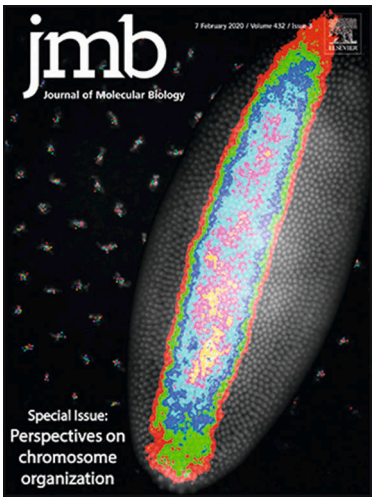
ProHistoneDB: a database of prokaryotic and viral histones

Schwab, S.; Olsthoorn, M.J.; Jansen, T.; Dame, R.T.

Journal Pre-proofs



Database

ProHistoneDB: A database of prokaryotic and viral histones

Samuel Schwab, Michel Olsthoorn, Tim Jansen, Remus T. Dame

PII: S0022-2836(26)00017-3
DOI: https://doi.org/10.1016/j.jmb.2026.169644
Reference: YJMBI 169644

To appear in: *Journal of Molecular Biology*

Accepted Date: 12 January 2026

Please cite this article as: S. Schwab, M. Olsthoorn, T. Jansen, R.T. Dame, ProHistoneDB: A database of prokaryotic and viral histones, *Journal of Molecular Biology* (2026), doi: https://doi.org/10.1016/j.jmb.2026.169644

ProHistoneDB: A database of prokaryotic and viral histones

Samuel Schwab[1,2,3], Michel Olsthoorn[1], Tim Jansen[1], Remus T. Dame[1,2,3,*]

[1]Leiden Institute of Chemistry, Leiden University, Einsteinweg 55, 2333CC Leiden, The Netherlands

[2]Centre for Microbial Cell Biology, Leiden University, Leiden, The Netherlands

[3]Centre for Interdisciplinary Genome Research, Leiden University, Leiden, The Netherlands

Abstract

Histones are one of the fundamental chromatin proteins of life. In eukaryotes and megaviruses, they form nucleosome structures that wrap DNA. However, in prokaryotes, histones are much more diverse in how they organize DNA. In bacteria, histones bend and wrap DNA while in archaea they wrap and bridge DNA. These differences in DNA organizing properties are primarily due to distinct modes of histone multimerization. Here we present ProHistoneDB, an online database describing and categorizing prokaryotic and viral histones. For each histone, monomer, dimer, tetramer, and hexamer predictions are viewable and downloadable. ProHistoneDB contains 7334 histones, categorized into 24 groups based on the multimer predictions. For each category, interactive phylogenetic trees and HMM profile logos are available to identify conserved residues and explore the relative evolutionary relationships of histones. ProHistoneDB can be accessed at `https://prohistonedb.org/`.

Keywords: Histone, Nucleoid, Chromatin, AlphaFold, ProHistoneDB

*Corresponding author: rtdame@chem.leidenuniv.nl

Introduction

Cells from all domains of life organize their genomes with DNA-binding proteins. In eukaryotes, the most important chromatin proteins are histones. Histones are small positively charged proteins that contain a characteristic histone-fold. The histone fold consists of three α-helices, $\alpha 1$, $\alpha 2$, and $\alpha 3$, which are connected through two short linkers. All eukaryotes encode the four core histones H2A, H2B, H3, and H4 [1], which heterodimerize into H2A-H2B and H3-H4. Histone dimers are the smallest functional unit that can bind to DNA. Together with DNA, the heterodimers oligomerize into larger complexes: two H3-H4 heterodimers form a V-shaped tetramer which together with two H2A-H2B heterodimers can wrap 146 base pairs of DNA into a protein-DNA complex called the nucleosome [2]. The nucleosome is the fundamental unit of chromatin in eukaryotes, and is involved in replication, repair, and gene expression. Importantly, the core histones have intrinsically disordered tails that can be post-translationally modified to regulate the behavior of the nucleosomes [3]. Megaviruses that infect eukaryotes also encode nucleosomal histones [4]. They encode the core histones, but often in unique arrangements. For example, members of the *Marseilleviridae* family have fused H2A and H2B together, as well as H3 and H4, to form pseudodimers, encoding two histone folds in one polypeptide chain [5]. These viruses use the histones to package their genome into nucleosome structures, which share high structural similarities with the eukaryotic nucleosomes [6, 7].

In prokaryotes, no single type of chromatin protein is conserved. Instead, prokaryotes encode various nucleoid-associated proteins (NAPs) and histones [8]. Like histones, NAPs are small and positively charged. However, instead of wrapping DNA, most NAPs bend DNA, bridge DNA, or form nucleofilaments along the DNA [9]. DNA wrapping histones are encoded by the majority of archaea. The best characterized canonical archaeal histones are HMfA and HMfB from *Methanothermus fervidus* and HTkA and HTkB from *Thermococcus kodakarensis* [10, 11, 12, 13, 14, 15, 16, 17]. They contain, like their eukaryotic counterparts, the histone fold but lack the disordered tails. Both the HMf and HTk histones form homodimers. The homodimers can wrap DNA into a superhelix similar to the eukaryotic nucleosome [15, 16]. However, unlike the nucleosome, the size of this superhelix is not limited to the octameric protein core, but can be extended indefinitely. This complex is referred to as the hypernucleosome [18, 19].

Archaeal nucleosomal histones have been studied since their discovery in 1990 [20]. Until recently, they were considered to be one of the few histone types in prokaryotes, with bacteria lacking histones altogether. However, recent studies have characterized non-canonical histones in archaea and bacteria that do not form hypernucleosomes, but instead bridge DNA or bend DNA, thereby functioning more like NAPs [21, 22, 23]. We recently performed an extensive bioinformatic

search and analysis of histones in prokaryotes [24]. We predicted monomer and multimer structures for every histone with AlphaFold2 and categorized them based on their quaternary structure, identifying in total 17 distinct groups of prokaryotic histones. Here, we present an online database, called ProHistoneDB, describing histones in prokaryotes and viruses. For each histone, predicted structures with their confidence values are available, as well as interactive phylogenetic trees, and interactive Hidden Markov Model logos which highlight important conserved residues. We extend the number of histone categories from 17 to 24, identifying new categories such as bacterial H2A and H2B histones, an archaeal bridging histone, and a new megavirus histone.

ProHistoneDB contents

Building on our previous work on histones in prokaryotes [24], we set up a new and easy to access database online. This database, ProHistoneDB, contains histone proteins from archaea, bacteria, and viruses, which we identify from the UniProt protein database. ProHistoneDB contains histones labelled by InterPro as "Histone-fold" (IPR009072), as well as histones we identified in UniProt with our HMM profiles, probabilistic models used to search sequence databases. In total, the database contains 7334 histones. The majority of histones are from archaea, and are either nucleosomal or face-to-face (FtF) histones (Fig. 1a-1b). The only archaeal phyla that commonly lack histones are Thermoproteota, Thermoplasmatota, and Candidates Marsarchaeota (Fig. S1). Archaea that do encode histones generally encode one to four histones, although some Asgard archaea can encode up to 12 histone proteins (Fig. S2a). In bacteria, phyla that commonly encode histones include Planctomycetota, Myxococcota, Spirochaetota, Elusimicrobiota, and Bdellovibrionota (Fig. S3). These bacteria generally encode one histone protein (Fig. S4). For each histone, we predicted monomer, dimer, tetramer, and hexamer homo-oligomer structures with AlphaFold2. Histones that form similar multimer structures were categorized together. Several of these predicted structures and properties have been confirmed experimentally, for instance HTkC and HLp form the predicted torus-shaped tetramer [24, 25], and HMfC can bridge DNA [24]. In total, ProHistoneDB contains 24 different histone categories (Fig. 1b and 2). Protein characteristics, such as protein length and isoelectric point, vary strongly between histone categories (Fig. 1c-1d). Histones longer than 70 amino acids commonly contain additional protein domains, additional histone folds, or intrinsically disordered tails. While most histones are characterized by a high isoelectric point, many histones in ProHistoneDB are, overall, strongly negatively charged. These histones are from halophiles, which are known to encode highly negatively charged proteins as adaptation to their high salt environment [26].

Most of the histone categories in ProHistoneDB overlap with the categories as defined by us in a previous study [24]. There are two super families of prokaryotic histones: nucleosomal histones and $\alpha 3$ histones. Histones in these two super families make up 63% of all prokaryotic histones. The nucleosomal histones are predicted to form a conventional histone tetramer structure, similar to eukaryotic histones H3 and H4 [2]. In archaea, the tetramer can be extended further with additional histone dimers, forming a nucleosome structure of undefined size, referred to as a hypernucleosome [15, 16]. Histones in the nucleosomal category are predicted to form structures similar to canonical hypernucleosome-forming histones, such as HMfA, HMfB, HTkA, and HTkB (Fig. 2). Nucleosomal histones generally lack N- and C-terminal tails, and are exclusively found in archaea (Fig. 1b and S5). Rare tails are found in Heimdallarchaeota, Thermoproteota, Verstraetearchaeota, Thorarchaeota, and some DPANN archaea (Fig. S5). Most archaea encode 1 or 2 nucleosomal histones, with as notable exceptions Asgard archaea and some Euryarchaeota (Fig. S2b). Asgard archaea commonly encode more than 2 nucleosomal histones. However, some Heimdallarchaeota completely lack nucleosomal histones, such as Heimdallarchaeota LC2, while others, such as Heimdallarchaeota LC3, encode up to 10. Similarly, Halobacteria archaea also lack nucleosomal histones. Halobacteria encode pseudodimer Halo doublet histones, which share characteristics of nucleosomal histones. Like nucleosomal histones, the Halo doublets are predicted to form a pseudotetramer similar to the nucleosomal tetramer. However, the Halo doublet pseudotetramer can not be further extended into a full nucleosome, as one side of the pseudodimer lacks the nucleosomal dimer-dimer interface, which in conventional nucleosomal histones facilitates nucleosome formation. Therefore, we predict that Halo doublets bind to DNA like a nucleosomal tetramer.

The other super family, $\alpha 3$ histones, are common in archaea, but are also encoded by several bacteria. While all other categories are defined by their unique quaternary structure, $\alpha 3$ histones are the only group defined by their secondary structure. The defining characteristic is a short $\alpha 3$ helix in their histone fold. We have subdivided the $\alpha 3$ super family into smaller categories, as defined by their predicted quaternary structures. These subgroups include FtF, bacterial dimer, ZZ, Rab GTPase, and phage histones (Fig. 2). FtF histones form a torus-shaped tetramer around which DNA can be wrapped [24, 25]. They are common in archaea and scarcely present in bacteria. Most archaea that encode FtF histones only encode one or two, while some Asgard archaea encode more, up to four FtF histones per proteome (Fig. S2c). Intriguingly, Halobacteria archaea, which lack nucleosomal histones, all encode at least one FtF histone. Similarly, the Heimdallarchaeota that lack nucleosomal histones, such as Heimdallarchaeota LC2, encode one or more FtF histones instead, suggesting that FtF histones per-

form similar functions in genome organization as nucleosomal histones. Archaeal FtF histones generally lack tails, with some exceptions in Heimdallarchaeota and Thorarchaeota, while bacterial FtF histones commonly have N- or C-terminal tails (Fig. S6). Bacterial dimer histones are a bacteria-exclusive histone category. They are relatively simple in that they form dimers that bend DNA. Closely related to bacterial dimers are the ZZ histones. ZZ histones contain a bacterial dimer histone-fold on the C-terminus and a ZZ-type zinc finger, which can possibly bind DNA, on the N-terminus. Rab GTPase histones are exclusive to Lokiarchaeota and contain an α3 histone fold on the C-terminus and a eukaryotic-like Rab GTPase on the N-terminus. Lastly, the phage histones contain an extra C-terminal α-helix and a zinc-binding site in the linker between the N-terminal histone fold and the C-terminal helix. The C-terminal helix is predicted to facilitate tetramerization and DNA-bridging.

The database contains several histone categories that are either predicted to bridge DNA or have been experimentally confirmed to bridge DNA. A common feature among the DNA bridging histones is a C-terminal α-helix which facilitates tetramerization. In contrast to nucleosomal histones, the dimeric histone folds of DNA bridging histones do not interact with each other in the predicted tetramer structures. As a result, the dimeric histone folds can each bind to a separate DNA duplex. The DNA-bridging categories include phage, coiled-coil (CC), Methanococcales, RdgC, Theion coiled-coil, and Nanohalo coiled-coil histones (Fig. 2). The largest DNA-bridging category are CC histones. CC histones are exclusive to archaea and are most common in DPANN and Euryarchaeota. They are defined by a long additional C-terminal α-helix which facilitates tetramerization by forming a coiled-coil. For the CC histone HMfC from *Methothermus fervidus*, DNA-bridging has been confirmed experimentally [24]. Negatively charged N- and C-terminal tails are commonly found in CC histones, however the function of these tails is unknown. Methanococcales histones are exclusive to Methanococcales archaea. For the model histone of this category, MJ1647 from *Methanocaldococcus jannaschii*, DNA-bridging has been confirmed experimentally [23]. Unlike CC histones, Methanococcales histones form a tetrameric handshake motif with their C-terminal helices. RdgC histones are a small group of hypothetical DNA-bridging histones present in Halobacteria archaea and Bacilli bacteria. Their additional C-terminal α-helix is predicted to form a tetrameric coiled-coil. The bacterial RdgC histones often have intrinsically disordered negatively charged N- and C-terminal tails. Interestingly, RdgC histones exhibit a conserved synteny [24]. The operon of the RdgC histones always includes an RdgC-like protein and an unknown transmembrane protein. The conserved synteny implies a functional, yet unknown, relationship between these three proteins. New additions to the DNA-bridging group are the Theion and the Nanohalo coiled-coil histones. Structurally,

they are predicted to be very similar to the conventional CC histones. However, we decided to keep them in separate categories due to the low sequence identity that they share with CC histones (Fig. S8). In addition, in the tetramer prediction, the topology of the coiled-coil motif in Theion coiled-coil histones differs from CC histones (Fig. S9).

The majority of Thermoplasmatota archaea do not encode histones. However, we identify two histone categories that are exclusive to Thermoplasmatota: Xer and Poseidoniia doublets (Fig. 2). Poseidoniia doublets are pseudodimeric histones exclusive to Candidatus Poseidoniia archaea. The two histone folds are on the N-terminus, while the C-terminus contains a degenerated pseudodimeric histone fold (Fig. S10-S11). Poseidoniia doublets are predicted to form dimers, whereby one pseudodimeric histone fold stacks on top of the other (Fig. S12). We hypothesize that Poseidoniia doublets bend DNA across the two pseudodimeric histone folds, similar to a nucleosomal tetramer. Xer histones are defined by their synteny and their elongated α2-helix, which is 2 turns longer than for conventional histones. Xer histones are always found in an operon that also includes a Xer-like recombinase. The largest multimer of Xer histones predicted by AlphaFold2 with confidence is a dimer. We therefore hypothesize that Xer histones bend DNA as dimers, similar to bacterial dimer histones, although the bending angle will be shallower due to the elongated form of the histone fold. Furthermore, we hypothesize that Xer histones are architectural co-factors in Xer-mediated recombination, similar to how IHF and HU facilitate genetic recombination in *Escherichia coli* [27, 28]. Another histone category that is defined by its synteny are IHF histones. IHF histones are always found in an operon that includes an IHF/HU protein (a bacterial NAP) [24]. They are exclusive to bacteria and contain two additional α-helices on the C-terminus. In the dimer predictions, the C-terminal α-helices interact with each other in a handshake motif.

ProHistoneDB contains three categories of histones that likely do not bind DNA. These include beta-propeller, DUF1931, and transmembrane histones (Fig. 2). Beta-propeller histones are exclusive to bacteria. They encode four histone folds on the N-terminus and a beta-propeller domain similar to the cartilage acidic protein 1 on the C-terminus (Fig. S13-S14). The four histone folds are predicted to form two pseudodimers that interact with each other. The way that these pseudodimers interact sequesters the α1-face of the histone folds from the solution. We therefore predict that beta-propeller histones do not bind DNA. DUF1931 are pseudodimeric histones, present in both bacteria and archaea [29]. They form dimers reminiscent of tetrameric nucleosomal or FtF histones. They lack conventional histone DNA-binding residues. In *Thermococcus kodakarensis*, DUF1931 is under the control of the heat shock regulator Phr, suggesting that these histones are involved in temperature adaptation [30]. Lastly, transmembrane histones en-

code a histone fold on their N-terminus and a transmembrane domain on their C-terminus. They are predominantly found in bacteria. Many transmembrane histones lack the α1-helix of the histone fold. As the α1-helix is critical for DNA-binding in conventional histones, we hypothesize that transmembrane histones do not bind to DNA. Membrane-related histones have been described before, such as the histone domain of the eukaryotic son of sevenless protein, which interacts with membranes [31]. Furthermore, eukaryotic histones co-localize with antimicrobial pore-forming peptides in neutrophil extracellular traps, resulting in synergistic antimicrobial activity [32]. Based on the structure predictions, the histone-fold of transmembrane histones functions as a dimerization domain (Fig. 2). While the in-vivo function of the transmembrane histones remains unclear, the histone-fold could potentially recruit other proteins. We find no conserved synteny near the transmembrane histone genes; it is therefore unclear what the identity of possible interaction partners might be. The precise localization of transmembrane histones is also not clear, as they lack a recognizable signal peptide.

New histones

With the construction of this database, we added new viral histones and identified several new histone categories. In *Streptomyces* bacteria we identify a pseudodimeric histone, which we name bacterial H2A H2B (Fig. 3a). Bacterial H2A H2B contains two histone folds, linked by a 48 amino acid linker, an intrinsically disordered N-terminal tail, and a 40 amino acid linker between the α2 and α3 helices of the C-terminal histone fold. As a monomer, the two histone folds are predicted to interact, forming a pseudodimer. The predicted dimer is structurally similar to the eukaryotic H2A H2B heterodimer, including the characteristic C-terminal helix of H2B (Fig. S15a). Despite this high similarity, the two histones share only 22% sequence identity with bacterial H2A H2B (Fig. S16-S17). Bacterial H2A H2B histones are common in *Streptomyces* but are not strictly conserved, with 17% of *Streptomyces* proteomes in UniProt containing a histone. *Streptomyces* encodes one H2A H2B histone per genome, and no H3 or H4 histones could be identified. The histones contain conserved positively charged amino acids on the α1 face of the pseudodimer, suggesting that they bind DNA (Fig. S15b-S15c and S18). Surprisingly, bacterial H2A H2B histones exhibit large sequence diversity, with sequence identities as low as 20%, even though they are found in closely related *Streptomyces* species. Bacterial H2A H2B histones cluster into five separate groups, sharing little sequence identity but identical predicted structures (Fig. S19-S20). The histone is found at the ends of *Streptomyces*' linear genome, a region where poorly conserved genes and horizontally acquired genes are commonly located. Therefore, *Streptomyces* likely acquired H2A H2B histones through hor-

izontal gene transfer, either from eukaryotes or from megaviruses. Why bacterial H2A H2B histones have seemingly evolved so rapidly in *Streptomyces*, forming the five separate groups, is not clear.

Megaviruses contain eukaryotic-like histones that form nucleosomes [33, 7, 34, 35]. We have added viral histones to ProHistoneDB and split them in four categories based on the number of histone folds that they contain: viral singlets, doublets, triplets, and quadruplets. Among megaviruses, double histones are the most common, followed by single, triple, and quadruple histones (Fig. 1b). Interestingly, the quadruplets contain all four core histones (H2A, H2B, H3, and H4), and thus are able to form a canonical eukaryotic-type nucleosome as a homo-dimer (Fig. S26). Among the Pacman-, Kaumoeba-, and Faustoviruses we identify a new group of viral histones, referred to as Pacman double histones. Pacman doublets contain two histone folds that self-dimerize (Fig. S27a). Like eukaryotic H2B, one of the histone folds contains an extra C-terminal helix (Fig. S27c). Therefore, Pacman doublets likely originate from a viral H2A H2B doublet. They are predicted to form dimers similar to a conventional nucleosomal tetramer (Fig. 3b). Pacman doublets contain strongly conserved lysines, arginines, and threonines on the α1 face of the histone, suggesting that they bind DNA similar to a conventional histone dimer. In contrast to conventional histones, Pacman doublets contain several extensions to their histone fold (Fig. S27b). These extensions vary between the Pacman-, Kaumoeba-, and Faustoviruses. In Pacmanviruses, the best characterized virus of the three [36, 37, 38], Pacman doublets have three extensions, two arms and one long "rear" α-helix. Arm 1 is inserted between α1 and α2 of the N-terminal histone fold, and contains two bent α-helices and an antiparallel β-sheet at the tip. Arm 2 is inserted between α3 and the H2B-like C-terminal helix of the C-terminal histone fold, and contains an antiparallel β-sheet, two α-helices of which one is only 3 amino acids long. Identifying conserved residues on these extensions is difficult as they vary significantly between Pacman-, Kaumoeba-, and Faustoviruses (Fig. S28). In Faustoviruses, arm 2 is only composed of two antiparallel β-sheets and arm 1 lacks the antiparallel β-sheet. Furthermore, the C-terminal histone fold lacks the H2B-like α-helix. Kaumoebaviruses lack arm 2 and the rear α-helix, while arm 1 lacks the antiparallel β-sheet. Based on an AlphaFold3 prediction, arms 1 and 2 interact and partially encircle the DNA bent around the histone folds (Fig. 3c). The rear helix extends back from the "rear" of the pseudodimeric histone folds, being located after the α3 of the N-terminal histone fold. The rear helix subsequently connects back to the α1 of the C-terminal histone fold through a 52 amino acid long intrinsically disordered linker. The C-terminal half of the helix interacts with the rear helix of the opposing Pacman doublet, possibly promoting dimerization. However, the rear helix is likely not critical for dimerization as Kaumoebaviruses lack the rear helix.

An overview of the website

The database is free to access at `prohistonedb.org` (Fig. S30). Histones can be searched by their UniProt accession codes or common names, gene/locus id, taxon name/id, proteome id, category name, and sequence. All identifiers are those as defined in UniProt. For example, to search histones from the phylum Heimdallarchaeota, you can search for "Heimdallarchaeota" or its UniProt taxon id "1936272". Each histone has a webpage which displays the predicted monomer and multimer structures (Fig. S31). By default, the prediction with the highest confidence is displayed (rank 1), colored by the predicted local difference distance test (pLDDT) value. Four less confident predictions from AlphaFold are also available. These structure files are downloadable in CIF format and come with all the files exported by ColabFold, including the multiple sequence alignment. Structures of multiple different histones can be downloaded together by adding them to the basket (Fig. S32). Above the structure, general information about the histone is displayed, including taxonomic lineage, UniProt accession, organism, gene name, and common name. Below the structure, plots of predicted aligned error (PAE), the sequence coverage, and the pLDDT values are shown, as exported by ColabFold (Fig. S33). The PAE and pLDDT are confidence values of importance when interpreting the predicted structure. The pLDDT values range from 0 to 100 and are a measure of local confidence in the secondary structure within domains. Low pLDDT values (<50) are generally viewed as a prediction of disorder. The PAE values are visualized in a 2-dimensional plot and range from 0 to 30 Å. The PAE value at (x,y) is the expected distance error of residue x relative to residue y when residue y would have been aligned to the true structure. In the case of multimer predictions, the residues of all the chains are appended on the y and x-axis and the different chains are separated by thick black lines. Low distance errors (<10Å) indicate that AlphaFold2 is confident in the relative position of the residues in question. Domains can be identified from the PAE plot by comparing the distance errors between residues from the same chain. Confident interfaces between domains can be identified from the PAE by comparing the distance errors between residues from different chains. For illustrative purposes, we use the homo-dimer prediction of histone Q6MIV3 as an example (Fig. S37). From the PAE plot, we identify two domains, one on the N-terminus and one on the C-terminus of both chains (Fig. S37a and S37b). Between these two domains is a disordered linker, supported by the low pLDDT values of these residues (Fig. S37c). The two protein chains interact through their C-terminal domains, evident from the low PAE values of the C-terminal residues between the two chains. On the other hand, their N-terminal domains have high PAE values with the C-terminal residues of both chains. This indicates that placement of the N-terminal residues in relation to the C-terminal residues is arbitrary. Lastly, both chains have a small disordered

C-terminal tail since these residues have very low pLDDT values.

Below the pLDDT plot are an interactive phylogenetic unrooted tree and an HMM logo, visualized with Phy3D and Skylign respectively (Fig. S34 and S35) [39, 40], of the histone category which the histone in question belongs to. The nodes of the tree are labelled with the UniProt accession and common name and are searchable with regular expression support. Furthermore, the nodes are colored by the phyla to which the histones belong to as defined by UniProt/NCBI. Clicking on a node provides a link to the ProHistoneDB page of the histone in question. Next to each node are the secondary structures of the histones visualized from N- to C-terminus. The alpha helices corresponding to the histone fold are colored as well as any additional category-defining secondary structures. All internal nodes contain support values as calculated by ultrafast bootstrap approximation from IQTREE2. The HMM logo shows the conserved residues among histones in the category. The logo gives the probabilities of each amino acid, occupancy and insertions, as well as insertion length. The tree and HMM files are downloadable from the category pages (Fig. S36). Together, the structures, phylogenetic trees, and HMM logos provide information on how these prokaryotic histones might function, which residues are important for their function, and how these histones relate to each other. Important to note is that all trees in ProHistoneDB are unrooted, therefore they do not show evolutionary pathways, only relative relationships. Lastly, at the bottom of the page a list of publications is displayed related to the histone in question.

Conclusion

With ProHistoneDB we have made a comprehensive online database of histone proteins in prokaryotes and viruses. The interest in non-conventional histones has grown in the last few years, with the discovery and characterization of bacterial DNA-bending histones, viral nucleosomes, and archaeal DNA-bridging histones [21, 22, 34, 23, 24]. While we have reported on new prokaryotic histones before [24], in ProhistoneDB we make this information easier to access and navigate. The proteins in ProHistoneDB originate from UniProt. As such, ProHistoneDB provides an additional layer of annotation on top of UniProt, which includes taxonomy information and links to the nucleotide sequences [41]. In terms of identifiers, ProHistoneDB mirrors UniProt, with UniProt's accession numbers being the main identifiers for each histone. For eukaryotic histones, nomenclature remains a topic of ongoing debate due to the complexity of post-translational modifications (PTMs) in eukaryotic chromatin [42]. In ProHistoneDB, nomenclature is kept relatively simple, with each histone category having its own unique name. The category names have been chosen to convey structural information (e.g. coiled-coil

histones), taxonomic information (e.g. Poseidoniia doublets), and conserved synteny information (e.g. RdgC histones). Expanded nomenclature will be necessary as more prokaryotic histones are experimentally characterized. This is especially relevant for the nucleosomal histones as these histones can contain PTMs [43] and disordered N-terminal tails [18]. The systematic and unambiguous naming as proposed for eukaryotic histones provides a useful reference framework [42]. As no literature exists for the majority of prokaryotic histones, the predicted structures, phylogenetic trees, and HMM profiles in ProHistoneDB provide a strong starting point for characterizing these histones. The structures, trees, and HMM profiles can be downloaded for further analysis. For example, the HMM profiles from our previous study have been used to identify histones in newly sequenced viral metagenomes [44]. Up-to-date information about the histones, including literature, are available on the histone category pages of ProHistoneDB. As more (meta)genomes are sequenced and more histones are characterized, we will regularly update the database with new histone entries and categories.

Methods

Identifying, predicting, and classifying histones

All histones were identified, predicted, and classified as described, with some minor deviations [24]. These deviations are described here. All histone structures were predicted with version 3 of AlphaFold2 and AlphaFold2-Multimer, using 12 recycles [45, 46]. Monomer, dimer, and tetramer structures were relaxed by AlphaFold's AMBER forcefield. The hexamer structures were not relaxed to save computation time. The complete dataset of predicted structures can be downloaded at `https://prohistonedb.org/about`. Contaminants were identified based on the UniRef50 cluster of the histones [41]. If the UniRef50 cluster contains more than 5 histones and if the histone in question is from a domain (archaea, bacteria, and viruses) which rarely appears in the UniRef50 cluster ($< 20\%$), then the histone in question is considered to be a possible contaminant.

Hidden Markov Profiles (HMM)

For each histone category, an HMM profile was generated. For each category, all histones were aligned in a multiple sequence alignment using Muscle v5 with the Super5 alignment algorithm [47]. From the multiple sequence alignments, an HMM profile was generated with HMMER3 (PyHMMER v0.10.12) [48]. With Skylign, a logo was generated for each HMM profile [40]. Each HMM logo was used with phmmer (`https://www.ebi.ac.uk/Tools/hmmer/search/phmmer`) to

identify histones that were not annotated by InterPro in UniProt [48, 49].

Phylogenetic trees

Multiple sequence alignments (MSAs) were made for each histone category using MUSCLE v5 with the Super5 alignment algorithm [47]. From the MSAs, phylogenetic trees were generated with IQ-Tree 2 [50]. For each tree, ModelFinder was used to determine the best-fit model and branch support was assessed with 1000 ultrafast bootstrap replicates.

CLANS

CLANS clustering of the Bacterial H2A H2B histones was performed with the MPI Bioinformatics Toolkit and the CLANS java application [51, 52]. Briefly, the online CLANS toolkit performed a pairwise sequence alignment of all histones. This sequence similarity matrix was loaded in CLANS (java application) and clustering was performed for 16976 rounds. Default parameters were used. From CLANS, the graph data and the attraction values between vertices were exported. The clustering was subsequently visualized in Python 3 with matplotlib. The histone categories, as determined based on the predicted structures, were visualized by generating alpha shapes.

AlphaFold3

The AlphaFold3 predictions of A0A1X6WFK8 and A0A481YTU5 were made through the web server: `https://alphafoldserver.com` [53]. For A0A1X6WFK8, two complementary DNA strands containing 55 nucleotides each (for sequence see below) and two chains of A0A1X6WFK8 were used for the prediction:

```
5’ TATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGC 3’
5’ GCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGAATCTGCAGGTGGATA 3’
```

For A0A481YTU5, two complementary DNA strands containing 146 nucleotides each (for sequence see below) and two chains of A0A481YTU5 were used for the prediction:

```
5’ ATCAATATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATG
   TTCAGCTGAATTCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGA
   ATCTGCAGGTGGATATTGAT 3’
5’ ATCAATATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATG
   TTCAGCTGAATTCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGA
   ATCTGCAGGTGGATATTGAT 3’
```

Website

The ProHistoneDB website is created in Python using the Flask micro web framework. Histone entry and category metadata are stored in an SQLite database based on JSON data. The Pandas Python library is used for faceted search functionality displayed on the main and search results pages. The website pages are rendered using Jinja HTML templates and the Bootstrap / Material Kit frontend framework. Phylogenetic trees, HMM profile logos, and protein structures are visualized with PhyD3, Skylign, and Mol* respectively [39, 40, 54].

Code and Data availability

The source code for the website is available at `https://github.com/Chromatin-Organization/prohistonedb`. The data and source code for generating the figures, phylogenetic trees, HMM profiles, PhyD3 secondary structure annotation, the contaminant identification, and the AlphaFold3 prediction files are available at 4TU: `https://doi.org/10.4121/8fd6e4f9-8b07-4821-9262-4616683a4a23`. AlphaFold2 predictions and their confidence metrics are available at `https://prohistonedb.org`.

References

[1] Valerie W. C. Soo and Tobias Warnecke. Slaying the last unicorn: discovery of histones in the microalga Nanochlorum eucaryotum. *Royal Society Open Science*, 8(2):202023, February 2021.

[2] Karolin Luger, Armin W. Mäder, et al. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, September 1997.

[3] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, February 2007.

[4] Paul B. Talbert, Karim-Jean Armache, et al. Viral histones: pickpocket's prize or primordial progenitor? *Epigenetics & Chromatin*, 15(1):21, May 2022.

[5] Albert J. Erives. Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: evidence of proto-eukaryotic provenance. *Epigenetics & Chromatin*, 10(1):55, November 2017.

[6] Marco Igor Valencia-Sánchez, Stephen Abini-Agbomson, et al. The structure of a virus-encoded nucleosome. *Nature Structural & Molecular Biology*, 28(5):413–417, May 2021.

[7] Yang Liu, Hugo Bisio, et al. Virus-encoded histone doublets are essential and form nucleosome-like structures. *Cell*, 184(16):4237–4250.e19, August 2021.

[8] Remus T. Dame, Fatema-Zahra M. Rashid, et al. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nature Reviews Genetics*, 21(4):227–242, April 2020.

[9] Martijn S. Luijsterburg, Malcolm F. White, et al. The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. *Critical Reviews in Biochemistry and Molecular Biology*, 43(6):393–418, 2008.

[10] Harsh M. Ranawat, Marc K. Cajili, et al. Cryo-EM reveals open and closed Asgard chromatin assemblies. *Molecular Cell*, 85(22):4152–4165.e5, November 2025.

[11] Samuel Schwab, Yimin Hu, et al. Histone-mediated chromatin organization in prokaryotes and viruses. *Trends in Biochemical Sciences*, 50(8):695–706, August 2025.

[12] Ilias Zarguit, Marc K. M. Cajili, et al. Modulation of archaeal hypernucleosome structure and stability by Mg2+. *Journal of Molecular Biology*, page 169533, November 2025.

[13] Amanda M. Erkelens, Bram Henneman, et al. Specific DNA binding of archaeal histones HMfA and HMfB. *Frontiers in Microbiology*, 14, April 2023.

[14] Kathryn M. Stevens and Tobias Warnecke. Histone variants in archaea - An undiscovered country. *Seminars in Cell & Developmental Biology*, 135:50–58, February 2023.

[15] Francesca Mattiroli, Sudipta Bhattacharyya, et al. Structure of histone-based chromatin in Archaea. *Science*, 357(6351):609–612, August 2017.

[16] Bram Henneman, Thomas B Brouwer, et al. Mechanical and structural properties of archaeal hypernucleosomes. *Nucleic Acids Research*, 49(8):4338–4349, May 2021.

[17] Samuel Bowerman, Jeff Wereszczynski, et al. Archaeal chromatin 'slinkies' are inherently dynamic complexes with deflected DNA wrapping pathways. *eLife*, 10:e65587, March 2021.

[18] Bram Henneman, Clara van Emmerik, et al. Structure and function of archaeal histones. *PLOS Genetics*, 14(9):e1007582, September 2018.

[19] Bram Henneman and Remus T. Dame. Archaeal histones : dynamic and versatile genome architects. *AIMS Microbiology*, 1:72–81, 2015.

[20] Kathleen Sandman, Joseph A. Krzycki, et al. HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon Methanothermus fervidus, is most closely related to histones. *Proceedings of the National Academy of Sciences of the United States of America*, 87(15):5788–5791, August 1990.

[21] Yimin Hu, Samuel Schwab, et al. Bacterial histone HBb from Bdellovibrio bacteriovorus compacts DNA by bending. *Nucleic Acids Research*, 52(14):8193–8204, August 2024.

[22] Antoine Hocher, Shawn P. Laursen, et al. Histones with an unconventional DNA-binding mode in vitro are major chromatin constituents in the bacterium Bdellovibrio bacteriovorus. *Nature Microbiology*, pages 1–14, October 2023.

[23] Sapir Ofer, Fabian Blombach, et al. DNA-bridging by an archaeal histone variant via a unique tetramerisation interface. *Communications Biology*, 6(1):1–16, September 2023.

[24] Samuel Schwab, Yimin Hu, et al. Histones and histone variant families in prokaryotes. *Nature Communications*, 15(1):7950, September 2024.

[25] Yimin Hu, Samuel Schwab, et al. DNA Wrapping by a tetrameric bacterial histone. *Nature Communications*, 16(1):11108, December 2025.

[26] Shiladitya DasSarma and Priya DasSarma. Halophiles and their enzymes: negativity put to good use. *Current Opinion in Microbiology*, 25:120–126, June 2015.

[27] Steven D. Goodman, Susan C. Nicholson, et al. Deformation of DNA during site-specific recombination of bacteriophage lambda: replacement of IHF protein by HU protein or sequence-directed bends. *Proceedings of the National Academy of Sciences of the United States of America*, 89(24):11910–11914, December 1992.

[28] Shisheng Li and Raymond Waters. Escherichia coli Strains Lacking Protein HU Are UV Sensitive due to a Role for HU in Homologous Recombination. *Journal of Bacteriology*, 180(15):3750–3756, August 1998.

[29] Yang Qiu, Valentina Tereshko, et al. The crystal structure of Aq_328 from the hyperthermophilic bacteria Aquifex aeolicus shows an ancestral histone fold. *Proteins*, 62(1):8–16, January 2006.

[30] Kathryn M Stevens, Antoine Hocher, et al. Deep Conservation of Histone Variants in Thermococcales Archaea. *Genome Biology and Evolution*, 14(1):evab274, December 2021.

[31] Jodi Gureasko, Olga Kuchment, et al. Role of the histone domain in the autoinhibition and activation of the Ras activator Son of Sevenless. *Proceedings of the National Academy of Sciences*, 107(8):3430–3435, February 2010.

[32] Tory Doolin, Henry M. Amir, et al. Mammalian histones facilitate antimicrobial synergy by disrupting the bacterial proton gradient and chromosome organization. *Nature Communications*, 11(1):3888, August 2020.

[33] Vincent Thomas, Claire Bertelli, et al. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology*, 13(6):1454–1466, June 2011.

[34] Chelsea M. Toner, Nicole M. Hoitsma, et al. Characterization of Medusavirus encoded histones reveals nucleosome-like structures and a unique linker histone. *Nature Communications*, 15(1):9138, October 2024.

[35] Terri D. Bryson, Pablo De Ioannes, et al. A giant virus genome is densely packaged by stable nucleosomes within virions. *Molecular Cell*, 82(23):4458–4470.e5, December 2022.

[36] Julien Andreani, Jacques Yaacoub Bou Khalil, et al. Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *Journal of Virology*, 91(14):10.1128/jvi.00212–17, June 2017.

[37] Sébastien Santini, Audrey Lartigue, et al. Pacmanvirus isolated from the Lost City hydrothermal field extends the concept of transpoviron beyond the family Mimiviridae. *The ISME Journal*, 19(1):wraf002, January 2025.

[38] Guillaume Blanc, Khalil Geballa-Koukoulas, et al. Pacmanvirus S19, the Second Pacmanvirus Isolated from Sewage Waters in Oran, Algeria. *Microbiology Resource Announcements*, 10(42), October 2021.

[39] Łukasz Kreft, Alexander Botzki, et al. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, 33(18):2946–2947, September 2017.

[40] Travis J. Wheeler, Jody Clements, et al. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1):7, January 2014.

[41] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025.

[42] Michael-Christopher Keogh, Genevieve Almouzni, et al. A needed nomenclature for nucleosomes. *Molecular Cell*, 85(19):3554–3561, October 2025.

[43] Béatrice Alpha-Bazin, Aurore Gorlas, et al. Lysine-specific acetylated proteome from the archaeon *Thermococcus gammatolerans* reveals the presence of acetylated histones. *Journal of Proteomics*, 232:104044, February 2021.

[44] Yang Liu, Zhuru Hou, et al. Metagenomic mining reveals novel viral histones in dsDNA viruses. *hLife*, February 2025.

[45] John Jumper, Richard Evans, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[46] Richard Evans, Michael O'Neill, et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021.

[47] Robert C. Edgar. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, 13(1):1–9, 2022.

[48] Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10), 2011.

[49] Simon C Potter, Aurélien Luciani, et al. HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1):W200–W204, July 2018.

[50] Bui Quang Minh, Heiko A Schmidt, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020.

[51] Tancred Frickey and Andrei Lupas. CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 20(18):3702–3704, 2004.

[52] Felix Gabler, Seung Zin Nam, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics*, 72(1), 2020.

[53] Josh Abramson, Jonas Adler, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.

[54] David Sehnal, Sebastian Bittrich, et al. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, July 2021.

Acknowledgments

Research on the topic of this database was funded by grants from the Netherlands Organisation for Scientific Research (VICI 016.160.613/533 and OCENW.GROOT.2019.012) to R.T.D.

The ALICE HPC cluster at Leiden University is kindly acknowledged for providing the infrastructure necessary to perform many of the computations described in this article.

UCSF ChimeraX is kindly acknowledged for the molecular graphics and analyses in this article.

Author contributions

Samuel Schwab: Writing - original draft, Software, Conceptualization, Visualization, Investigation
Michel Olsthoorn: Software (Front-end development of ProHistoneDB website), Writing - review & editing
Tim Jansen: Software (Back-end development of ProHistoneDB website), Writing - review & editing
Remus T. Dame: Writing - review & editing, Supervision, Funding acquisition, Conceptualization

Conflict of interest statement

The authors declare no conflict of interest.

Ethics statement

All authors agree with the contents of this article.

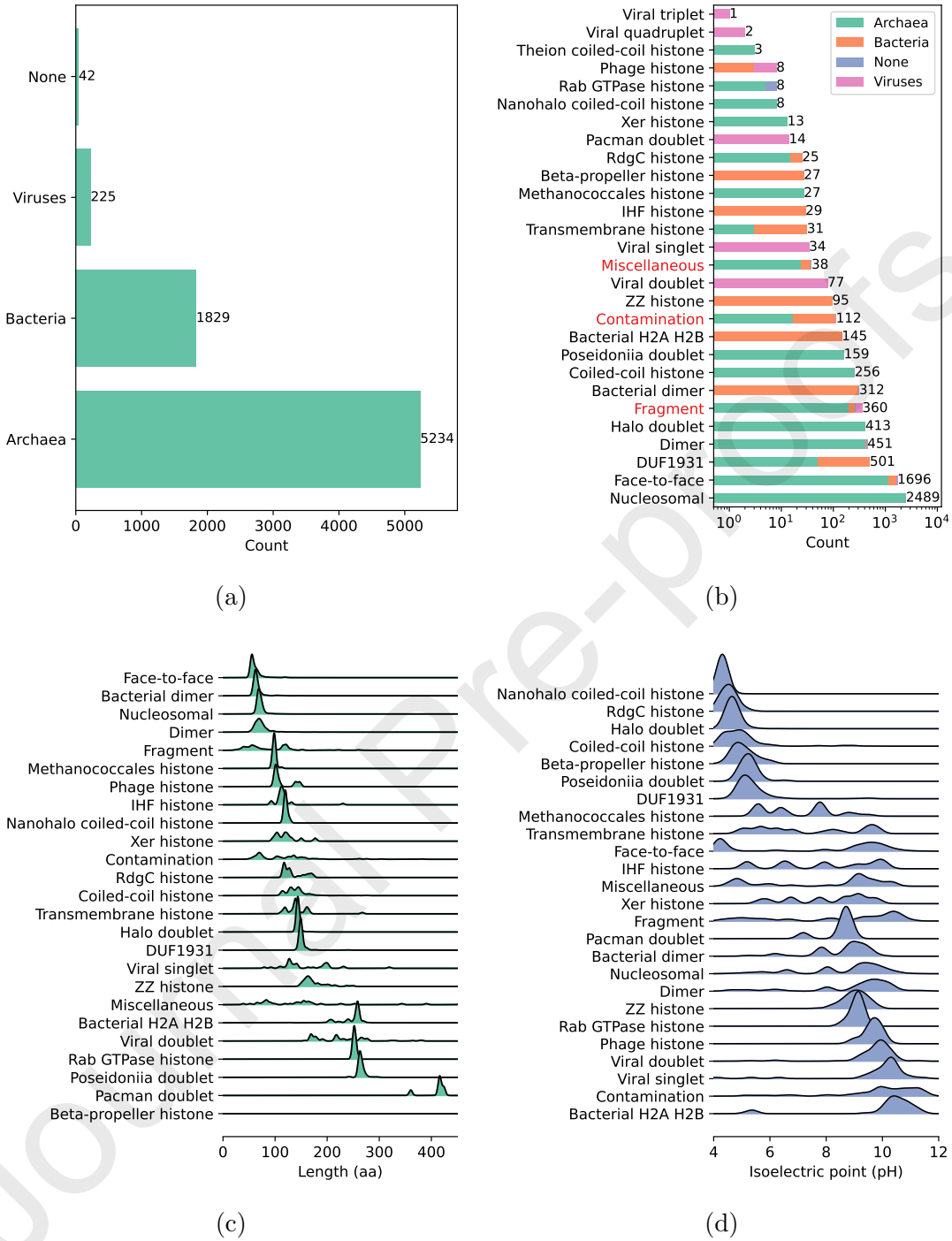


Figure 1: Overview of the composition of ProHistoneDB. (a) Number of histones in ProHistoneDB from archaea, bacteria, and viruses. Histones in "None" have no domain classification. (b) Number of histones in each of the different histone categories. Histones that do not fit any category, that are fragments, or that are contaminations are grouped in 'Miscellaneous', 'Fragment', and 'Contamination' respectively and are highlighted in red. (c,d) The length (c) and isoelectric point (d) distributions of different histone categories. Only categories containing more than 5 histones are shown.

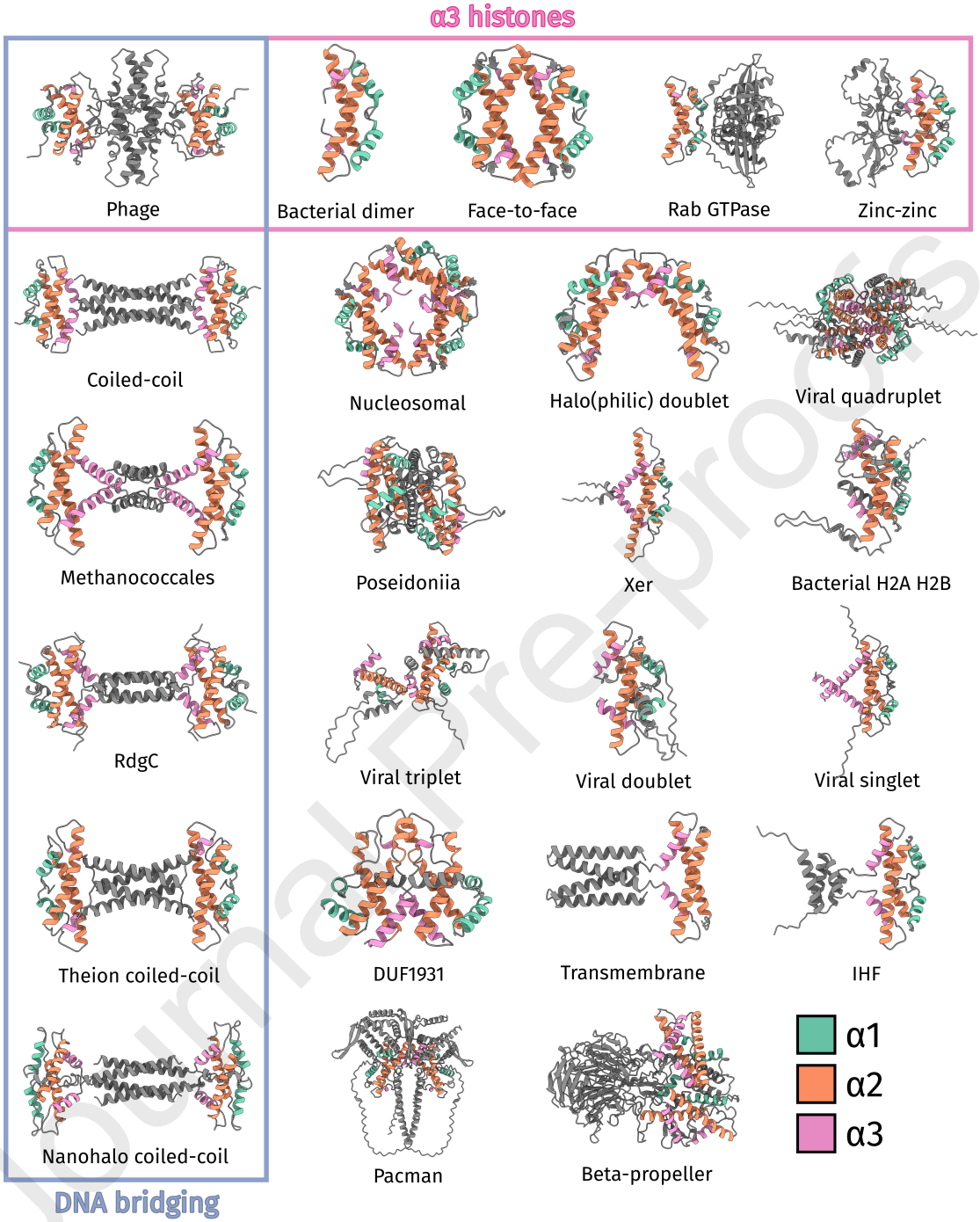


Figure 2: The 24 different histone categories of ProHistoneDB and their defining multimer structure. The structures of face-to-face, bacterial dimer, nucleosomal histones are from PDB entries 9F2C, 8CMP, and 5T5K respectively [24, 21, 15]. All other structures are predicted structures from AlphaFold2-Multimer [45, 46].

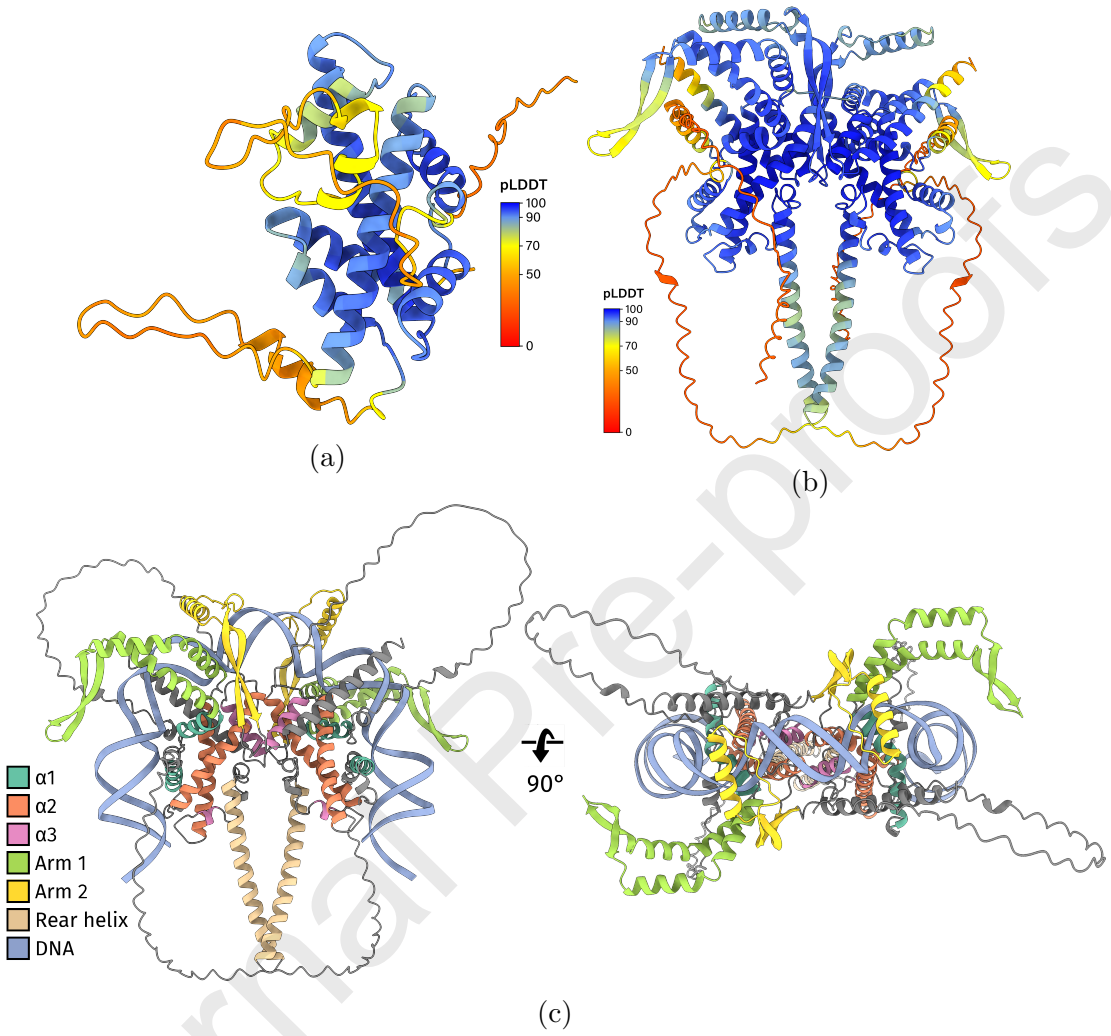


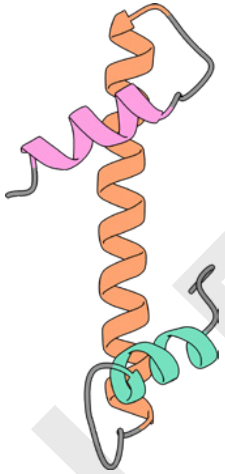
Figure 3: New histones in bacteria and viruses. (a) The monomer of bacterial H2A H2B histone Q9AK10 from *Streptomyces coelicolor* as predicted by AlphaFold2, colored by predicted local difference distance test (pLDDT) [46, 45]. (b) The dimer of pacman histone A0A1X6WFK8 from *Pacmanvirus A23* as predicted by AlphaFold2, colored by predicted local difference distance test (pLDDT) [46, 45]. (c) The A0A1X6WFK8 dimer bound to 55 base pairs of DNA as predicted by AlphaFold3, colored by structural features [53].

The authors declare no conflict of interest.

- ProHistoneDB is an online database of bacterial, archaeal, and viral histones.
- ProHistoneDB contains 7334 histones and 24 different histone categories.
- Predicted monomer and multimer structures are available for each histone.
- Interactive phylogenetic trees and HMM logos are available for each histone category.
- ProHistoneDB is available at https://prohistonedb.org/.

Graphical Abstract (for review)

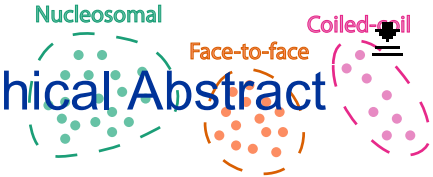
UniProt



ProHistoneDB

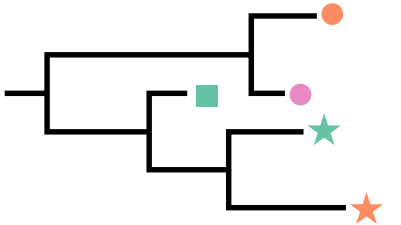
7334 histones
from prokaryotes and viruses

Monomer and multimer
structure predictions (AF2)



Classification of
24 histone groups

HMM profiles



Phylogenetic trees