-HMGA1-3) which possess 4 to 6 AT-hook motifs. All three were detected in our analysis of the nuclear proteome of an Arabidopsis T87 cell suspension culture (Table S1, http://proteome.arabidopsis.pl). Interestingly, the Arabidopsis GH1-HMGA cluster also includes a protein with no AT–hook domains (AT5G08780.1, named GH1-HMGA-related4 in our proposed nomenclature). We were unable to detect this protein in our T87 nuclear proteome (Table S1), but its transcript was present in an Arabidopsis transcriptome derived by RNA–Seq analysis (Table S1). Its GH1 sequence places GH1-HMGA-related4 distantly from the rest of the Arabidopsis H1-HMGA sub-group. Comparison of the charged amino acid profiles of non–GH1 fragments of Arabidopsis GH1–containing proteins demonstrated that the CTDs of GH1-HMGA1-3 have an island–like distribution of positively and negatively charged residues, with mostly the latter present in fragments directly adjacent to GH1 (Fig. S2). The corresponding profile for GH1-HMGA-related4 is significantly different. Secondary structure predictions suggest a potentially novel domain which lacks sequence similarity to any other protein domain of known or unknown structure/function. Interestingly, similar sequences are present in proteins from other species of the order Brassicales, in which they are also accompanied by GH1. The phylogenetic tree of GH1s from model plant proteomes identifies a distinct cluster comprised of Arabidopsis GH1-HMGA-related4 and similar proteins from other species. Importantly, according to the InterPro database (http://www.ebi.ac.uk/interpro/), some of the proteins from other species belonging to this cluster retained AT-hook motifs.
The fusion of genuine GH1 and multiple AT–hook motifs which occurred in angiosperm plants, can also be found in phylogenetic groups outside plant kingdom, e.g. in numerous fish species, in *Trichoplax adhaerens*, the only extant representative of the phylum Placozoa (a primitive group of multicellular animals), as well as in some yeast, nematode and insect species. The fish and *T. adhaerens* genomes encode very large proteins (up to 2900 aa) in which GH1 and AT–hook motifs co–occur with RING and PHD domains. The other mentioned organisms possess simpler proteins in which GH1 co–exists exclusively with AT–hook motifs. The phylogenetic relationships among these extremely diverse organisms suggest that multiple evolutionary events have resulted in the co–occurrence of GH1 and AT–hook motifs within their proteins.

Surprisingly given the fundamental functions of HMGA proteins in animals, the functional significance of the GH1/multiple AT–hook motif fusion has never been studied, despite its being referred to in all the major literature concerning plant HMG proteins. Notably, in several prokaryotes in which either HMGA–like or histone H1 CTD–like domains are present in important hub proteins regulating critical cellular processes, these two domains were found functionally equivalent and could be interchanged without any phenotypic consequences. Moreover, even chimeras in which the AT–hook domain was substituted by the human histone H1 CTD or full–length human H1, functioned properly in prokaryotic hosts (García-Heras et al., 2009). Thus, Arabidopsis GH1-HMGA proteins may be considered as a highly specialized derivatives of H1 in which the typical CTD of H1 has been replaced by HMGA. To try and verify such a possibility, we re–examined the long held view that Arabidopsis is devoid of canonical HMGA proteins. Using the SMART tool (Schultz et al., 2000; http://smart.embl-heidelberg.de), we identified 48 Arabidopsis proteins containing AT–hook motifs, 23 of which, unlike typical HMGA members, contain only a single AT–hook. Most of the identified proteins, including those of the H1-HMGA sub–group, contain additional domains. Only two proteins, the predicted products of the alternatively spliced *At1g48610* gene, contain 4 AT-hook motifs and no other domain. *At1g48610.1* encodes a relatively small protein (212 amino acids, about 21.6 kDa) with a high isoelectric point (pI = 11.6) – features typical for HMGA. *At1g48610.2* (transcript retains the last intron) encodes a shorter protein with a pI of 11.4. The other putative proteins with the AT-hook motif are significantly larger and their isoelectric point, unlike that of canonical HMGA, is below 10. Interestingly, a protein encoded by *At1g48610* was detected in our analyses of the nuclear proteome of Arabidopsis T87 cells, with a score and peptide number similar to those of core– and linker histones, which indicated a substantial concentration in nuclei (http://proteome.arabidopsis.pl). Moreover, and probably due to its high isoelectric point, it was co–purified during the isolation of Arabidopsis linker histones by extraction with 4.5% PCA and cation exchange chromatography (Kotliński, et al. 2016). In both analyses, the larger version of AT1G48610 had a higher number of peptides and higher score than
the smaller form (100% and 92% of sequence coverage, respectively). Using 4 different proteases (Trypsin, ArgC, Termolysin and Pepsin), we identified 516 peptides unique for AT1G48610.1, i.e. matching the last 29 amino acids of this protein, including peptides spanning the exon–exon junction. However, we detected no peptides unique for the smaller AT1G48610.2 form, i.e. matching the last 14 amino acids that are different in this variant. Similarly, RNA–Seq analysis revealed multiple reads spanning the junction of the last two exons of the gene, but only one low quality read within the intron retained in At1g48610.2. These data indicate that the larger version of the protein (AT1G48610.1) is the main product of this gene. According to the BAR Toronto database (Toufighi et al., 2005), expression of At1g48610 is strongest in the central, rib and peripheral zones of the shoot apical meristem, in pistil tissue primarily consisting of ovaries, and in phloem companion cells at the border of the meristematic and elongation zones of the root. This suggests that AT1G48610, which we believe to be a true Arabidopsis HMGA protein, is important in the differentiation of stem cells, a role highly reminiscent of that played by animal HMGA–type proteins (Ozturk et al., 2014). Interestingly, the At1g48610 locus in chromosome 1 is located next to that encoding the H1-HMGA2 protein.

GH1-Myb-/GH1-Myb-related

This sub–group comprises five proteins with an additional N–terminal Myb domain accompanied by a 17–18–amino acid–long Myb extension–like domain. They seem to be as evolutionary old as H1s, as in addition to angiosperms, they occur in representatives of green algae, bryophytes, lycophytes and gymnosperms (Fig. S1). They are known as ‘single myb histone’ (SMH) or ‘telomere repeat binding’ (TRB) proteins and two of them GH1-Myb-TRB1 and GH1-Myb-TRB2 were shown to bind Arabidopsis telomeric repeats \textit{in vitro} through a Myb domain of the telo-box (telomere motif AAACCCTAA) – type ( Marian et al., 2003; Schrumpfová et al., 2004). The demonstration of \textit{in vivo} interactions of these proteins with Arabidopsis telomerase support a suggestion that they are part of the greater plant telomeric interactome (Procházková Schrumpfová et al., 2014). However, a recent mapping by chromatin immunoprecipitation sequencing (ChIP–seq) of genome–wide distribution of TRB1:GFP revealed its presence in over 7800 genomic loci. The majority of these loci contained telo–box–related motifs located at the transcription start sites (TSS), with additional loci spreading across gene bodies as well as distal promoter regions. Moreover, it was shown by genome–wide expression (RNAseq) analysis that TRB1, by binding at these loci, plays a role of transcriptional regulator which is independent of its role in telomere maintenance (Zhou et al., 2016). Given such wide–spread occurrence, it seems highly probable that at least in some of the detected loci TRB1 through its GH1 domain, competes for nucleosome binding with H1s.
Since GH1-Myb-TRB3 is very similar to GH1-Myb-TRB1 and GH1-Myb-TRB2 – all three locate in the same branch on phylogenetic tree (Fig. S1), it may perform the same function. GH1-Myb-TRB1 was identified in our proteomic analysis of Arabidopsis nuclei, while GH1-Myb-TRB2 and GH1-Myb-TRB3 were detected below the established threshold (Table S1). Transcripts encoding GH1-Myb-TRB1-3 were all present in our RNA–seq data. Two other GH1-Myb proteins GH1-Myb4 and GH1-Myb5 (AT1G17520.1 and AT1G72740.1, respectively) are more distantly related to GH1-Myb-TRB1-3 (Fig. S1). The three other proteins of this sub–group (GH1-Myb-related 6-8) lack the Myb domain, although the transcript of one them (AT1G54260.1) contains a Myb–coding sequence in front of the start codon, suggesting the loss of this domain during evolution. AT1G54260.1 also contains a strongly diverged and truncated GH1 domain at the C–terminal side of its regular GH1 domain. According to secondary structure predictions, the two other proteins lacking the Myb domain (AT1G54230 and AT1G54240) have α–helical regions within their CTDs. Interestingly, all three proteins lacking Myb are encoded by neighboring genes on chromosome 1. The N– and C–terminal domains of all proteins from the GH1-Myb-/GH1-Myb-related sub–group are mostly negatively charged.

A rationale for the proposed new nomenclature of Arabidopsis GH1–containing proteins

At first glance, the evolutionary diversification of H1s into well distinguished and conserved sub–types seems to be less pronounced in angiosperm plants than in animals, particularly vertebrates. The most distinct structural and functional diversification of plant H1s coincided with the appearance of angiosperms (ca. 140 MYA) and resulted in two major sub–types that have been maintained ever since: the main– and stress–inducible H1s. Regarding H1s, the case of Arabidopsis shows that two main variants and a single stress–inducible variant are sufficient to support the basic processes of growth and development in a typical flowering plant. While this does not rule out the functional significance of more subtle variation within these two major sub–types observed in systematically distant families and species, proof of such significance has yet to be provided. The above notwithstanding, the impression of a seemingly limited diversification of H1s during the evolution of plants may be misleading and result from biased classification rules. These rules were adopted from studies on typical animal H1s, and do not take into account the fundamentally different life strategies and vastly different selection pressures shaping major chromatin structural proteins in plants and animals during their long histories of separate evolution. The GH1-HMGA-/GH1-HMGA-related and GH1-Myb-/GH1-Myb-related sub–groups could be the end result of such specific selection pressures in the plant kingdom. The concept that proteins of these two sub–groups represent highly diverged and specialized derivatives of plant H1 that use GH1 as a common motif
for targeting nucleosomes, is supported by the conserved phylogenetic relationships among plant
GH1–containing proteins, a recently demonstrated wide–spread occurrence of GH1-Myb-TRB1 in
chromatin and its likely involvement in transcriptional regulation, as well as by the identification of
a candidate for a true Arabidopsis HMGA protein that does not contain GH1 domain. This concept
is by no means equivalent to suggesting that all plant GH1-containing proteins are *bona fide* ‘H1
variants’, in a sense ascribed to this subcategory in animal studies. Its main purpose is to draw
attention to the fact, that in plants the competition-based removal of H1 from chromatin may be
dependent on more diversified and specialized group of competitors than in animals, suggesting a
novel plant-specific mechanisms of chromatin regulation. We therefore propose a unified
nomenclature for plant GH1–containing proteins built simply on their GH1–based phylogenetic
relationships, as shown in Fig. 1. We further propose to distinguish proteins possessing two
characteristic domains (GH1-HMGA and GH1-Myb) and proteins belonging to the same sub–
groups due to the phylogenetic position of their GH1, but lacking the second characteristic domain
– HMGA or Myb. We name these latter proteins GH1-HMGA-related and GH1-Myb-related,
respectively (they are marked by lighter color in Fig. S1). It is important to remember that proteins
of these two types from other species still retain their AT–hook motifs and Myb domains. Since the
GH1-Myb-TRB1-2 proteins have been experimentally confirmed to bind telomere repeats and were
therefore named Telomere Repeat Binding 1 and 2 (TRB1 and TRB2), we propose to retain this
functional reference in their names (as GH1-Myb-TRB) for the sake of clarity and tradition. The
same applies to GH1-Myb-TRB3, a very similar protein that has previously been described as
TRB3. With regard to GH1-Myb4 and GH1-Myb5, which are also described as TRB-proteins in
many databases, we suggest removing the designation TRB from their names. In Arabidopsis GH1
evolutionary tree, both of these proteins group in a clade separate from that of TRB1-3, suggesting
a greater evolutionary distance. Moreover, and unlike GH1-Myb-TRB1-3, they both contain a Myb
extension-like sequence different than GH1-Myb-TRB1-3, so their binding preferences may be
different. We have also indicated (Table S1 and, in parentheses, in Fig. S1) the former names of
GH1-Myb proteins as Single Myb Histone (SMH) that were used in the discontinued ChromDB and
in Maize genomic databases. Importantly, our inspection in SMART/UniProt of the domain
structure of all proteins included in the tree in Fig. S1, revealed some singularities. In *Medicago
trunca
tula* a GH1-Myb protein has an additional RNA-recognition motif. Another GH1-Myb of this
species has strongly changed GH1 domain. Both *Brassica rapa* and *Oryza sativa*, have a GH1-Myb
protein carrying additional domain, and *Zea mays* contains a GH1-HMGA protein with S/T kinase
domain. Moreover, in *Z. mays* H1 group there is a protein with two AT-hook motifs (indicative that
such fusions are not unusual in plants). While exception proves the rule, it cannot be excluded that
at least some of the above singularities resulted from errors in genomes assemblies or gene models.
We believe that the proposed phylogeny– and structure–supported system of classification, apart from practical convenience, will foster novel approaches in studies on the functional roles of GH1–containing proteins in plants.

ACKNOWLEDGEMENTS

We thank Fred Berger, Franziska Turck, Fredy Barneche, J. Mark Cock, Lars Hennig, Motoaki Seki, Anna Amtmann and Doris Wagner for comments and fruitful discussions. This work was supported by grants from the European Cooperation in Science and Technology (MNiSW 212/N–COST/2008/0) and the Ministry of Science and Higher Education (MNiSW 212/N–COST/2008/0) to A. Jerzmanowski and M. Kotlinski, the Foundation for Polish Science (TEAM) and the National Science Centre (2011/02/A/NZ2/00014, 2014/15/B/NZ1/03357) to K. Ginals, and the National Science Centre grant (2012/07/D/NZ2/04286), the scholarship for outstanding young researchers from Ministry of Science and Higher Educations to A. Muszewska and Swiss National Science Fundation (SNSF grant 31003A_149974/1) to C. Baroux.

FIGURE LEGENDS

Table 1. Accession numbers and descriptions of Arabidopsis canonical linker histones in different databases: TAIR10, UniProt, NCBI and ChromDB (a copy of this discontinued database in the web archive was used).

Fig. 1 Maximum–likelihood phylogenetic tree and domain architecture of Arabidopsis thaliana GH1-containing proteins. Protein sequences were aligned with the local pair iterative algorithm implemented in Mafft (Yamada et al., 2016). Conserved columns from each multiple sequence alignment were selected manually. The phylogenetic analysis was performed with PhyML (Guindon et al., 2005), with the JTT model of amino acid substitutions and three random starting trees. Approximate likelihood ratio test SH-like branch supports above 50% are shown. The tree was rooted using GH1-Myb as an internal sister outgroup for both GH1-HMGA and Histone H1 clades. The tree image was prepared with iTol (Letunic and Bork, 2011). Domain architecture analysis was carried out using the SMART (Letunic et al., 2015) and GeneSilico (Kurowski and Bujnicki, 2003) webserver and Meta-BASIC (Ginals et al., 2004).

SUPPLEMENTAL DATA
Supplementary Table S1. Accession numbers and descriptions of *Arabidopsis thaliana* proteins containing a GH1 domain from different databases (TAIR10, UniProt, NCBI and ChromDB*).

Supplementary Table S2. List of articles referring to the role of plant linker histones.

Supplementary Table S3. List of GH1 containing protein identifiers in selected model plants.

Supplementary Fig. S1 Maximum-likelihood phylogenetic tree of GH1-containing proteins from selected plants.

Supplementary Fig. S2. Moving sum plot of net charge for N- and C-terminal domains of all Arabidopsis GH1-containing proteins.

Supplementary Fig. S3. Relative expression levels of GH1-containing protein coding genes in Arabidopsis, across 74 tissue– or cell–specific microarrays.

Supplemental Methods

Supplementary Table S1. Accession numbers and descriptions of *Arabidopsis thaliana* proteins containing a GH1 domain from different databases (TAIR10, UniProt, NCBI and ChromDB*). This table is supplemented with data concerning the occurrence, rank and scores of proteins identified in the nuclear proteome of Arabidopsis T-87 suspension culture cells (6753 proteins identified in total, www.proteome.arabidopsis.pl) and localization of gene expression according to the BAR Toronto database (Fucile et al., 2011)(http://bar.utoronto.ca/).

*a* a copy of this discontinued database in the web archive was used.

Supplementary Table S2. List of articles referring to the role of plant linker histones.

Supplementary Table S3. List of GH1 containing protein identifiers in selected model plants. GH1 protein identifiers and names listed for 18 complete plant proteomes in a tabular format.

Supplementary Fig. S1 Maximum-likelihood phylogenetic tree of GH1-containing proteins from selected plants.

Supplementary Fig. S2. Moving sum plot of net charge for N- and C-terminal domains of all Arabidopsis GH1–containing proteins. The net charge (y-axis) is summed in a 20–aa sliding window, with the position along the N– and C–terminal domains, with respect to GH1, denoted on the x-axis. For each N- and C-terminus, the percentages of both positively (K, R) and negatively (D, E) charged residues, total charge and theoretical isoelectric point (pI, calculated with http://web.expasy.org/compute_pi) are also shown.
Supplementary Fig. S3. Relative expression levels of GH1-containing protein coding genes in Arabidopsis, across 74 tissue– or cell–specific microarrays (as used in Schmidt et al., 2011). The arrangement of genes and samples is based on Euclidean distance and hierarchical agglomerative clustering. Colors are scaled per row. Red and green ranges correspond to high and low expression levels, respectively. The pictograms indicate the cell and tissue types.

Supplementary methods

Database screen

All proteins from TAIR (http://arabidopsis.org) database and protein records from Arabidopsis thaliana deposited in NCBI (https://www.ncbi.nlm.nih.gov/) and UniProt (http://www.uniprot.org/) databases were searched with use of BLAST (Altschul et al., 1990) for proteins containing GH1 domain. Sequences of GH1 from all 15 Arabidopsis GH1-containing proteins were used as query. All records found are included in Tables 1 and S1. Additionally, full genomic sequence from TAIR repository was translated in six reading frames and searched by position specific BLAST (Altschul et al., 1997). All 15 GH1 sequences from known proteins were used as query. We have not found any new GH1-containing proteins in Arabidopsis.

Domain architecture

Domain architecture analysis was carried out for all A. thaliana proteins containing GH1 domain using Meta-BASIC (Ginalski et al., 2004) as well as SMART (Letunic et al., 2015) and GeneSilico (Kurowski and Bujnicki, 2003) webservers. The regions with no detectable homology to known protein domains, yet with conserved sequence and predicted secondary structures (with PSIPRED (Jones, 1999)), have been additionally denoted as potential new domains.

For proteins previously assigned to GH1-Myb subfamily, yet lacking Myb domain, nucleotide upstream/downstream sequence of coded genes were verified using both manual translations and data from TAIR gene model and exon confidence ranking system (https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_gene_confidence_ranking/DOCUMENTATION_TAIR_Gene_Confidence.pdf). Truncated GH1 domain detected in AT1G54260 was verified in similar manner.

Moving sum plot

Moving sum plot of net charge was generated for both N- and C-terminal regions (with respect to GH1 domain) of all Arabidopsis GH1-containing proteins. The net charge was summed in a 20-aa sliding window along N- and C-terminal regions, starting from GH1 domain. For each
region, the percentages of both positively (K, R) and negatively (D, E) charged residues, total charge and theoretical isoelectric point (pI, calculated with http://web.expasy.org/compute_pi) were also calculated.

**Phylogenetic analyses**
Protein sequences for model plants were collected *via* phmmer (Finn et al., 2011) available from Ensembl Plants website (Kersey et al., 2014). Ensembl database was chosen to ensure data quality, limiting the dataset to well studied organisms with possibly complete proteomes. Such dataset enables observations of specific subfamily expansions (due to consecutive duplications) in some angiosperms from Brassicaceae and Fabaceae. For better taxon sampling, the following representatives of missing major taxon groups were added: *Auxenochlorella protothecoides*, *Coccomyxa subellipsoidea*, *Marchantia polymorpha*, *Picea sitchensis*, *Pinus taeda* (from UNIPROT) and *Klebsormidium flaccidum* (from NCBI genomes).

Sequence searches were performed using H1.2, GH1-HMGA2, GH1-Myb-TRB1 and TRB1 from *Arabidopsis thaliana* as queries. All hits were mapped on Uniprot identifiers (http://www.uniprot.org), except for *Physcomitrella patens* (which lacks Uniprot ids for 2 out of 9 analyzed sequences). Subsequently, representative plants were chosen with emphasis on Brassicaceae (3 taxa) and including all basal plant model organisms present in the aforementioned database (for a list of identifiers and names see Supplementary Table S3). Incomplete truncated sequences were discarded. Phylogenetic trees were inferred both for *A. thaliana* GH1 proteins (see Fig 1) and for 282 representative plant sequences (Fig S1).

Sequences of all GH1-containing proteins used for phylogenetic comparison were screened with SMART (Schultz et al., 2000; Letunic et al., 2015) for presence of any additional domains or loss of domains (other than GH1). The results are included in Fig. S1.
Table 1. Accession numbers and descriptions of Arabidopsis canonical linker histones in different databases: TAIR10, UniProt, NCBI and ChromDB (a copy of this discontinued database in the web archive was used).

<table>
<thead>
<tr>
<th>Gene spl. var.</th>
<th>TAIR10 id</th>
<th>TAIR10 description</th>
<th>UniProt id</th>
<th>UniProt description</th>
<th>NCBI id</th>
<th>NCBI description</th>
<th>Chrom DB</th>
<th>Length [aa]</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1.1</td>
<td>AT1G06760.1</td>
<td>winged-helix DNA-binding transcription factor family protein</td>
<td>P26568</td>
<td>Histone H1.1</td>
<td>NP_172161.1</td>
<td>Histone H1.1; histone H1-1</td>
<td>HON1</td>
<td>274</td>
</tr>
<tr>
<td>H1.2</td>
<td>AT2G30620.1</td>
<td>winged-helix DNA-binding transcription factor family protein</td>
<td>P26569</td>
<td>Histone H1.2</td>
<td>NP_180620.1</td>
<td>Histone H1.2; putative histone H1 protein; histone H1</td>
<td>HON2</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td>AT2G30620.2</td>
<td>winged-helix DNA-binding transcription factor family protein</td>
<td>F4INW2</td>
<td>Histone H1.2</td>
<td>NP_0011896</td>
<td>histone H1.2</td>
<td>HON2</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>C0Z3A1</td>
<td>AT2G30620 protein</td>
<td>BAH57180.1</td>
<td>AT2G30620</td>
<td>208</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>AT2G18050.1</td>
<td>HIS1-3</td>
<td>histone H1-3</td>
<td>P94109</td>
<td>Histone H1</td>
<td>His1-3</td>
<td>histone H1-3; histone H1</td>
<td>HON3</td>
</tr>
<tr>
<td>2</td>
<td>AT2G18050.2</td>
<td>HIS1-3</td>
<td>histone H1-3</td>
<td>Q3EBY3</td>
<td>Histone H1-3</td>
<td>HIS1-3</td>
<td>histone H1-3</td>
<td>138</td>
</tr>
</tbody>
</table>

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ascenzi R, Gantt JS (1997) A drought-stress-inducible histone gene in Arabidopsis thaliana is a member of a distinct class of plant linker histone variants. Plant Mol Biol 34: 629-641

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Over RS, Michaels SD (2014) Open and closed: the roles of linker histones in plants and animals. Mol Plant 7: 481-491

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title


