Unraveling the evolutionary origins of Histone fold motif (HFM) in plants

Amish Kumar
National Institute of Plant Genome Research (NIPGR)
Gitanjali Yadav (gy@nipgr.ac.in)
University of Cambridge https://orcid.org/0000-0001-6591-9964

Research article

Keywords: Histone Fold Motif, Plant genome, HMM, core histones, evolution, Nuclear Factor-Y, TAF, DR1

Posted Date: September 8th, 2019

DOI: https://doi.org/10.21203/rs.2.14114/v1

Abstract

Background The three helical Histone Fold Motif (HFM) of core histone proteins provides an evolutionarily favoured site for the protein-DNA interface. Despite significant variation in sequence, the HFM retains a distinctive structural fold that has diversified into several non-histone protein families in. In this work we explore the ancestry of non-histone HFM containing families in the plant kingdom.

Results A sequence search algorithm was developed using iterative profile Hidden Markov Models to identify remote homologs of core-histone proteins. The resulting hits were functionally annotated, classified into families, and subjected to comprehensive phylogenetic analyses via Maximum likelihood and Bayesian methods. We have identified over 4000 HFM containing proteins in the plant kingdom that are not histones, mostly existing as diverse transcription factor families, distributed widely within and across taxonomic groups.

Conclusion Patterns of homology suggest that core histone subunit H2A has evolved into newer families like NF-YC and DrAp1, whereas the H2B subunit of core histones shares a common ancestry with NF-YB and Dr1 class of TFs. Core histone subunits H3 and H4 were found to have evolved into DPE and TAF proteins, respectively. Taken together these results provide insights into diversification events during the evolution of the histone fold motif, including sub-functionalization and neo-functionalization of the HFM.

Background

Histones are one of the most evolutionary conserved and ubiquitous proteins among all eukaryotes, fundamental to formation of DNA compaction units called ‘nucleosomes’, which, in turn are known to organize genomic DNA into chromatin [1]. Each nucleosome comprises of a 146bp long DNA molecule, along with two copies each, of four core histone subunits namely; H2A, H2B, H3 and H4, forming an octameric protein and DNA complex. Despite differences in primary sequence, all four core histone subunits share a distinctive secondary structural fold with three helices separated by two loops/sheets. This “histone fold” enables pairwise dimerization of histone subunits (H2A with H2B; H3 with H4) in handshake fashion as shown in Figure 1B and 1C, followed by dimerization of dimers into tetramers $(H2A\text{-}H2B)_2$ $(H3\text{-}H4)_2$ (Figure 1A) and finally, the Histone octamer [2]. These series of paired dimerization are guided precisely by the *h*istone *f*old *m*otif (HFM), which also provides a binding site for DNA. The presence of the HFM in all four classes of histones suggests that evolution has favored considerable variation in primary structure but only to the extent that secondary structure remains conserved [3]. The high conservation in HFM across eukaryotes (Figure 1D), combined with existence of histone like proteins in archaebacteria, strongly support a common “protohistone” ancestor [4].

As shown in Figure 1A, most nucleosomal HFM residues stay within close interacting range of partner histone subunits or the DNA nucleosomes, resulting in a compact complex, and this compaction, in turn, has led to slow evolution of the HFM regions [3]. The presence of HFM in almost all lineages of the Euryarchaeota indicates the origin of HFM back to the archaeal DNA binding proteins HMf, HMt, and

HMv. Most of these archeal proteins (*Methanobacterium clade)* have one histone fold which is very similar to the eukaryotic histone fold but some archaea (*Methanopyrus kandleri* and *Halobacterium* species) are known to have twice the size, or a doublet histone fold, possibly to provide an opportunity for the eukaryotic HFM to evolve its multi-dimeric design [4].

Apart from core histones and archaeal histone-like proteins, the HFM has been detected in a large number of eukaryotic non-histone proteins, often transcription factors [1]. For example, the general transcription factor TFIID subunits, also known as *T*BP-*A*ssociated *f*actors (TAFs) of *Drosophilla* and humans have significant primary as well as secondary structural overlap with the HFM of core histones [5, 6]. TAFs are particularly well studied in *Drosophila,* humans, and yeast. *In-vitro* studies have revealed TAFs to be co-activators for transcriptional activators, playing important roles in transcriptional initiation and regulation [7]. Another TF family well known to have an HFM, is the Nuclear Factor Y (NF-Y) or CBF. It is a CCAAT-binding heterotrimeric protein complex known to activate a large number of eukaryotic promoters [8]. NF-Y has three subunits, namely the NF-YA, NF-YB and NF-YC, respectively, all of which are necessary for DNA binding. Of these three sub-units, the NF-YB and NF-YC form a heterodimer very similar to the histone H2A-H2B complex [9].

In the plant kingdom, only the NF-Y class of functional non-histone HFM proteins has been characterized (NF-Ys), whereas other known classes like TAFs, Dr1/DrAp1, H3-like Centromeric proteins have been investigated inadequately, and still other families like the Chromatin accessibility complex, DNA polymerase epsilon subunits, and Centromeric protein S are very poorly known [10]. In order to overcome this gap in knowledge, we have performed a large scale identification of all possible non-histone HFM containing proteins across 52 plant species, using sequence and structural homology, followed by functional annotation, and assessment of shared ancestry with core histone sub-units. Over the years, there have been several discrepancies in assigning ancestry to these proteins based on HFM diversification, mainly because the HFM homology across currently known non-histone proteins varies drastically across functional classes, especially in terms of distance from core histone subunits. In this work, we have attempted to resolve this issue for plant proteins through a comprehensive phylogenetic analysis within and across various functional non-histone HFM classes.

Results

Identification of HFM in plants and annotation

A total of 663 HFM sequences (Supplementary Table S2) were obtained after expanding the plant origin seed set via three iterations. These were then used to identify novel HFM sequences on the complete plant proteome representing all major taxa of green plants as described in Methods (Figure 2B). After filtering redundant and non-significant hits, a total of 8264 unique HFM containing sequences were obtained. Functional annotation of these proteins was performed to remove true histone homologs and to retain only the non-histone HFM sequences for downstream analyses. Out of the 8264 HFM homologs, 4376 sequences were identified as core histones and the remaining 3888 hits were treated as non-histone

proteins. These non-histone HFM proteins (Table 1) were assigned to four parental histone classes as described in Methods. Of these, the H2B-HFM class was found to have highest number of non-histone protein homologs (1634), followed by H2A-HFM (1182) and then H4-HFM (911). The least number of non-histone HFM proteins (160) were found to match with H3-HFM as shown in Table 1.

Functional annotation

Functional annotation of the newly identified 3888 non-histone HFM proteins in the plant kingdom revealed their presence in various protein subfamilies, many of which are transcription factors or activators or co-regulators. In all, we find *N*uclear *F*actor *Y* Subunits B (NF-YB) & C (NF-YC), *D*own *r*egulator *1* (Dr1) and *D*own *r*egulator *a*ssociated *p*rotein *1* (DrAp1) also known as Negative Co-regulator α and β respectively (NCα/NCβ), Chromatin accessibility complex (CHRAC), DNA polymerase epsilon subunit 3 (DPE–3) and DNA polymerase epsilon subunit C (DPE-C), Centromeric histone 3 (CENH3), Centromeric proteins S (CEN-S), Bromodomain transcription factor and different subunits of *T*BP *a*ssociated *f*actors (TAFs). More than half (52%) of the identified non-histone HFM proteins in plants belong to the large multi gene family of NF-Ys, followed by several classes of TAFs (23%) and then by Dr1/DrAp1 (10%) class, as described below.

Assignment of Parental Histone classes

All identified non-histone HFM containing proteins were assigned to four parental groups of core histones, namely H2A, H2B, H3 and H4, based on e-value of the Histone Fold Motif (HFM) in each HMM profile. The non-histone HFM containing proteins in the Nuclear Factor Y transcription factor subunits were found to resemble most closely to H2A and H2B classes of core histones. Of these, the majority of NF-YB could be assigned to H2B class of histones, while the other subunit NF-YC was assigned as an H2A homolog. Since these NF-Ys belong to a very large gene family in plants, our analysis also reveals these two subunits of NF-Ys to be represented in very large number. In all, we found 839 NF-YC proteins and assigned these to H2A class of HFM, while 1194 NF-YBs were assigned to H2B class of HFM. Another major group of non-histone proteins classified to H2B and H2A class of HFM, namely the *D*own *r*egulator *1* (Dr1) and *D*own *R*egulator *a*ssociated *p*rotein *1* (DrAp1) respectively. We assigned 226 DrAp1 to H2A class of HFM, and 150 Dr1 to H2B (Table 1). It is interesting to note that both NF-YB/NF-YC and Dr1/DrAP1 constitute cognate pairing partners [8, 11], just as in case of H2A/H2B, and pairwise protein-protein interactions are maintained between NF-YB and C, as well as between Dr1and DrAP1 classes of non-histone proteins, mediated via the conserved HFM. We are currently trying to match the 3D pairing interface of non-histone partners to the well characterized core histone residues involved in binding at the H2A/H2B interaction interface, to understand the extent to which binding patterns have been preserved during evolution.

Members of CHRAC and DPE-C functional classes of non-histone HFM proteins were found to be homologous to H2A-HFM, with 23 CHRAC and 71 DPE-C members being identified in this study, as shown in Table 1. Other non-histone proteins with homology to H2A-HFM belong to DNA repair proteins, Calmodulin binding proteins, Neurofilament heavy polypeptide like proteins. These are in very small numbers and thus assigned as 'others' in Table 1. In contrast to the previous pairwise interactors, we did not find any domains that may be interacting counterparts of DPE or CHRACs. Another subunit of DNA polymerase epsilon (DPE–3) was indeed found among the non-histones, but all 87 members were assigned to the H4 subunit.

For the H2B class, apart from NF-YB and Dr1 described above, we assigned a majority of non-histone proteins that are TBP associated Factor subunit 12 (TAF12). TAF12 is one of the subunits of multi-subunit general transcription factor TFIID[5]. Members of TAF12 were found in significant number (237), amounting to almost 10% of the total H2B homologs. Surprisingly, even though TAF subunits are often known to dimerize with specific pairing preferences, we could not identify any potential TAF subunits that could be assigned to the H2A-HFM, as partners of TAF12.

The lowest number of non-histone proteins were assigned to the H3-HFM parental class, and these included Centromeric proteins, 60s Ribosomal protein, and few unknown and poorly annotated proteins assigned to 'other' classes Table 1.

A total of 911 non-histone HFM containing proteins were assigned to the ancestral H4-HFM class. These non-histone HFM proteins belong to different protein families. Majority of these belong to different subunits of TAFs; namely TAF10, TAF9, TAF8, TAF6 and TAF11. Each of these TAFs were found in significant numbers as shown in Table 1, except for TAF11 class for which only 16 members were identified. TAFs are not well studied in plant system but there is significant literature on this family in Yeast, Drosophilaand Humans. TAFs have also been reported in TBP less acetyl-transferase complex and SAGA complexes [12]. In vitro, TAFs are known to form dimers similar to the histone heterodimers, and five such dimers have been identified between TAFs, namely TAF3–10, TAF6–9, TAF4–12, TAF8–10 and TAF11–13 [12–14]. However, we could not find significant HFM homology in any member of TAF13 and TAF4 in the plant kingdom. Furthermore, we could only assign the H4 ancestral core histone class to most TAFs, with the exception of TAF12 which was found to be homologous to H2B-HFM as mentioned above. Two other proteins families, Centromere protein S and DNA polymerase epsilon subunit 3 were also found in significant numbers (39 and 87 respectively), in H4-HFM parental group.

Phylogenetic analysis

Phylogenetic analysis was performed for three sets as discussed in methods. Set I HFM sequences of NF-YB/NF-YC, Dr1/DrAp1, H2A and H2B were analyzed together by Bayesian analysis and distance based phylogeny and the results are depicted in Figure 3A. The phylogeny suggests HFM of NF-YB share very close homology to the Dr1 and NF-YC share close homology to DrAp1. Both core histones make a distinct

group. The non-histones in this set (Set I) share common ancestral nodes with H2A and H2B and may have evolved parallel to the H2A and H2B lineage.

Set II HFM sequences belonged to all the identified class of TAFs, namely TAF6, TAF8, TAF9, TAF10 and TAF12, analyzed together with core histones H2B, while remaining TAF class were found to group to with H4 as shown in figure 3B. TAF12 was found to share ancestry with H2B while TAF6, TAF8, TAF10 and TAF9 share ancestral groups with H4 as shown in Figure 3B. TAF6, TAF8, TAF10 and H4 show single point branching or polytomy that suggest these TAFs and H4 evolved from a common ancestor and that TAFs may not have evolved from H4 at all, despite very close homology of HFM among TAF6, TAF8, TAF9 and TAF10. In contrast, TAF12 shares HFM homology with H2B (Figure. 3B).

The phylogeny of Set III containing DPE–3, DPE-C, H2B and H4 HFM sequences, is shown in Figure 3C, and reveals that DPE–3 and DPE-C share the ancestral group with H4 and H2A respectively.

In summary, the dendrograms corroborate our findings from the profile hidden markov models. In order to further confirm ancestral core-histone assignment and evolutionary associations, we performed a detailed physicochemical analysis of each non-histone domain with the four core-histone classes as discussed below.

Physicochemical properties & DNA binding patterns

The physico-chemical properties of the HFM regions of newly identified non-histone proteins in the plant kingdom were measured to find the extent of similarity with respective core histone counterparts, as described in Methods. The HFM with in core histones is basic and positively charged, and this feature facilitates binding to DNA[15]. The range of charges found on non-histone HFM sequences are shown in Figure 4A, and these values suggest the extent of their ability to interact with DNA. In general, HFMs with the strongest positive charge belong to CENH–3, DRAP1, DPE-C and CHRAC, suggesting high potential for binding to DNA. In contrast, charges on DPE–3 and Dr1 are highly negative, while most of the TAF subunits range from neutral to negative (except TAF10 which showed neutral to positive) charge on the HFM. Charges on the NF-YB and NF-YC are wide ranging, with median close to neutral (Figure 4A). These features indicate an affinity for binding other biomolecules with complex charge patterns, a premise supported by the fact that TAFs and NF-Ys are both components of large multi-subunit protein complexes. In order to validate some of these findings, we checked residue level compositions. Figure 4B shows boxplots to compare core histones with HFM regions of non-histone proteins. Core-histone HFMs have more than 20% basic and less than 10% acidic amino acids, a tendency that matches DPE-C, DRAP1 and TAF10 among the non-histone HFM regions, further validating our observations in the residue level charge patterns. Detailed class-wise amino acid content in the non-histone HFMs is shown in Supplementary Figure S1.

Residue conservation at the binding interface

3D structural models were constructed for selected non-histone HFM domains using the crystal structure templates of *A.thaliana* NF-Ys, namely AtNF-YB6 (AT5G47670) and AtNF-YC3 (AT1G54830) with PDB-ID 5G49 [16]. Comparison of non-histome HFMs with the crystal structure reveals greater variability at exposed surface regions of the dimer, and high conservation of residues at the protein-protein interface as shown in Figure 5. Positively charged residues near the surface of HFM dimer (site for DNA binding in core histone HFM) are much more variable than core histone sequences, or the protein-protein interface which occurs in the long middle helix. This pattern is not limited only to NF-Ys, as shown in Figure 5; other non-histone HFM proteins, DR1 and DRAP1 also have positively charged residues at either end of the HFM tertiary fold structure. The conservation heat map shows greater variability at the C-terminal of the HFM, while residues making dimer contact are conserved. The HFM models of TAF proteins also show conservation across the tertiary fold of HFM, with positively charged residues at either end of the HFM. The exceptions to this pattern are TAF6 and TAF9 where arginine was found in the long middle helix, the potential dimer forming interface, reminiscent of the H4 structure which also has two arginine and one lysine at the long middle helix (Figure 5).

Discussion

In this work, we have identified and assessed the Histone Fold Motif (HFM) of non-histone proteins in the plant kingdom. Approximately 4000 such proteins were identified and homology was found to be widely variable among different protein families. Some non-histone proteins share very close homology with core histones while other are far more diversified. The NF-YB and NF-YC are among the closest homologs of core histones and we were able to easily identify and classify these into H2A type (NF-YC) and H2B type (NF-YB). Other class of proteins like DNA polymerase epsilon subunit 3, DNA polymerase epsilon subunit C, and the TAFs are distant relatives and could not be identified by conventional HMM based homology search. HHblits based search made it possible to identify most of these remotely homologous non-histone HFM containing protein families. Overall, we could assign the non-histone classes into HFMs with mutual pairing preferences, and those without any potential pairing partners, as described below:

Paired interactors NF-YB/C and Dr1/DrAp1

More than 50% of all newly identified non-histone proteins belong to transcription factor families annotated as NF-Y Subunit B and subunit C. This was also expected since the NF-Y transcription factors are a huge multigene family well known to be expanded among plants; for review [17]. NF-Ys are often confused with other HFM containing protein families, most particularly with the Down regulator1 and Down regulator associated protein 1 (Dr1/DrAp1), which are sometimes wrongly annotated as NF-YB and NF-YC respectively [18]. The Dr1/DrAp1 families are closely homologous to the H2B and H2A HFM respectively, just as they are to the NF-Ys, despite being a distinct class of proteins [19]. We performed a detailed phylogenetic study and found both classes to form separate monophyletic groups during the phylogenetic analysis (Figure 3). Apart from Dr1/DrAp1 other classes like DPE–3 and DPE-C are also

misunderstood with NF-YB and NF-YC respectively [18], and our work has been able to resolve this misperception.

Previously available annotations have assigned 26 proteins annotated as NF-Ys in *A.thaliana* (13 NF-YB and 13 NF-YC) [20] and there was no annotation for any Dr1, DrAp1, DPE–3 and DPE-C. We have corrected the annotation of all 26 Arabidopsis non-histone proteins, revealing two Dr1 (AT5G08190 and AT5G23090) and one DPE–3 (AT2G27470) which were earlier wrongly annotated as NF-YB. We have also found two DPE-C (AT5G43250, AT1G07980) and one DrAp1 (AT3G12480) in Arabidopsis, which were previously annotated as NF-YC. Based on the comprehensive phylogenetic analysis in this work, we have been able to identify 20 NF-Ys in Arabidopsis; ten NF-YB and ten NF-YC, listed in Supplementary Table S3.

UnPaired HFMs with ancestry to H2A and H2B

Other than NF-YB/NF-YC and Dr1/DrAp1, other classes of non-histone proteins were also found to have HFM regions homologous to H2A-HFM. These include Chromatin accessibility complex (CHRAC) and DNA polymerase epsilon C (DPE-C) but both of these classes lack interacting counterparts in H2B-HFM homologs. It is quite possible that these proteins have evolved to function without any need for dimerization. Accordingly, these are found to be part of the chromatin assembly/remodeling and DNA replication machinery. In humans there are two proteins CHRAC15 and CHRAC17, reported to form complexes with ACF1-ISWI complex (ATP-dependent chromatin assembly and remodeling factor [ACF]) and thereby aid the ATP-dependent nucleosome sliding on the DNA [21, 22].

Ancestry of plant TAFs

We found the phylogeny of plant TAF subunits to be quite interesting. Most of these proteins were found to be homologous to H4-HFM ancestors, with the only exception being TAF12, that was assigned to H2B-HFM ancestral class. Surprisingly, no TAF subunits showed significant homology with H3 or H2A core histone classes. In Drosophila and humans, the paired interactors TAF6 and TAF9 were found to have HFMs that resembled H4 and H3 respectively, and both were found to dimerize like H3-H4 heterodimer [5, 6, 14]. Our data does not provide any evidence for such an interaction. It was initially postulated that TAF12 may form homodimers, in the absence of any known H2A homolog to interact with the H2B-HFM of TAF12 [6]. However, this view was rejected later when H2A homolog TAF4 was identified, lacking the α3 helix which surprisingly did not hinder the interaction with TAF12, resulting into a TAF4/TAF12 heterodimer in yeast [12, 14]. Based on these heterodimer structures, TAFs were shown to form histone like octameric structure with two copy of each TAF4/TAF12 and TAF6/TAF9. However, the quaternary arrangement of TAF octamer is very different form the histone octamer [7]. In the present study, we did not identify any such pairwise relationships within the plant kingdom. The non-histone HFM containing proteins mapping to H3 and H2A ancestral classes were found devoid of any TAF homologs. In order to confirm and support our findings, we checked the ancestry of human TAF6/TAF9 an d TAF4/TAF12 and

found these to map to human core histones H4 and H3 respectively, suggesting that plant non-histone sequences have indeed diverged much more than metazoan lineages.

The alternative is that present phylogenetic data may not be strong enough to assign classes appropriately, at least in case of the TAFs. Figure 3 shows H2B and TAF12 forming a common ancestral group while H4 and other TAFs are forming a different monophyletic group with polytomy, suggesting equal homology among H4, TAF6, H3, TAF10, TFA9 and TAF8. The tree also depicts the uncertainty of a common ancestor to these TAFs and H4. Neighbor Joining tree had resolved the polytomy but branch supports were too low to be considered (data not shown). This polytomy further implies that evolutionary information for clear diversification for TAFs from histones are not sufficient in current set of plant HFM sequences. It likely is that TAFs have probably evolved from the common ancestor parallel to the histones in such way that they share similarity to the histone HFM structure but clear similarity to any one histone is very diluted. This finding provides an incentive to explore plant TAF family in greater detail.

Resolving Nomenclature of non-histone HFM families in plants

In this work, DNA polymerase epsilon subunit 3 (DPE–3) and subunit C (DPE-C) were mapped to the ancestral histone classes of H4 and H2A respectively. Phylogenetic analysis also supports this classification with good Bayesian probability and bootstrap support of the branches. It may be noted that in the past, both DPE–3 and DPE-C have been mis-annotated as NF-Ys [18] and the present work resolves the distinct identities. Furthermore, this work has also resolved ancestry of Centromeric histone H3 and Centromeric protein S, homologous to H3 and H4 core histones respectively. Centromeric protein S (CEN-S) is a homolog of human MHF1 (FANC*M* –associated *H*istone-*F*old protein 1) protein also known to form histone H3-H4 like heterodimer with MHF2 in yeast [23]. In Arabidopsis AtMHF1 and AtMHF2 are known to limit the crossover formation at meiosis [24] but in our work, only AtMHF1 (AT5G50930) was found to have HFM which belongs to Centromeric protein S subfamily. The AtMHF2 does not appear to have HFM and there is no report for their interaction in plants.

DNA binding potential of non-histone HFM proteins

3D models of selected HFMs of non-histones reveal these to form similar heterodimeric handshake interfaces like their core histone HFM protein complexes. However, the overall charge propensities on the HFM of non-histones reveal diversions from the core histone HFM, ranging from positive (CHRAC, DPE-C, DRAP1, CEN-H3, TAF10) to near neutral (NF-YB, NF-YC) and negative (DPE–3, Dr1, TAF6, TAF8, TAF9 and TAF12) as shown in Figure 4. The conservation heat map of the amino acid residues on the tertiary structure of the non-histone HFMs shows that the low conservation residues are more likely in positively charged residues (blue coloured sphere in Figure 5) are located at the terminal topography of the HFM, the region known to be involved in protein-DNA interaction, whereas highly conserved residues form the

long middle helix, involved primarily in dimer formation, a region that lacks positive charge. This suggests that the HFMs in non-histones forming canonical handshake dimer have compromised the DNA binding ability to varying extant across protein families to perform their present day role, while the dimer forming abilities of these HFM was not compromised in any of the non-histone protein families.

This conservation heat map on the protein family of CEN-S and TAF9 have less conserved amino acid residues, suggesting that both of these non-histone domains have hugely diverged in functional roles, that require greater sequence variability. In contrast, the DrAP1 set of non-histones having a greater measure of positive charge at the DNA-protein interface, appear to have retained the ability to interact with DNA. This premise is also supported by existing literature reports in rice where OsDrAp1 showed DNA binding, while OsDr1 could not do so on its own [11]. This is supported by our findings as the charge on the DRAP1 is positive and Dr1 HFM is negative, as shown in Figure 4.

Furthermore, we also found reduced proportion of basic amino acids on HFM of NF-YB and NF-YC as compared to their core histone counterparts (Table 2), suggesting their inability to bind to DNA on their own. This finding explains and supports available reports wherein, binding of NF-YB/NF-YC complex to DNA has been shown to be mediated by a third subunit NF-YA, as revealed by the 3-D crystal structure of NF-Ys [16, 25]. In both NF-YB/NF-YC and DR1/DRAP1 complex, there is a third partner NF-YA and TBP respectively to provide sequence specific DNA binding ability, our data also suggest that DNA binding ability of HFM is compromised in non-histone proteins in sequence specific manner in order to utilize this motif in transcriptional regulation and other sequence specific DNA metabolism. At the same time sequence non-specific binding of these HFMs are still being utilized in order to stabilize the DNA complex formation, further supporting the notion of a controlled loss of DNA binding ability of the HFMs in non-histone proteins which varies among different families (Table 4).

We find that TAFs have retained the three dimensional HFM structural fold and the physico-chemical properties required for heterodimerization, as well as the ability to form larger complexes as shown in Figure 4. However, the ability of most plant TAFs (with the exception of TAF10) to interact with DNA appears to be severely compromised in terms of charge propensities and residue compositions. This inference is further supported by reports showing that TAFs are multi-subunit complexes, dependent on TAF10 and other DNA binding proteins for their recruitment on DNA in general transcription machinery or in DNA methylation SAGA complexes.

Conclusion

The Histone Fold Motif (HFM) is a conserved 70–90 amino acid sequence that has been variously identified in a large number of various non-histone proteins in several organisms. In this study, we perform a comprehensive evolutionary analysis of non-histone HFM containing proteins in the plant kingdom. We find that the HFM in non-histones has evolved parallel to core histones and possibly from a common ancestor, enabling us to assign ancestry to most of the non-histone HFM containing plant proteins. With the help of detailed phylogenetics analyses, we have resolved the distinct identities of

several non-histone HFM proteins such as Dr1, DRAP1, DPE–3 and DPE-C. The HFM of non-histone proteins has diversified to perform new functions involving protein-DNA interactions to variable extent, depending upon the extent of conservation of specific residues at the protein-protein and protein-DNA interface. Comparative data across nine families of non-histone proteins suggests that evolution of non-histone HFM classes has generally dis-favored binding to DNA, with only few selected classes having retained DNA binding ability in terms of residue potentials. Overall, the data imply that that protein-protein interaction was favored over the protein DNA interaction during the evolution of HFM. These insights have elucidated molecular interactions at the HFM protein-protein interface and we hope that our findings will pave the way for accurate predictions of cognate partners of HFM based interactions among hugely divergent non-histone protein families.

Methods

Data collection

Plant core histone protein sequences were obtained by taxonomic filtering of the UNIPROT database (http://www.uniprot.org/). This was used as training dataset for the HFM search algorithm. Complete plant proteome data was downloaded from Joint Genome Institute website Phytozome 11 (https://phytozome.jgi.doe.gov/), comprising data for 51 full plant proteome sequences as well as 13 early-release plant proteomes. A taxon based list of these 63 plant species is provided in Supplementary Table S1, and the HFM search algorithm was tested on these sequences for identification of HFM containing proteins, as described in the next section.

Identification of HFM

The HFM of core histone subunits is tightly conserved (>95% identity) among eukaryotes. However, the homology is known to drop steeply for Non histone proteins. Accordingly, simple sequence based homology searches using the core histone dataset were not sufficient to identify non-histone HFM proteins. It was thus decided to add diversity into our original conserved histone sequence alignments by using profile-profile search method, since this is known to be the most sensitive class of sequence-search methods [26]. For this we used the HHblits tool as shown in Figure 2A (https://toolkit.tuebingen.mpg.de/hhblits) [27]. Core histone sequences of the training data set were aligned with clustal omega [28] and then histone fold motif (HFM) region was marked using espript (ESPript—http://espript.ibcp.fr) [29]. These HFM regions were used as seed for profile-profile HMM search by HHblits over ‘refseq-nr20’ database for three iterations to get more distant HFM containing relatives. After each iteration, the highly conserved seed set for each core histone class was expanded by adding the HFM regions from newly found plant origin sequences from the previous iteration. The expanded seed set was then used to search the complete plant proteome for HFM containing homologs. This was done using *HMMsearch* program of HMMER3.1b (http://hmmer.org/) [30]. This output was filtered to remove redundancy at 0.001 e-value.

Phylogenetic analysis

Phylogenetic analysis was done in three separate sets based on the sequence similarity and overlaps with ancestral core histones. The First set included NF-YB, NF-YC, Dr1, DrAp1, H2A, and H2B. This set was designed to study the relationship of Dr1-DrAp1 to NF-YB-NF-YC including respective parental histones H2A and H2B. The Second set included TAF6, TAF8, TAF9, TAF10, TAF12 along with their parental group histones H4 and H2B. The third set included: DNP-C, DNP–3, H2A and H4.

Representative HFM sequences of each set were taken by removing redundancy within functional classes at 90%. These sequences were aligned using clustal omega online [28] with five guide tree and HMM iterations each. The alignment was trimmed using trimAl tool to remove gaps that are more than 95% of the aligned positions [31] in order to avoid extended loop regions of individual sequences. The alignments were manually adjusted using Jalview2.8 [32]. Phylogenetic analysis was done by Bayesian (MrBayes v3.2) and distance based neighbor joining method. For neighbor joining phylogeny, Phylip 3.6v [33] suite was used, analysis was performed with Jones-Taylor-Thornton matrix with 1000 bootstraps, unrooted tree was calculated by neighbor joining and final consensus tree was drawn by majority extended rule. Bayesian analysis (BA) was conducted using MrBayes v3.2 [34], Markov chain Monte Carlo (MCMC) runs with four chains and fixed Jones model was performed for 100,000 generations unless specified. In Set1, NF-Y and histone HFM sequences were run for 100,00,00 generations. Convergence was verified, and potential scale reduction factor (PSRF) was confirmed to reach 1.0 (PSRF of fully converged analysis is 1.0 so PSRF of any analysis have to be reasonably close to 1.0) as runs converged. Tree was visualized and edited by Figtree [35]. For Set2, H2B, H4 and TAFs were run for 200,000 generations and convergence was verified by PSRF being 1.003. Set3 DNP-C, DNP–3, H2A and H4 sequences were run for 100,000 generations convergence was verified by PSRF being 0.998.

Physicochemical Properties

Physicochemical properties of HFM sequences were analyzed using the R package “Peptides” [36]. Physiological Charge on HFM was calculated using EMBOSS pKa scale and box plots were plotted in R version 3.5.1.Amino acid composition was also calculated and boxplot of percent non-polar, polar, basic and acidic amino acid was generated using R version 3.5.1.

Structural analyses

3D Structural models of Arabidopsis HFM containing non-histone proteins were done using I-TASSER (Iterative Threading ASSEmbly Refinement) web program [37] and ConSurf server for mapping conserved region on the HFM three dimensional structural fold [38].

List Of Abbreviations

HFM: Histone Fold Motif, NF-Y: Nuclear Factor Y, TAF: TBP Associated Factors, DR1: Down Regulator protein 1, DRAP1: Down Regulator Associated Protein 1, DPE–3: DNA Polymerase Epsilon Subunit 3, DPE-C: DNA Polymerase Epsilon Subunit C.

Declarations

Acknowledgements

Authors would like to Acknowledge DBT-BTIS-NET and DBT-SERB for grant. AK thanks CSIR-UGC and National Institute of Plant Genome Research, New Delhi for providing fellowship.

Ethics approval and consent to participate

Not aplicable. This manuscript does not report on or involves use of any animal or human participants, human data or human tissue.

Consent for publication

Not applicable

Availability of data and materials

All data analyzed in this study are included in this published article and its supplementary file.

Competing interests

Auther declares that in present study there is no financial and non-financial competing interest.

Funding

This work was funded by DBT-BTISNET and DST-SERB grant, AK obtained fellowship grant from UGC-CSIR and NIPGR.

Authors’ contributions

GY have concieved the idea and AK have planned and executed the work. MS was paired by AK and corrections rectifications was done by GY.

References

1.Baxevanis AD, Arents G, Moudrianakis EN, Landsman D: *A variety of DNA-binding and multimeric proteins contain the histone fold motif. Nucleic acids research* 1995, *23(*14):2685–2691.

2.Frouws TD, Barth PD, Richmond TJ: *Site-Specific Disulfide Crosslinked Nucleosomes with Enhanced Stability. Journal of molecular biology* 2018, *430(*1):45–57.

3.Arents G, Moudrianakis EN: *The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proceedings of the National Academy of Sciences of the United States of America* 1995, *92(*24):11170–11174.

4.Malik HS, Henikoff S: *Phylogenomics of the nucleosome. Nature Structural Biology* 2003, *10:*882.

5.Hoffmann A, Chiang C-M, Oelgeschläger T, Xie X, Burley SK, Nakatani Y, Roeder RG: *A histone octamer-like structure within TFIID. Nature* 1996, *380:*356.

6.Xie X, Kokubo T, Cohen SL, Mirza UA, Hoffmann A, Chait BT, Roeder RG, Nakatani Y, Burley SK: *Structural similarity between TAFs and the heterotetrameric core of the histone octamer. Nature* 1996, *380:*316.

7.William Selleck RH, Qiaojun Fang, Vladimir Podolny, Michael G.Fried, Stephen Buratowski and Song Tan: *A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. Nature Structural Biology* 2002, *9:*231.

8.Bellorini M, Lee DK, Dantonel JC, Zemzoumi K, Roeder RG, Tora L, Mantovani R: *CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues. Nucleic acids research* 1997, *25(*11):2174–2181.

9.Romier C, Cocchiarella F, Mantovani R, Moras D: *The NF-YB/NF-YC Structure Gives Insight into DNA Binding and Transcription Regulation by CCAAT Factor NF-Y. Journal of Biological Chemistry* 2003, *278(*2):1336–1345.

10.Kumar A, Yadav G: *Diversification of the Histone Fold Motif in Plants: Evolution of New Functional Roles. Defence Life Science Journal* 2016, *1(*1):63–68.

11.Song W, Solimeo H, Rupert RA, Yadav NS, Zhu Q: *Functional Dissection of a Rice Dr1/DrAp1 Transcriptional Repression Complex. The Plant Cell* 2002, *14(*1):181.

12.Gangloff YG, Werten S, Romier C, Carré L, Poch O, Moras D, Davidson I: *The human TFIID components TAF(II)135 and TAF(II)20 and the yeast SAGA components ADA1 and TAF(II)68 heterodimerize to form histone-like pairs. Molecular and cellular biology* 2000, *20(*1):340–351.

13.Trowitzsch S, Viola C, Scheer E, Conic S, Chavant V, Fournier M, Papai G, Ebong I-O, Schaffitzel C, Zou J *et al: Cytoplasmic TAF2–TAF8–TAF10 complex provides evidence for nuclear holo–TFIID assembly from preformed submodules. Nature Communications* 2015, *6:*6011.

14.Werten S, Mitschler A, Romier C, Gangloff Y-G, Thuault S, Davidson I, Moras D: *Crystal Structure of a Subcomplex of Human Transcription Factor TFIID Formed by TATA Binding Protein-associated Factors hTAF4 (hTAFII135) and hTAF12 (hTAFII20). Journal of Biological Chemistry* 2002, *277(*47):45502–45509.

15.Arents G, Moudrianakis EN: *Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. Proceedings of the National Academy of Sciences* 1993, *90(*22):10489–10493.

16.Gnesutta N, Saad D, Chaves-Sanjuan A, Mantovani R, Nardini M: *Crystal Structure of the Arabidopsis thaliana L1L/NF-YC3 Histone-fold Dimer Reveals Specificities of the LEC1 Family of NF-Y Subunits in Plants. Molecular Plant* 2017, *10(*4):645–648.

17.Laloum T, De Mita S, Gamas P, Baudin M, Niebel A: *CCAAT-box binding transcription factors in plants: Y so many? Trends in Plant Science* 2013, *18(*3):157–166.

18.Petroni K, Kumimoto RW, Gnesutta N, Calvenzani V, Fornari M, Tonelli C, Holt BF, Mantovani R: *The Promiscuous Life of Plant NUCLEAR FACTOR Y Transcription Factors. The Plant Cell* 2012, *24(*12):4777.

19.Kim S, Na JG, Hampsey M, Reinberg D: *The Dr1/DRAP1 heterodimer is a global repressor of transcription in vivo. Proceedings of the National Academy of Sciences of the United States of America* 1997, *94(*3):820–825.

20.Hackenberg D, Wu Y, Voigt A, Adams R, Schramm P, Grimm B: *Studies on Differential Nuclear Translocation Mechanism and Assembly of the Three Subunits of the Arabidopsis thaliana Transcription Factor NF-Y. Molecular Plant* 2012, *5(*4):876–888.

21.Kukimoto I, Elderkin S, Grimaldi M, Oelgeschläger T, Varga-Weisz PD: *The Histone-Fold Protein Complex CHRAC–15/17 Enhances Nucleosome Sliding and Assembly Mediated by ACF. Molecular Cell* 2004, *13(*2):265–277.

22.Wang Y-L, Faiola F, Xu M, Pan S, Martinez E: *Human ATAC Is a GCN5/PCAF-containing Acetylase Complex with a Novel NC2-like Histone Fold Module That Interacts with the TATA-binding Protein. Journal of Biological Chemistry* 2008, *283(*49):33808–33815.

23.Yang H, Zhang T, Tao Y, Wu L, Li H-t, Zhou J-q, Zhong C, Ding J: *Saccharomyces Cerevisiae MHF Complex Structurally Resembles the Histones (H3-H4)2 Heterotetramer and Functions as a Heterotetramer. Structure* 2012, *20(*2):364–370.

24.Girard C, Crismani W, Froger N, Mazel J, Lemhemdi A, Horlow C, Mercier R: *FANCM-associated proteins MHF1 and MHF2, but not the other Fanconi anemia factors, limit meiotic crossovers. Nucleic acids research* 2014, *42(*14):9087–9095.

25.Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C, Moras D, Romier C, Bolognesi M *et al: Sequence-Specific Transcription Factor NF-Y Displays Histone-like DNA Binding and H2B-like Ubiquitination. Cell* 2013, *152(*1):132–143.

26.Söding J: *Protein homology detection by HMM–HMM comparison. Bioinformatics* 2004, *21(7*):951–960.

27.Remmert M, Biegert A, Hauser A, Söding J: *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods* 2011, *9:*173.

28.Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J *et al: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology* 2011, *7:*539–539.

29.Gouet P, Courcelle E, Stuart DI, F. M: *ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics* 1999, *15(*4):305–308.

30.Finn RD, Clements J, Eddy SR: *HMMER web server: interactive sequence similarity searching. Nucleic acids research* 2011, *39(*Web Server issue):W29-W37.

31.Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England)* 2009, *25(*15):1972–1973.

32.Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ: *Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford, England)* 2009, *25(*9):1189–1191.

33.Tuimala J: *A primer to phylogenetic analysis using the PHYLIP package. Espoo Finl Cent Sci Comput Ltd* 2006:6:55.

34.Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: *MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology* 2012, *61(*3):539–542.

35.Rambaut A: *FigTree, a graphical viewer of phylogenetic trees.* 2007.

36.Osorio D, Rondon-Villarreal P, Torres R: *Peptides: A Package for Data Mining of Antimicrobial Peptides. R Journal* 2015, *7(*1):4–14.

37.Zhang Y: *I-TASSER server for protein 3D structure prediction. BMC Bioinformatics* 2008, *9(*1):40.

38.Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic acids research* 2005, *33(*Web Server issue):W299-W302.

Tables

Table 1. Showing number of proteins from different functional protein classes in each parental HFM group.

Core Histone group	HFM containing Non-histone protein class	Number of proteins	Total Proteins
H2A	NF-YC	839	1182
	DrAp1	226	
	CHRAC	23	
	DPE-C	71	
	Other	23	
H2B	NF-YB	1194	1634
	Dr1	150	
	TAF12	237	
	Other	53	
H3	CENH3	91	161
	Other	70	
H4	CEN-S	60	911
	DPE-3	87	
	TAF (6,8,9,10)	657	
	Bromodomain	18	
	Other	89	

Table 2. Showing length of HFMs of various non-histone proteins and percent positively charged residues at different conservation level.

	Total Residues in HFM	Positive Amino acids (%)			
		Total Positive	Variable	Average	Conserved
NF-YB	95	17.9	6.3	4.2	7.4
NF-YC	92	19.8	4.4	4.4	11.0
Dr1	99	14.1	2.0	7.1	5.1
DRAP1	79	19.0	5.1	1.3	12.7
TAF12b	80	16.3	2.5	5.0	8.8
TAF6	74	12.2	1.4	2.7	8.1
TAF9	73	15.1	1.4	9.6	4.1
DPE-C	74	20.3	0.0	5.4	14.9
DPE-3	81	17.3	3.7	2.5	11.1
CEN-S	75	18.7	2.7	4.0	12.0

Figures

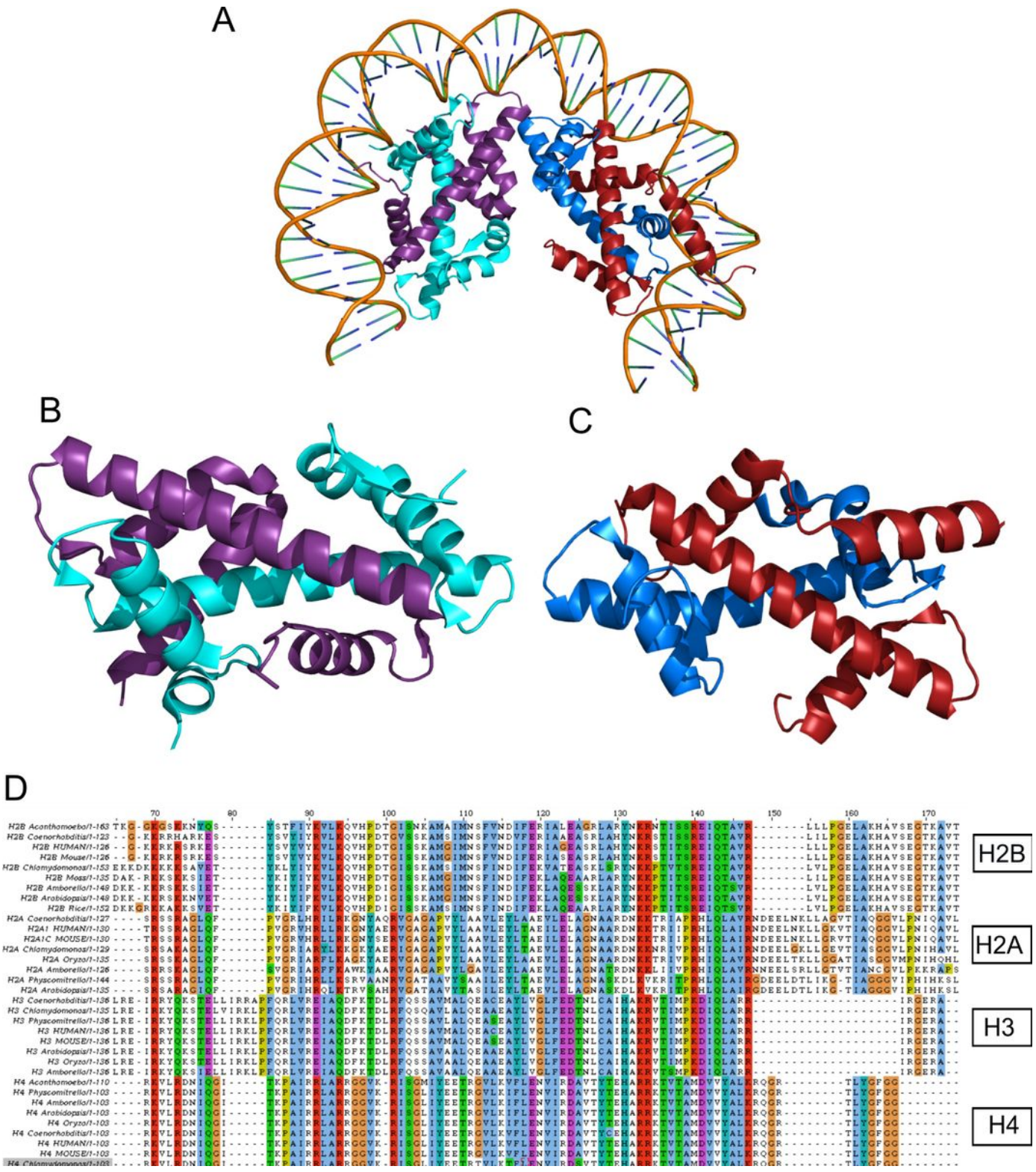

Figure 1

Structure of (A) Showing interaction of HFM of core histone with DNA, (B) H2A-H2B and (C) H3-H4 dimer showing hand shake heterodimerization. (D) Shows conservation in the Multiple Sequence Alignments of HFM sequence of core histones across eukaryotes.

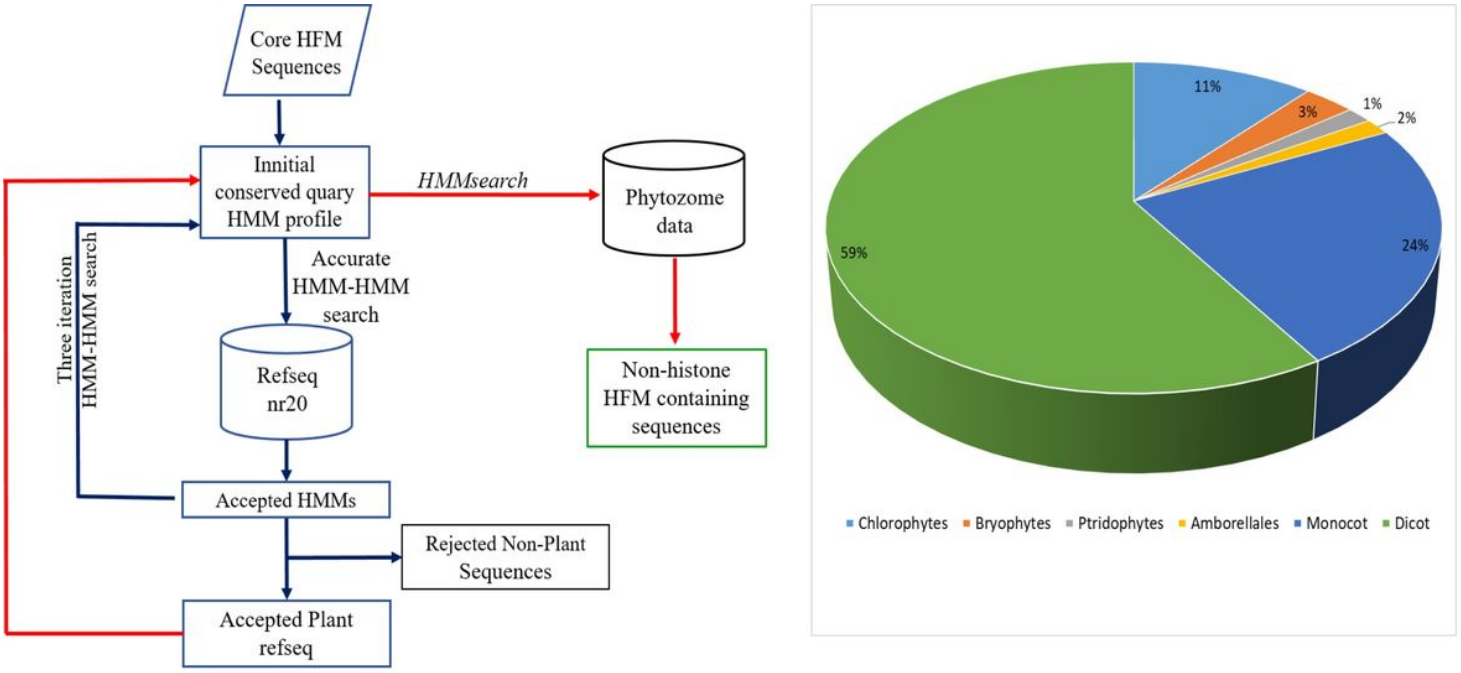

Figure 2

HFM identification pipeline (A). Pie chart showing proportion of various taxa in test data set (B).

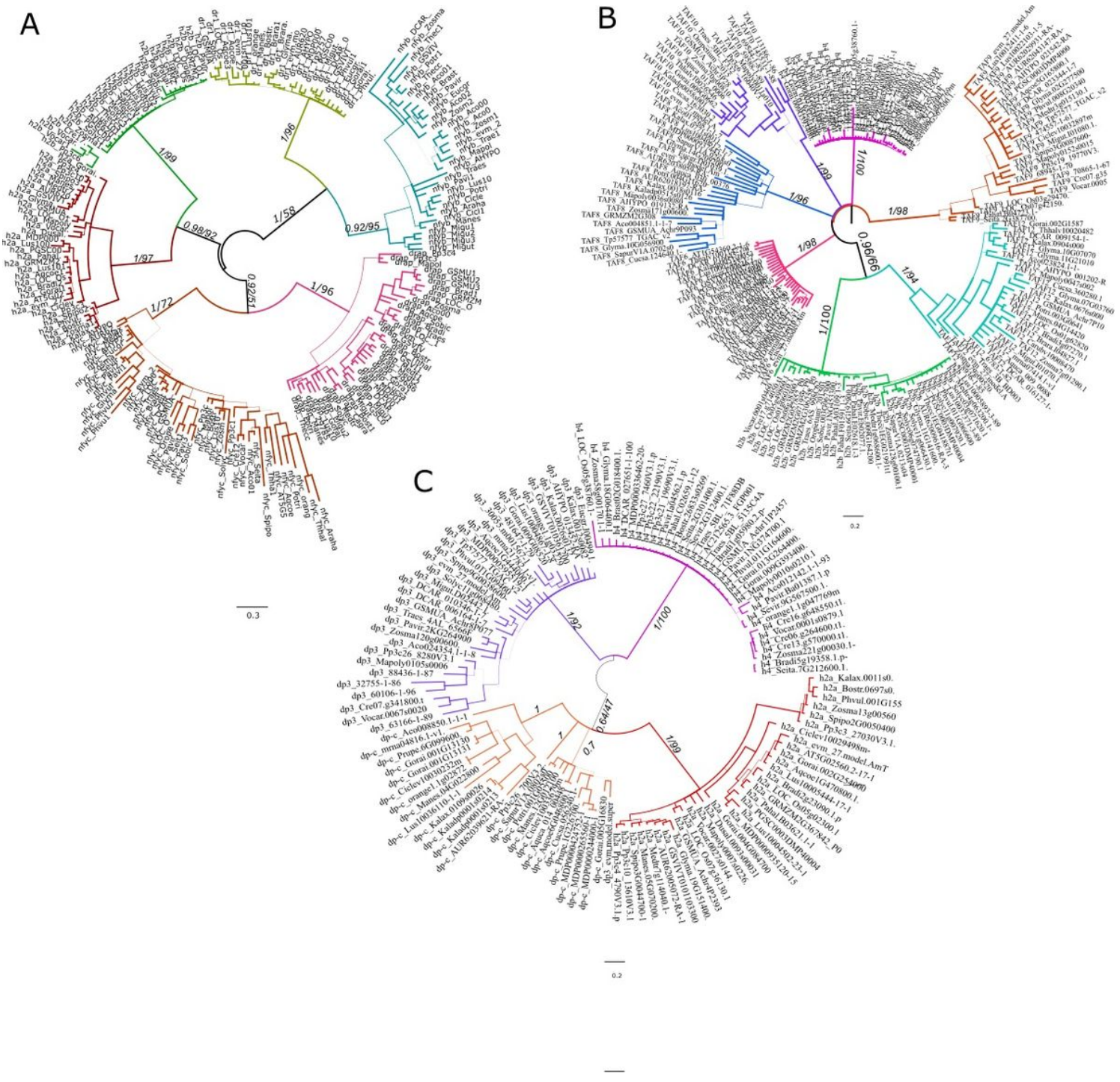

Figure 3

Phylogenetic tree (A) showing NF-YC, DrAp1 forming different group while NF-YB, Dr1 forming another group. Histone H2A and H2B forms different group suggesting that HFM of NF-YB/C and Dr1/DrAp1 evolved parallel to the core histone H2A and H2B. (B) Phylogenetic tree of HFM sequences of TAF6, TAF8, TAF9, TAF10, TAF12, H2B and H4 where TAF6, TAF8, TAF9, TAF10 and H4 are forming phylotomy where as TAF12 show close homology to H2B. (C) Phylogenetic tree sowing homology relation of H4 to the DNA polymerase epsilon subunit 3 and H2A to DNA polymerase epsilon subunit C. The width of the colored

branches represents the confidence score. Confidence score are given in a/b where a= probability of Bayesian analysis and b=% branching in neighbor joining.

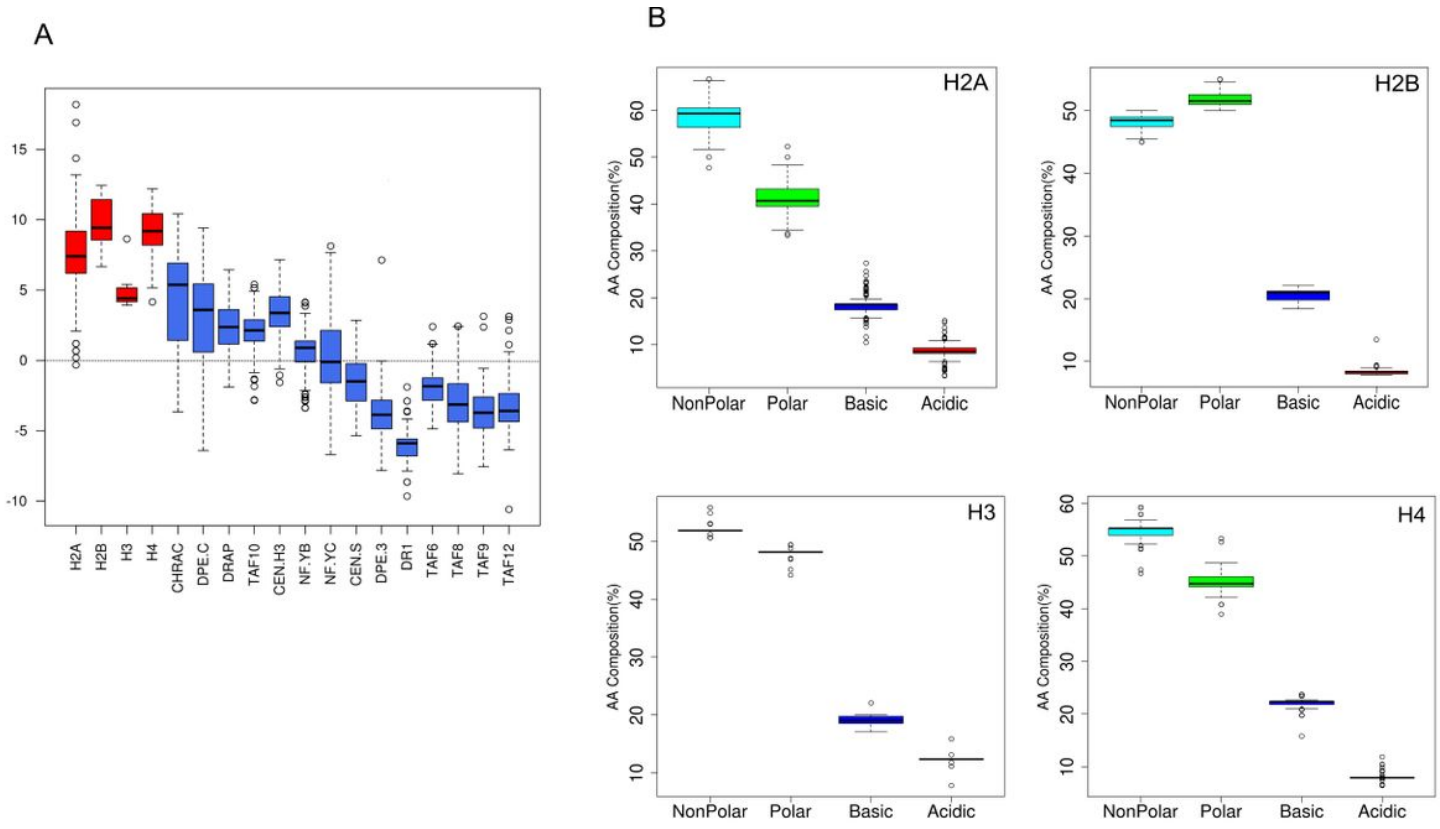

Figure 4

(A) Box plot showing charge on HFM of core histone in red and non-histone proteins in blue. HFM of histones have higher positive charge while HFM of non-histones have charge below core histone HFM and it vary from positive (CHRAC, DPE-C, DRAP1, CEN-H3, TAF10) to near neutral (NF-YB, NF-YC) and negative (CEN-S, DPE-3, Dr1, TAF6, TAF8, TAF9 and TAF12). (B) Box plot showing percent amino acid composition (Non-Polar, Polar, Basic, Acidic) of core histone HFM. All histone HFM have more than 55% Non-polar, ~20% basic and <10% acidic amino acids. Polar amino acids are found 40-50 % except H2B where it is >50%.

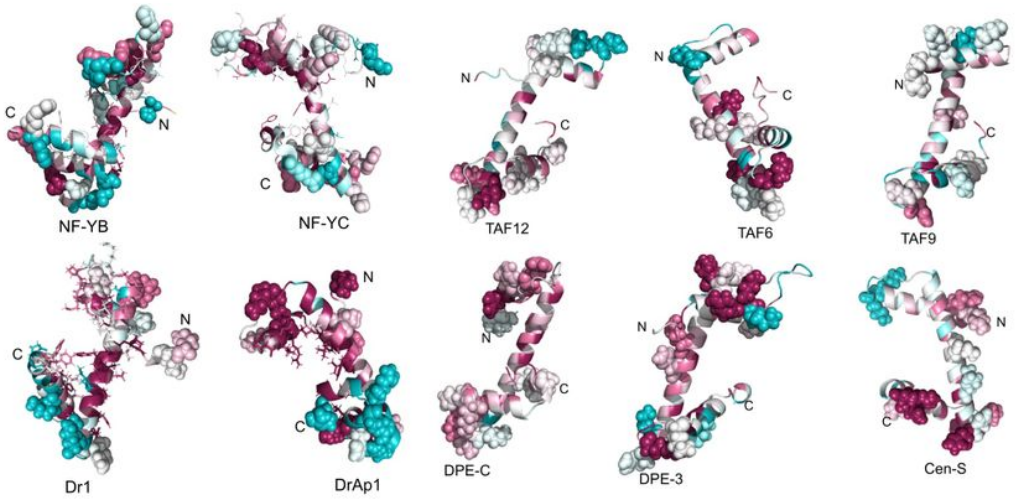

Figure 5

Conservation heat map of non-histone HFMs residues. Spheres represent positively charged residues while sticks show residues involved in dimer formation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.docx