Novel histones and histone variant families in prokaryotes

Samuel Schwab[1], Aimee L. Boyle[1], Remus T. Dame[1,*]

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

Abstract

Histones are important chromatin-organising proteins in eukaryotes and archaea. They form superhelical structures around which DNA is wrapped. Recent studies have shown that some archaea and bacteria contain alternative histones which exhibit different DNA binding properties, in addition to highly divergent sequences. However, the vast majority of these new histones are identified in metagenomes and thus are difficult to study in vivo. The recent revolutionary breakthroughs in *in-silico* protein structure prediction by AlphaFold2 and RoseTTAfold allow for unprecedented insights into the possible function and structure of currently unstudied proteins. Here, we categorise the complete prokaryotic histone space into 17 distinct groups based on AlphaFold predictions. We identify a new superfamily of histones, termed $\alpha 3$ histones, which are common in archaea and present in several bacteria. Importantly, we establish the existence of a large family of histones throughout archaea and in some bacteriophages that, instead of wrapping DNA, bridge DNA, therewith diverging from conventional nucleosomal histones.

Introduction

Cells from the two domains of life, the archaeal and bacterial domains, use chromatin proteins to structure and organise their genomic DNA. This stems from the basic necessity of fitting the relatively large chromosomal DNA into a cell which is orders of magnitude smaller. Chromatin proteins are varied with no single

*Corresponding author: rtdame@chem.leidenuniv.nl

group being common in all domains of life. However, their architectural properties are largely conserved. Chromatin proteins are roughly divided into three groups based on how they structure DNA: proteins that locally deform DNA by bending or wrapping the DNA, those that bridge DNA across short and long distances, and those that form long filaments on the DNA [1]. In eukaryotes, structural maintenance of chromosomes (SMC) proteins facilitate long-range interactions by bridging DNA while histones bend and wrap DNA. The four core histones, H2A, H2B, H3, and H4, are universal to all eukaryotes and together form the octameric nucleosome which wraps 150 base pairs of DNA [2]. While the four core histones diverge largely in sequence, they all share the same fundamental histone fold. This histone fold consists of approximately 65 amino acids which form three alpha helices linked together by two short linkers. Positively charged tails are present on both the N- and C-termini and are targets for post-translational modifications. The smallest functional unit of histones are dimers, which bind and bend 30 base pairs of DNA. The nucleosome is formed by consecutively linking H2A/H2B and H3/H4 heterodimers into a superhelix.

Histones are not exclusive to eukaryotes. They are also present in the majority of archaeal phyla. The model archaeal histones are HMfA/B from *Methanothermus fervidus* and HTkA/B from *Thermococcus kodakarensis* [3, 4, 5, 6, 7, 8, 9, 10]. Archaeal histones differ from their eukaryotic counterparts in that they generally lack tails and can form homodimers. However, histones of the Asgard archaea, a sister clade of eukaryotes, often contain eukaryotic-like positively charged N-terminal tails. Archaeal histones form similar nucleosomal structures called hypernucleosomes [11, 12, 13, 14]. However, while eukaryotic nucleosomes are limited to an octamer, archaeal hypernucleosomes can theoretically extend infinitely. How the size or the positioning of hypernucleosomes is regulated is not well understood. Hypernucleosome-forming histones show a preference for sequences with alternating A/T and G/C-rich regions, similar to the eukaryotic histones [15]. An artificial high-affinity sequence, "Clone20", at which tetrameric nucleosomes are formed, was identified for HMfA and HMfB. The biological relevance of the Clone20 sequence is uncertain as similar sequences have not yet been found in archaeal genomes [15, 16]. Histones have, until recently, been considered to only wrap DNA. However, histone MJ1647 from *Methanocaldococcus jannaschii* diverges from conventional histones in that it bridges DNA [17]. MJ1647's bridging ability facilitates long-range DNA-interactions, a feature which is not found in any other histone. These findings suggest that more histones might exist with alternative DNA binding properties.

Bacteria and viruses are generally considered to lack histones. Instead, bacteria contain nucleoid-associated proteins (NAPs). NAPs are histone-like in that they are generally small, highly expressed proteins with functionally similar roles to

histones. However, the protein-DNA complexes that NAPs form differ. While histones strongly wrap DNA, NAPs generally bend, or bridge DNA, and form protein-DNA filaments. Prime examples are the histone-like nucleoid structuring protein (H-NS) from *Escherichia coli* which bridges DNA and forms protein-DNA filaments, the integration host factor (IHF) which bends DNA at an angle of almost 180°, and the factor for inversion stimulation (Fis) which bends DNA at a shallower angle of 60° [18].

Histones are not altogether absent in bacteria and viruses [19, 20]. Recent studies have shown that some do indeed use histones to organise their DNA. The viral family *Marseilleviridae* contains double histone variants of the core eukaryotic histones [21]. These viral histones form nucleosomes which are structurally almost identical to the eukaryotic nucleosomes. Furthermore, these viral nucleosomes are used to package viral DNA within the virions [22]. In bacteria, *Bdellovibrio bacteriovorus* contains a highly expressed and essential histone, Bd0055 [23, 24]. Compared to eukaryotic and archaeal histones, it forms vastly different quaternary structures. A recent crystal structure suggests that Bd0055 forms long protein-DNA filaments and does not bend or wrap DNA [23].

The novel histones from *Bdellovibrio bacteriovorus* and *Methanocaldococcus jannaschii* are the first examples of histones with alternative architectural properties, from the conventional wrapping of (hyper)nucleosomal histones to bridging and filament formation. Interestingly, while these histones radically changed their DNA binding mode, their histone folds are almost identical to the nucleosomal histones. Therefore, the histone fold might represent a fundamental chromatin protein fold which can easily be changed to accommodate different architectural needs. The enormous growth of the available metagenomic sequencing data over the last decade has given us an unprecedented view of protein diversity throughout life. However, the lack of structural data has limited the functional analysis of protein homologues until recently. With the release of AlphaFold2 and RoseTTAFold, we are now able to predict monomeric and multimeric structures with atomic accuracy [25, 26, 27]. We set out to predict monomer and multimer structures of all prokaryotic histone fold proteins in the UniProt database. Here, we completely map out a largely undiscovered prokaryotic histone world, categorising all prokaryotic histones into 17 distinct groups. Alongside nucleosomal histones, prokaryotes contain a second major histone family, termed $\alpha 3$ histones, found in almost all archaeal phyla and several bacteria. Furthermore, we find various novel histones that bridge DNA, showing that histones with alternative architectural properties are widespread in prokaryotes.

Material and Methods

Identifying and predicting histones

To generate a list of histones in prokaryotes based on sequence, all bacterial, archaeal, and viral histone-fold proteins in InterPro (IPR009072) were retrieved [28]. Multiple sequence alignments (MSAs) were generated with MMseqs2 at a search-sensitive parameter of 8 [29]. Target databases used for the MSAs were constructed by the ColabFold team (https://colabfold.mmseqs.com/) and include UniRef30, BFD, Mgnify, MetaEuk, SMAG, TOPAZ, MGV, GPD, and MetaClust2 [30]. The constructed MSAs were used as an input for LocalColabFold to predict the protein structures. Monomer, dimer, tetramer, and hexamer structures were predicted for all histones with AlphaFold2 and AlphaFold2-Multimer (v2) [25, 26]. No templates were used. The monomer, dimer, and tetramer predictions were done with 6 recycles and the structures were relaxed by AlphaFold's AMBER forcefield. The hexamer predictions were done with 3 recycles and no relaxation to save computation time. Both MMseqs2 and LocalColabFold were run on the high-performance computing facility ALICE at Leiden University.

Histone classification

All histones were classified based on their predicted structures. Classification was done manually by looking at all the predicted monomer and multimer structures, the predicted alignment errors (PAE), and the predicted local distance difference test (pLDDT). Proteins that could not be classified based on their predicted structures, i.e. histones with only confident monomer and dimer predictions and no defining features in the monomer structure, were classified as "undefined". The only exception are histones we classified as "bacterial dimers", which are sequentially highly related to each other and are exclusive to bacteria. Histones which contained eukaryotic sequences in their multiple sequence alignments with a sequence identity above 60% were classified as "eukaryotic-like" and likely represent contaminations.

CLANS clustering

CLANS clustering was performed on the MPI Bioinformatics Toolkit and with the CLANS java application [31, 32]. 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 16407 cycles. Default parameters were used except for the minimal attraction parameter, which was set to 50. From CLANS, the graph data and the attraction values between

vertices were exported. The clustering was subsequently visualised in Python 3 with matplotlib. The categories were visualised by generating alpha shapes.

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 algorithm [33]. From the multiple sequence alignments an HMM profile was generated using HMMER (v3.3.2) [34]. With Skylign a logo was generated for each HMM profile [35]. The HMM profiles were also used to search against UniProt to find additional histones that were not annotated by InterPro.

Phylogenetic trees

The coiled-coil and $\alpha 3$ histone phylogenetic trees were generated with RAxML Next Generation (v1.1.0) [36]. Multiple sequence alignments were made with the coiled-coil and $\alpha 3$ histone nucleotide sequences using Muscle v5 (Super5 algorithm) [33]. For both trees, a GTR model with 4 free rates was used with a maximum likelihood estimate of the base frequencies (GTR+R4+FO). For the $\alpha 3$ tree, 40 tree searches were performed using 20 random and 20 parsimony starting trees. For the coiled-coil tree, 60 tree searches were performed using 30 random and 30 parsimony starting trees. 1000 and 600 bootstraps were performed for the coiled-coil and the $\alpha 3$ trees respectively. The bootstrap trees were used to compute the Transfer Bootstrap Expectation (TBE) support metric. The trees were visualised in R with the ggtree package [37].

Cladogram

The cladogram of the bacterial superkingdom was generated in R with the ggtree package [37]. The tree data from GTDB version 207 was used [38]. To determine which histone belongs to which bacterial phylum in the GTDB database, the GenBank assembly accession of each histone was searched against the GTDB database.

Structural similarity search

For the identification of unknown domains or proteins, FoldSeek was used [39]. The structure data of the domains in question, as predicted by AlphaFold, were used as inputs for FoldSeek. They were searched against FoldSeek's PDB100 and the AlphaFold Swiss-Prot databases using the webserver's default parameter. Structural similar proteins were subsequently analysed in ChimeraX.

Gene clusters

For gene clustering, Clinker was used [40]. For each histone, the genomic context, 5000 nucleotides before and after the histone gene, was retrieved from the NCBI database using Entrez. These genome files were used as inputs for the Clinker web server using default parameters. Briefly, Clinker performs an all vs all global alignment and automatically detects gene clusters based on sequence similarity. The resulting alignment and gene clusters were manually analysed and genes were classified based on sequence and predicted AlphaFold structures.

Results

Identifying new prokaryotic histones

To find new histones in archaea and bacteria, we made use of the protein annotation database InterPro [28, 41]. InterPro classifies all proteins within the UniProt database into families, domains, and important sites based on their sequences. For histones, InterPro has the histone-fold homologous superfamily, which is an all-encompassing category for all histone-fold-containing proteins. We requested all sequences that are part of this histone-fold homologous superfamily. To validate that these sequences indeed contain a histone fold and to gain insight into their potential function, we predicted monomer and multimer structures with AlphaFold2 [25, 26]. We find a total of 5823 histones in prokaryotes, 25% of which are from bacteria. Half of the 5823 histones have not been previously identified. All new histones contain the characteristic histone fold structure. The characteristic histone fold is defined by approximately 65 amino acids which form three alpha helices linked together by two short linkers (Fig. 1a). Histones form dimers in solution, whereby the $\alpha 2$ helices cross each other and the $\alpha 1$ and $\alpha 3$ helices end up on opposite faces of the dimer (Fig. 1b). The dimer is the smallest functional unit and can bind 30 base pairs of DNA at its $\alpha 1$ face. Conventional histones form nucleosome structures by consecutively linking dimers through their $\alpha 3$ helices and the last 8 C-terminal residues of their $\alpha 2$ helices (Figs. 1c and 1d). Conventional nucleosomal histones in eukaryotes contain long N-terminal tails. These tails are lacking in archaeal nucleosomal histones, although we find several nucleosomal histones in the Asgard archaea phylum that have long disordered N-terminal tails (Supplementary Fig. S1). All new histones differ significantly from conventional nucleosomal histones in sequence and quaternary structure. We categorised new histones based on their predicted multimer structures and visualised their position in sequence space using the clustering software CLANS [31] (Fig. 2a). In CLANS, each sequence is a point mass in 2-dimensional space. Sequences attract each other based on their sequence similarity. As a result, similar sequences group together

and form clusters. We subdivided all prokaryotic histones into 17 categories, 13 of which are completely new, based on their predicted multimer structures. These histone categories are visualised in the CLANS clustering and overlap well with the sequence clusters.

From CLANS, we identify two major prokaryotic histone families, the nucleosomal histones and a new family of histones, the $\alpha3$-truncated histones, which we refer to as $\alpha3$ histones from hereon in. The $\alpha3$ family can be subdivided into five smaller categories, each of which likely fulfils a different function, based on predicted quaternary structure and the presence of additional domains: fold-to-fold, bacterial dimer, ZZ, Rab GTPase, and phage histones. These categories are discussed in more detail below. $\alpha3$ histones are defined by a truncated $\alpha3$ helix. This $\alpha3$ helix is 3 to 4 amino acids long, which is severely truncated compared to the 10 amino acid long nucleosomal $\alpha3$ helix. Furthermore, their $\alpha2$ helix is also truncated by 4 to 5 amino acids compared to nucleosomal histones. While nucleosomal histones exist exclusively in archaea, a substantial number of these $\alpha3$ histones, 40%, are from bacteria (Supplementary Fig. S3). However, they are not well conserved within the bacterial domain as only 1.15% of bacterial proteomes in UniProt contain an $\alpha3$ histone, in contrast to archaea where $\alpha3$ histones are found in almost all phyla. Determining the amount of conservation is complicated by the fact that most bacterial histones are found in incomplete metagenomes which are often only classified up to the phylum or class level. While nucleosomal and $\alpha3$ histones form the two major families, only $\sim$69% of histones belong to either of these two families. The remaining histones are part of minor histone categories. Diversity among these minor categories is large. Some minor histones seem to have dramatically changed their architectural properties by bridging DNA instead of wrapping it. Other minor histones seem to have lost their DNA binding ability, as they lack identifiable DNA binding residues and instead gained transmembrane domains (Supplementary Fig. S4). Sequence identity scores can range from 45% down to the 'twilight zone' of 15% between histones, underlining the difficulty in finding novel histones based on sequence alone. In the subsequent sections, we go further into detail about some of these new histones. As we can not discuss every new histone type, we will focus on the most prominent new histone families, $\alpha3$ histones, and histones that likely bridge DNA.

Fold-to-fold histones

Fold-to-fold (FtF) histones make up the largest subcategory of $\alpha3$ histones (83%) and are, after nucleosomal histones, the largest group of histones in prokaryotes. FtF histones are found in all archaeal phyla except Huberarchaeota and Hydrothermarchaeota (Figure 2b). In bacteria, FtF histones are predominantly found in the phyla Spirochaetota, Planctomycetota, and Proteobacteria (Supplementary Figs.

S2 and S3). FtF histones are defined by their predicted tetramer structure (Fig. 3a). In the tetramer, two dimers interact in a manner similar to nucleosomal histones, via their $\alpha 3$ helices and the last $\sim$8 amino acids of their $\alpha 2$ helix. Unlike nucleosomal histones, both sides of the dimer interact with each other in this manner, forming a torus. To gain more insight into important residues, we aligned all FtF histones, constructed an HMM profile, and visualised this profile as an HMM logo using Skylign (Supplementary Fig. S5). The HMM logo shows which residues are strongly conserved between FtF histones, which is indicative of functional importance. The strongest conserved residues are found at the C-terminus of the protein and include residues 48, 52, 54, 56, and 61. Residues R54, T56, and D61 form an RxTxxxxD motif which is also present in nucleosomal histones from archaea (Fig. 3b and Supplementary Fig. S6). The arginine and the aspartate are responsible for structuring the L2 loop, while the threonine, which is located in the L2 loop, is involved in DNA binding. Compared to nucleosomal histones, FtF histones lack a conserved DNA-binding arginine or lysine at position 55 (R**K**TxxxxD). Residues R48 and N52 are responsible for the dimer-dimer interactions at the dyad. N52 is located at the 'front' of the dimer-dimer interface and can form hydrogens bonds with R54 of the opposing dimer. R48 is located further back within the dimer-dimer interface and forms salt bridges with the carboxyl group of residue 66 of the opposing dimer (Fig. 3c). On the $\alpha 1$ helix, a lysine is strongly conserved at position 16. Lysines can also be found in positions 12, 14, 16, and 20, although with lower frequency.

To gain more insight into the possible function of FtF histones, we examined transcriptome data from *Halobacteria*, *Thermococci*, and *Leptospira*, three taxa that contain a conserved FtF histone. Halobacteria contain one FtF histone on their chromosome and additional ones on their plasmids. The chromosomal FtF histones in *Haloferax volcanii* (*H. volcanii*) and *Halobacterium salinarum* (*H. salinarum*), HVO0196 and VNG2273H respectively, are among the top 2% of the highest expressed genes across three out of four transcriptome datasets (Supplementary Fig. S7a-S7d). As the FtF histone is highly expressed, it is a likely candidate to be the unknown protein that causes nucleosome-like organisation of chromatin in Halobacteria. Electron microscopy images of the chromosomal fibers of *H. salinarum* show beads-on-a-string-like structures [42]. These beads were estimated to have a diameter of 8.1 ± 0.6 nm, somewhat smaller than the eukaryotic nucleosomes of 11 nm. Micrococcal nuclease digestion of crosslinked *H. volcanii* chromatin showed protected DNA fragments of 50 to 60 base pairs [43], suggesting that this unknown protein binds 50 to 60 base pairs of DNA. The expected expression of this unknown protein is high as the *H. volcanii* genome is estimated to contain 14.2 nucleosomes per kilobase, 2.7 times higher than the 5.2 nucleosomes per kilobase in *Saccharomyces cerevisiae* [43]. Not all organisms with

FtF histones show extremely high expression levels. The FtF histone in *Thermococcus kodakarensis* is part of the top 7% of the highest expressed genes, with an expression level 34 times lower than that of hypernucleosome histone HTkA (Supplementary Fig. S7e). The transcriptome of the related organism *Thermococcus onnurineus* (*T. onnurineus*) was measured in three different conditions: in yeast extract-peptone-sulfur (YPS), modified minimal-CO (MMC), and modified minimal-formate (MMF) media (Supplementary Fig. S7f-S7h). In YPS and MMF, the FtF histone shows a two to three times lower expression level than its other two nucleosomal-like histones, B6YSY3 and B6YXB0. In MMF however, the FtF histone is expressed twice as highly as the hypernucleosome histones, indicating that environmental factors might play a role in regulating the expression of FtF histones in hypernucleosome-containing archaea. Similar to Halobacteria, the FtF histone in the pathogenic bacterium *Leptospira interrogans serovar Lai* is among the top 2% of highest expressed genes (Supplementary Fig. S7i). As we find no other known NAPs among the top 10% of expressed genes, the FtF histone might be Leptospira's main architectural protein. The FtF histone is essential to *Leptospira interrogans serovar Lai* as attempts to delete the gene from its genome have failed [23, 24]. The histone is found in both pathogenic and free-living saprophytic Leptospira and always contains an N-terminal tail. Tails are common in bacterial FtF histones as opposed to archaeal FtF histones which generally lack tails (Supplementary Fig. S8). There is no conserved tail sequence across bacteria, however the majority are positively charged.

We propose two possible models for how FtF histones bind to DNA. The first model is based on the previously mentioned studies on *H. volcanii*. If the FtF histone is indeed the main architectural protein of Halobacteria, it binds 50 to 60 base pairs of DNA and forms beads-on-a-string-like structures with a diameter of $8.1 \pm 0.6\,\mathrm{nm}$. In order to fit 50 base pairs, the DNA would have to be wrapped around the tetramer, similarly to how nucleosomal histones bind DNA (Fig. 3d). Furthermore, the diameter of this DNA-wrapped tetramer would be 8.3 nm, highly similar to the observed beads-on-a-string-like structures. However, one of the main problems with this model is the high amount of steric hindrance caused by bending DNA molecules close to each other. We therefore also propose a second model, with less steric hindrance, which is based on studies on the histone Bd0055 from *Bdellovibrio bacteriovorus*. A recent crystal structure of Bd0055 suggests that it binds DNA only at its L1 and L2 loops instead of across the whole dimer [23]. It belongs to the bacterial dimer group, which is, like FtF histones, a member of the $\alpha 3$ superfamily. Given that FtF histones are more similar in sequence and structure to bacterial dimers than to nucleosomal histones, it too might bind DNA exclusively at the L1 and L2 loops. As such, FtF histones might bind the DNA around the dyads of the tetramer where the L1 and L2 loops are located (Fig. 3e).

In this model, FtF histones bridge DNA instead of wrapping it as the tetramer contains two dyads on opposite ends that each independently can bind a separate DNA duplex.

Minor α3 histones

Within the α3 histone family there are four minor categories: bacterial dimers, ZZ histones, Rab GTPase histones, and phage histones. The bacterial dimers are found in Proteobacteria, Elusimicrobiota, Spirochaetota, Planctomycetota, and Chlamydiota (Supplementary Fig. S2). AlphaFold does not produce a confident multimer prediction larger than dimers for these histones, hence the name bacterial dimers (Supplementary Fig. S9a). Closer inspection of the HMM logo shows that bacterial dimers lack the conserved residues R48 and N52 from the FtF histones, which facilitate the dimer-dimer interaction in FtF histones (Supplementary Fig. S10). Similar to nucleosomal histones, we find the RKTxxxD motif in bacterial dimers (Supplementary Fig. S9b). They also contain conserved lysines on their α1 helix at positions 11, 13, and 17. K17 is also conserved in nucleosomal histones as is an arginine (R20), while K13 is similarly conserved in FtF histones (K16) (Supplementary Figs. S5, S6, and S10). The bacterial dimer Bd0055 in *Bdellovibrio bacteriovorus* HD100 is highly expressed, being part of the top 6% of highest expressed genes during its growth phase with an absolute expression level similar to IHF, HU, and SMC (Supplementary Fig. S11a). Bd0055 is essential to *Bdellovibrio bacteriovorus* and a recent crystal structure suggests it forms long protein:DNA filaments instead of wrapping DNA [23].

Very closely related to bacterial dimers are the ZZ histones. They are found predominantly in Proteobacteria (Supplementary Fig. S2). All ZZ histones have a ZZ-type zinc finger domain on their N-terminus and a bacterial-dimer-like histone on their C-terminus (Fig. 4a). This is the first ZZ-domain identified in bacteria. The ZZ-domain contains two conserved zinc binding sites, a C4 and a C2H2 site, and shows close structural similarity to the eukaryotic ZZ-domains of HERC2, p300, and ZZZ3 (Figs. 4b-4c and Supplementary Figs. S12-S13). HERC2, p300, and ZZZ3 are a ubiquitin protein ligase, histone acetyltransferase, and a histone H3 reader respectively. All three are involved in post-translational modifications of histones. Interestingly, these eukaryotic ZZ-domains bind to the tail of histone H3 which may help them localise to nucleosomes [44, 45, 46]. If the function of the ZZ-domain is conserved with its structure, it might bind H3-like tails in bacteria. Two strains that contain a ZZ-domain histone, from *Haliangium ochraceum* SMP-2 and *Halobacteriovorax marinus* SJ, contain a second histone which belongs to the FtF histone group. Both these FtF histones have N-terminal tails that look similar to the tail from H3 (Supplementary Figs. S14 and S15). The binding pocket residues of the eukaryotic ZZ-domain are, however, not conserved in the bacterial variant.

The binding pocket in HERC2, p300, and ZZZ3 is strongly negatively charged, unlike the bacterial binding pocket which is positively charged (Supplementary Fig. S16). Combined with the fact that only 50% of proteomes that have a ZZ histone also contain a second histone and that bacterial histone tails do not share a conserved sequence, we think it is unlikely that the ZZ-domain binds to tails of other histones. Transcriptome data from *Bdellovibrio bacteriovorus* HD100 shows that the ZZ histone has very low expression in both its attack and growth phases (top 43 to 70% of highest expressed genes) (Supplementary Fig. S11).

The last two minor $\alpha 3$ histone types are the Rab GTPase and phage histones. Rab GTPase histones are found exclusively in *Lokiarchaeota*. They contain a FtF-like histone on their C-terminus and a Rab GTPase domain, a subfamily of Ras GTPases which are involved in regulating membrane trafficking pathways, on their N-terminus (Fig. 5a). The closest homologues of these Rab GTPase domains are the small Rab GTPases from eukaryotes, such as Rab-1A and Rab-1B from humans. GTPase-related histones are also found in eukaryotes. The eukaryotic Ras activator Son of Sevenless (SOS) contains a double histone domain [47]. This histone domain binds to lipids and regulates SOS activity to activate Ras GTPases [48, 49]. Whether the archaeal Rab GTPase histones can bind to lipids like the histone domain in SOS is unclear as the histone domain in SOS has no sequence identity with the Rab GTPase histones. Structurally the SOS histone domain is more similar to the H2A/H2B heterodimer than to the Rab GTPase histones. The phage histones are found in prokaryotic dsDNA virus metagenomes and bacterial metagenomes. In the viral metagenomes we find tail proteins, suggesting that these bacteriophages are part of the Caudovirales order. Since some bacterial genomes also contain the histone, the bacteriophage might be a prophage. The phage histones contain an $\alpha 3$ histone on their N-terminus and an alpha helix domain on their C-terminus. In the tetramer prediction, the C-terminal domains form a tetramerisation domain (Fig. 5b). The histone domain lacks the RxTxxxD motif and shows, in general low, sequence identity with other histone types (Supplementary Fig. S17). It contains two conserved DNA binding residues: R10, and R49. These two residues correspond to the DNA binding residues R20 and K63 in the nucleosomal HMM logo. Interestingly, one of the most strongly conserved residues is an arginine on the side of the histone dimer (Supplementary Fig. S18). As the phage histone forms a tetramer structure through its C-terminal domain instead of through the histone folds, it might bridge two DNA duplexes.

DNA bridging histones

Among the 17 prokaryotic histone categories, three might bridge DNA. These are the Methanococcales, coiled-coil, and the aforementioned phage histones. The Methanococcales (Mc) histones are exclusively found in the Methanococcales or-

der and contain a tetramerisation domain on the C-terminus which facilitates DNA bridging. The DNA bridging ability of Mc histone MJ1647 from *Methanocaldococcus jannaschii* has been confirmed experimentally [17]. The coiled-coil (CC) histones are more widely distributed through Archaea, being found in Methanobacteriota, Micrarchaeota, Iainarchaeota, Altiarchaeota, Nanoarchaeota, Aenigmatarchaeota, Nanohaloarchaeota, Undinarchaeota, and Methanofastidiosa (Fig. 6a). CC histones have a long alpha helix at the C-terminal end of the histone fold, which is predicted to form a coiled-coil helical bundle in the tetramer (Fig. 6b). The tetramer structure of CC histones shows structural similarities to the predicted tetramer structure of Mc histones, whereby the two dimers stand opposite each other and interact through their C-terminal domains. However, CC and Mc histones share little sequence identity. From the HMM logo, we identify 4 possible DNA binding residues on the $\alpha 1$ helix face of CC histones: R32, K35, R45, and K49 (Supplementary Fig. S19). Two of these, R32 and K35, correspond to the DNA binding residues R20 and K23 in the nucleosomal HMM logo. We suggest that CC histones bridge DNA by binding two separate DNA duplexes at opposing histone dimers. This would be a major divergence from how conventional histones bind DNA. A common feature of CC histones are highly negatively charged tails (Supplementary Figs. S20 and S21). These tails can be present at the C- and/or N-terminus and are predicted by AlphaFold to be disordered. Interestingly, the CC histone of model archaeon *Methanothermus fervidus* lacks these tails. The function of the tails is not clear. As they are highly negatively charged, they might act as intramolecular inhibitors by occluding the DNA-binding $\alpha 1$ face of the histone. Transcriptome data of *Methanobrevibacter smithii* PS shows low expression for its CC histone, being in the top 53% of highest expressed genes (Supplementary Fig. S22).

Two additional histone categories might bridge DNA: the RdgC and Nanohaloarchaea coiled-coil histones. Both are structurally identical to CC histones in that they contain a large C-terminal helix which forms a tetrameric coiled-coil helical bundle (Fig. 6c). However, they share only 15 to 20% sequence identity with CC histones or each other. Nanohaloarchaea coiled-coil (NCC) histones are exclusive to Nanohaloarchaeota. Based on the HMM logo, residues R25, R28, and R31 are likely DNA binding residues (Supplementary Fig. S23). Two of these, R28 and R31, correspond to the DNA binding residues R20 and K23 in the nucleosomal HMM logo. Unlike CC histones, NCC lack tails. RdgC histones are present in some *Bacillus* and *Halobacteria* species. RdgC histones are special in that their genes appear in multi-gene operons. The RdgC histone is always the first gene in its operon, followed by an RdgC-like protein and a novel transmembrane (TM) protein (Supplementary Fig. S24). We believe that these three proteins are functionally related as they stay together within this multi-gene operon across vastly

different organisms. The RdgC-like protein is structurally similar to RdgC (Supplementary Fig. S25). RdgC is found in Proteobacteria and forms a ring structure as a dimer, through which it might bind DNA [50]. Functionally, it is involved in recombination and is thought to modulate RecA activity [51]. The RdgC-like protein differs from RdgC in that it contains two additional small domains, one on the C-terminus and the other on the N-terminus. Sequence identity between the two RdgC proteins is low at 18%. The TM protein is exclusively found within the context of the RdgC histone. It consists of 4 domains: a winged helix domain, an exonuclease domain, an unknown domain found in various membrane proteins, and a transmembrane domain (Supplementary Fig. S26). Identifying functional residues of the RdgC histone is difficult as sequences vary greatly across species (Supplementary Fig. S27). One of the few conserved DNA binding residues is K34, which corresponds to the DNA binding residue K23 in the nucleosomal HMM logo. Although the sequence is not strongly conserved, the $\alpha 1$ face is always positively charged, suggesting that they all bind DNA. In transcriptome data from *H. volcanii* and *Bacillus cereus* strain ATCC 10987 the RdgC histone has low expression, being present in the top 14 to 37% of highest expressed genes (Supplementary Figs. S7a,S7b,S28).

IHF-related histones

Genes of nucleoid associated proteins are generally present as single-gene operons [52]. This is true for model histones HMfA/HMfB and HTkA/HTkB but also for the discussed $\alpha 3$ histones and coiled-coil histones. However, some of our novel histones consistently appear in multi-gene operons across different organisms. One such example are the previously mentioned RdgC histones. Another more interesting example is what we refer to as IHF-related or IHF histones (Fig. 7a). IHF histones are scarcely found in bacterial metagenomes, with only 19 IHF histones in all of UniProt. The histone fold of IHF histones is most similar to nucleosomal histones with a normal length of the $\alpha 3$ helix. One of their defining structural features is a structured C-terminal tail which dimerises with another tail. The IHF histones function most likely as dimers as AlphaFold does not confidently predict larger homo-oligomer structures. They appear in metagenomes from the phyla Proteobacteria, Elusimicrobia, Candidatus Omnitrophica, Armatimonadetes, Bacteroidetes, and Candidatus Wallbacteria. In all cases there is an integration host factor-like (IHF-like) protein within the same operon as the IHF histones, hence the name IHF histones (Figs. 7b and 7c). The orientation of the two genes is always the same, with the gene for the IHF histone preceding the IHF-like gene. The fact that these proteins always appear within the same operon suggests that they are functionally related. However, what this function specifically entails is unclear.

Both the IHF histones and the IHF-like proteins show low sequence similarities across organisms, ranging from 95% identity between closely related organisms down to 25% between organisms from different phyla. The most strongly conserved residues in the histone fold are the RxTxxxxD motif and "the sprocket" R24 (Fig. 7d and Supplementary Fig. S29). There are no conserved residues on the $\alpha1$ helix except for hydrophobic residues involved in the hydrophobic packing of the dimer. Therefore, it is not clear if the histone fold binds DNA across the $\alpha1$ helices, similar to nucleosomal histones, or if it binds DNA similar to the bacterial dimer from *Bdellovibrio bacteriovorus*. The C-tail contains two alpha helices which dimerise in a handshake motif. The strongest conserved solvent-exposed residues in the C-tail are D94, S99, and Y105. The function of the C-tail is unknown. Unlike the C-tails of coiled-coil histones, AlphaFold does not predict a dimer-dimer interface at the C-tail. As it contains no strongly conserved positively charged residues, it likely does not bind to DNA. The C-terminal tail might be a platform to recruit other proteins, possibly the IHF-like protein although AlphaFold does not predict an interaction interface between the two proteins. The IHF-like protein shows low sequence similarity with either HUα/β or IHFα/β. However, the IHF-like protein might function similar to IHFβ based on the strong conservation of R45 (Supplementary Fig. S30); One of the defining features of IHFβ is its sequence specificity for the DNA motif TTR. The sequence specificity comes from R46, which is not present in HUα/β or IHFα but is conserved in the IHF-like protein (R45) [53]. Furthermore, we believe that the IHF-like protein can strongly bend DNA, similar to IHF, based on the conservation of intercalating and DNA binding residues on the beta arms (Supplementary Fig. S31).

Discussion

We have discovered a prokaryotic histone world that is more diverse than previously thought. We have categorised all prokaryotic histones into 17 distinct groups based on their predicted AlphaFold structures. All histones contain a conventional histone fold, however they differ in their predicted quaternary structures. We identify a new major histone superfamily, called $\alpha3$ histones, which is found in almost all archaeal phyla, several bacteria, and some bacteriophages. This is the first histone from a bacteriophage reported to date. Furthermore, we find histones throughout archaea that likely exhibit DNA bridging properties, a major divergence from the DNA wrapping properties of nucleosomal histones.

We believe that $\alpha3$ histones likely originate from archaea, given the more widespread appearance of these histones within the archaeal domain compared to the bacterial domain. $\alpha3$ histones likely share a common ancestor with nucleosomal histones given the conservation of the nucleosomal RxTxxxxD motif. Bacteria

might have acquired fold-to-fold (FtF) histones through horizontal gene transfer. Our phylogenetic tree of $\alpha 3$ histones shows several examples where horizontal gene transfer likely occurred between archaea and bacteria, such as the common ancestor histone between Leptospira and Iainarchaeota FtF histones. The bacteria exclusive $\alpha 3$ histones, such as the bacterial dimer and ZZ histones, might have evolved from these horizontally acquired FtF histones.

The origin of some novel histone groups is still rather unclear. The IHF histones, for example, are found in only 19 metagenomes across vastly different organisms. We refer to these types of histones as "exotic" histones as they are rare and they do not seem to be conserved by any known group of organisms. However, proteins that are not conserved by any group do not exist as they will die out over time. Therefore, these exotic histones might be conserved in as-yet-undiscovered organisms. Time will tell as the amount of metagenomic data increases.

Assigning functions to these novel histones is difficult based on structure alone. To approximate the function, we compared the predicted structures and the conserved residues to literature. In the case of the coiled-coil histones, the predicted tetramer is highly similar to the predicted structure of the bridging histone MJ1647, hence our expectation that coiled-coil histones also bridge DNA. However, there is currently limited literature available on prokaryotic histones. We hope that our findings will provide the scientific community with a useful resource to investigate these histones using experimental methods. One of the main hurdles is that several of these novel histone groups are found either exclusively in metagenomes or in archaea that are notoriously difficult to culture. The most accessible new histone groups are the fold-to-fold, coiled-coil, RgdC, bacterial dimer, and ZZ histones as these can be found in culturable bacteria or archaea.

Lastly, while our bioinformatical analysis is limited to only histones, the application of AlphaFold2 to find distant protein homologues and gain insight into their function is applicable to all proteins. This is possible as almost all of UniProt now has AlphaFold predictions. While traditional structural alignment programmes are far too slow to search through this massive amount of structural data, the recent development of FoldSeek allows anyone to easily find structural similar proteins within a few seconds. These structural analyses will bring new insights into the conservation of proteins across life and the relationships between structure, function, and sequence.

Acknowledgments

Vikram Alva, Yimin Hu, Finn Werner, and Fabian Blombach are kindly acknowledged for thorough reading and comments on the manuscript.

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 performed all computational analyses described in this article.
Samuel Schwab wrote the manuscript.
Remus T. Dame and Aimee L. Boyle supervised the project and reviewed and corrected the manuscript.

References

[1] Martijn S. Luijsterburg, Malcolm F. White, Roel Van Driel, and Remus T. Dame. The Major Architects of Chromatin : Architectural Proteins in Bacteria , Archaea and Eukaryotes. *Critical Reviews in Biochemistry and Molecular Biology*, 9238, 2009.

[2] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[3] Kathleen Sandman, Joseph A. Krzycki, Beate Dobrinski, Rudi Lurz, and John N. Reeve. 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*, 1990.

[4] Kathleen Sandman, Rowan A. Grayling, Beate Dobrinski, Rudi Lurz, and John N. Reeve. Growth-phase-dependent synthesis of histones in the archaeon Methanothermus fervidus. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12624–12628, 1994.

[5] Mary R. Starich, Kathleen Sandman, John N. Reeve, and Michael F. Summers. NMR structure of HMfB from the hyperthermophile, Methanothermus fervidus, confirms that this archaeal protein is a histone. *Journal of Molecular Biology*, 255(1):187–203, 1996.

[6] Klaas Decanniere, Kathleen Sandman, John N. Reeve, and Udo Heinemann. Crystallization and preliminary X-ray characterization of the Methanothermus fervidus histones HMfA and HMfB. *Proteins: Structure, Function and Genetics*, 24(2):269–271, 1996.

[7] Hiroki Higashibata, Shinsuke Fujiwara, Masahiro Takagi, and Tadayuki Imanaka. Analysis of DNA compaction profile and intracellular contents of archaeal histones from Pyrococcus kodakaraensis KOD1. *Biochemical and Biophysical Research Communications*, 258(2):416–424, 1999.

[8] Klaas Decanniere, Allampura M. Babu, Kathleen Sandman, John N. Reeve, and Udo Heinemann. Crystal structures of recombinant histones HMfA and HMfB, from the hyperthermophilic archaeon Methanothermus fervidus. *Journal of Molecular Biology*, 303(1):35–47, 2000.

[9] Divya J. Soares, Kathleen Sandman, and John N. Reeve. Mutational analysis of archaeal histone-DNA interactions. *Journal of Molecular Biology*, 297(1):39–47, 2000.

[10] Lubomira Čuboňová, Masahiro Katano, Tamotsu Kanai, Haruyuki Atomi, John N. Reeve, and Thomas J. Santangelo. An archaeal histone is required for transformation of Thermococcus kodakarensis. *Journal of Bacteriology*, 194(24):6864–6874, 2012.

[11] Francesca Mattiroli, Sudipta Bhattacharyya, Pamela N. Dyer, Alison E. White, Kathleen Sandman, Brett W. Burkhart, Kyle R. Byrne, Thomas Lee, Natalie G. Ahn, Thomas J. Santangelo, John N. Reeve, and Karolin Luger. Structure of histone-based chromatin in Archaea. *Science*, 357(6351):609–612, 2017.

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

[13] Bram Henneman, Clara van Emmerik, Hugo van Ingen, and Remus T. Dame. Structure and function of archaeal histones. *PLoS Genetics*, 14(9):1–21, 2018.

[14] Bram Henneman, Thomas B. Brouwer, Amanda M. Erkelens, Gert-Jan Kuijntjes, Clara van Emmerik, Ramon A. van der Valk, Monika Timmer, Nancy C. S. Kirolos, Hugo van Ingen, John van Noort, and Remus T. Dame. Mechanical and structural properties of archaeal hypernucleosomes. *Nucleic Acids Research*, 49(8):4338–4349, 2021.

[15] Kathryn A. Bailey, Suzette L. Pereira, Jonathan Widom, and John N. Reeve. Archaeal Histone Selection of Nucleosome Positioning Sequences and the Procaryotic Origin of Histone-dependent Genome Evolution. *Journal of Molecular Biology*, pages 25–34, 2000.

[16] Amanda M. Erkelens, Bram Henneman, Ramon A. Van Der Valk, Nancy C. S. Kirolos, and Remus T. Dame. Specific DNA binding of archaeal histones HMfA and HMfB. *Frontiers in Microbiology*, (April):1–9, 2023.

[17] Sapir Ofer, Fabian Blombach, Amanda Erkelens, Declan Barker, Samuel Schwab, Katherine Smollett, Dorota Matelska, Thomas Fouqueau, Nicholas Kent, Remus Dame, and Finn Werner. DNA-bridging by an archaeal histone variant via a unique tetramerisation interface. *Research Square*, 2023.

[18] Remus T. Dame, Fatema Zahra M. Rashid, and David C. Grainger. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nature Reviews Genetics*, 21(4):227–242, 2020.

[19] Vikram Alva and Andrei N. Lupas. Histones predate the split between bacteria and archaea. *Bioinformatics*, 35(14):2349–2353, 2019.

[20] Paul B. Talbert, Karim Jean Armache, and Steven Henikoff. Viral histones: pickpocket's prize or primordial progenitor? *Epigenetics and Chromatin*, 15(1):1–20, 2022.

[21] Yang Liu, Hugo Bisio, Chelsea Marie Toner, Sandra Jeudy, Nadege Philippe, Keda Zhou, Samuel Bowerman, Alison White, Garrett Edwards, Chantal Abergel, and Karolin Luger. Virus-encoded histone doublets are essential and form nucleosome-like structures. *Cell*, 184(16):4237–4250.e19, 2021.

[22] Terri D. Bryson, Pablo De Ioannes, Marco Igor Valencia-Sánchez, Jorja G. Henikoff, Paul B. Talbert, Rachel Lee, Bernard La Scola, Karim Jean Armache, and Steven Henikoff. A giant virus genome is densely packaged by stable nucleosomes within virions. *Molecular Cell*, 82(23):4458–4470.e5, 2022.

[23] Antoine Hocher, Shawn P. Laursen, Paul Radford, Jess Tyson, Carey Lambert, Kathryn M. Stevens, Mathieu Picardeau, R. Elizabeth Sockett, Karolin Luger, and Tobias Warnecke. Histone-organized chromatin in bacteria. *bioRxiv*, 2023.

[24] Yimin Hu, Silvia Deiss, Joe D. Joiner, Marcus D. Hartmann, Andrei N. Lupas, Birte Hernandez Alvarez, and Vikram Alva. Atomic resolution structure of a DNA-binding histone protein from the bacterium Bdellovibrio bacteriovorus. *bioRxiv*, page 2023.02.26.530074, 2023.

[25] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[26] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021.

[27] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. Van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[28] Matthias Blum, Hsin Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A. Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H. Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A. Natale, Marco Necci, Christine A. Orengo, Arun P. Pandurangan, Catherine Rivoire, Christian J.A. Sigrist, Ian Sillitoe, Narmada Thanki, Paul D. Thomas, Silvio C.E. Tosatto, Cathy H. Wu, Alex Bateman, and Robert D. Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, 2021.

[29] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[30] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022.

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

[32] Felix Gabler, Seung Zin Nam, Sebastian Till, Milot Mirdita, Martin Steinegger, Johannes Söding, Andrei N. Lupas, and Vikram Alva. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics*, 72(1), 2020.

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

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

[35] Travis J. Wheeler, Jody Clements, and Robert D. Finn. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1):1–9, 2014.

[36] Alexey M. Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, 2019.

[37] Shuangbin Xu, Lin Li, Xiao Luo, Meijun Chen, Wenli Tang, Li Zhan, Zehan Dai, Tommy T. Lam, Yi Guan, and Guangchuang Yu. Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta*, 1(4):5–8, 2022.

[38] Donovan H. Parks, Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre Alain Chaumeil, and Philip Hugenholtz. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 2022.

[39] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Cameron L.M. Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, (3):2022.02.07.479398, 2022.

[40] Cameron L.M. Gilchrist and Yit Heng Chooi. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, 37(16):2473–2475, 2021.

[41] Alex Bateman, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Ray Coetzee, Austra Cukura, Alan da Silva, Paul Denny, Tunca Dogan, Thank God Ebenezer, Jun Fan, Leyla Garcia Castro, Penelope Garmiri, George Georghiou, Leonardo Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Petteri Jokinen, Vishal Joshi, Dushyanth Jyothi, Antonia Lock, Rodrigo Lopez, Aurelien Luciani, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Manuela Menchi, Alok Mishra, Katie Moulang, Andrew Nightingale, Carla Susana Oliveira, Sangya Pundir, Guoying Qi, Shriya Raj, Daniel Rice, Milagros Rodriguez Lopez, Rabie Saidi, Joseph Sampson, Tony Sawford, Elena Speretta, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Vladimir Volynkin, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Edouard de Castro, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Arnaud Kerhornou, Vicente Lara, Philippe Le Mercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Batista Neto, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Christian Sigrist, Karin Sonesson, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, Jian Zhang, Patrick Ruch, and Douglas Teodoro.

UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[42] S. Takayanagi, S. Morimura, H. Kusaoke, Y. Yokoyama, K. Kano, and M. Shioda. Chromosomal structure of the halophilic archaebacterium Halobacterium salinarium. *Journal of Bacteriology*, 174(22):7207–7216, 1992.

[43] Ron Ammar, Dax Torti, Kyle Tsui, Marinella Gebbia, Tanja Durbic, Gary D. Bader, Guri Giaever, and Corey Nislow. Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife*, 2012(1):1–11, 2012.

[44] Jiuyang Liu, Zhaoyu Xue, Yi Zhang, Kendra R. Vann, Xiaobing Shi, and Tatiana G. Kutateladze. Structural Insight into Binding of the ZZ Domain of HERC2 to Histone H3 and SUMO1. *Structure*, 28(11):1225–1230.e3, 2020.

[45] Yi Zhang, Yongming Xue, Jiejun Shi, Jae Woo Ahn, Wenyi Mi, Muzaffar Ali, Xiaolu Wang, Brianna J. Klein, Hong Wen, Wei Li, Xiaobing Shi, and Tatiana G. Kutateladze. The ZZ domain of p300 mediates specificity of the adjacent HAT domain for histone H3. *Nature Structural and Molecular Biology*, 25(9):841–849, 2018.

[46] Wenyi Mi, Yi Zhang, Jie Lyu, Xiaolu Wang, Qiong Tong, Danni Peng, Yongming Xue, Adam H. Tencer, Hong Wen, Wei Li, Tatiana G. Kutateladze, and Xiaobing Shi. The ZZ-type zinc finger of ZZZ3 modulates the ATAC complex-mediated histone acetylation and gene activation. *Nature Communications*, 9(1):1–9, 2018.

[47] Holger Sondermann, Stephen M. Soisson, Dafna Bar-Sagi, and John Kuriyan. Tandem Histone Folds in the Structure of the N-Terminal Segment of the Ras Activator Son of Sevenless. *Structure*, 11(12):1583–1593, 2003.

[48] Holger Sondermann, Bhushan Nagar, Dafna Bar-Sagi, and John Kuriyan. Computational docking and solution x-ray scattering predict a membrane-interacting role for the histone domain of the Ras activator son of sevenless. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16632–16637, 2005.

[49] Jodi Gureasko, Olga Kuchment, Debora Lika Makino, Holger Sondermann, Dafna Bar-Sagi, and John Kuriyan. Role of the histone domain in the autoinhibition and activation of the Ras activator Son of Sevenless. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3430–3435, 2010.

[50] Geoffrey S. Briggs, Paul A. McEwan, Jing Yu, Timothy Moore, Jonas Emsley, and Robert G. Lloyd. Ring structure of the Escherichia coli DNA-binding protein RdgC associated with recombination and replication fork repair. *Journal of Biological Chemistry*, 282(17):12353–12357, 2007.

[51] Julia C. Drees, Sindhu Chitteni-Pattu, Darrell R. McCaslin, Ross B. Inman, and Michael M. Cox. Inhibition of RecA protein function by the RdgC protein from Escherichia coli. *Journal of Biological Chemistry*, 281(8):4708–4717, 2006.

[52] Antoine Hocher, Guillaume Borrel, Khaled Fadhlaoui, Jean François Brugère, Simonetta Gribaldo, and Tobias Warnecke. Growth temperature and chromatinization in archaea. *Nature Microbiology*, 7(11):1932–1942, 2022.

[53] L. Zulianello, P. Van Ulsen, P. Van de Putte, and N. Goosen. Participation of the flank regions of the integration host factor protein in the specificity and stability of DNA binding. *Journal of Biological Chemistry*, 270(30):17902–17907, 1995.

[54] Fabian Blombach, Darya Ausiannikava, Angelo Miguel Figueiredo, Zoja Soloviev, Tanya Prentice, Mark Zhang, Nanruoyi Zhou, Konstantinos Thalassinos, Thorsten Allers, and Finn Werner. Structural and functional adaptation of Haloferax volcanii TFEa/b. *Nucleic Acids Research*, 46(5):2308–2320, 2018.

[55] Adrián López García de Lomana, Ulrike Kusebauch, Arjun V. Raman, Min Pan, Serdar Turkarslan, Alan P. R. Lorenzetti, Robert L. Moritz, and Nitin S. Baliga. Selective Translation of Low Abundance and Upregulated Transcripts in Halobacterium salinarum. *mSystems*, 5(4), 2020.

[56] Saaz Sakrikar and Amy K. Schmid. An archaeal histone-like protein regulates gene expression in response to salt stress. *Nucleic acids research*, 49(22):12732–12743, 2021.

[57] Dominik Jäger, Konrad U. Förstner, Cynthia M. Sharma, Thomas J. Santangelo, and John N. Reeve. Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics*, 15(1):1–15, 2014.

[58] Suhyung Cho, Min Sik Kim, Yujin Jeong, Bo Rahm Lee, Jung Hyun Lee, Sung Gyun Kang, and Byung Kwan Cho. Genome-wide primary transcriptome analysis of H 2 -producing archaeon Thermococcus onnurineus NA1. *Scientific Reports*, 7(January):1–12, 2017.

[59] Feng Xue, Haiyan Dong, Jinyu Wu, Zuowei Wu, Weilin Hu, Aihua Sun, Bryan Troxell, X. Frank Yang, and Jie Yan. Transcriptional responses of Leptospira interrogans to host innate immunity: Significant changes in metabolism, oxygen tolerance, and outer membrane. *PLoS Neglected Tropical Diseases*, 4(10), 2010.

[60] Iris Karunker, Or Rotem, Mally Dori-Bachash, Edouard Jurkevitch, and Rotem Sorek. A Global Transcriptional Switch between the Attack and Growth Forms of Bdellovibrio bacteriovorus. *PLoS ONE*, 8(4), 2013.

[61] Cédric Notredame, Desmond G. Higgins, and Jaap Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000.

[62] Elizabeth E. Hansen, Catherine A. Lozupone, Federico E. Rey, Meng Wu, Janaki L. Guruge, Aneesha Narra, Jonathan Goodfellow, Jesse R. Zaneveld, Daniel T. McDonald, Julia A. Goodrich, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences of the United States of America*, 108(SUPPL. 1):4599–4606, 2011.

[63] Simen M. Kristoffersen, Chad Haase, M. Ryan Weil, Karla D. Passalacqua, Faheem Niazi, Stephen K. Hutchison, Brian Desany, Anne Brit Kolstø, Nicolas J. Tourasse, Timothy D. Read, and Ole Andreas Økstad. Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome biology*, 13(4), 2012.

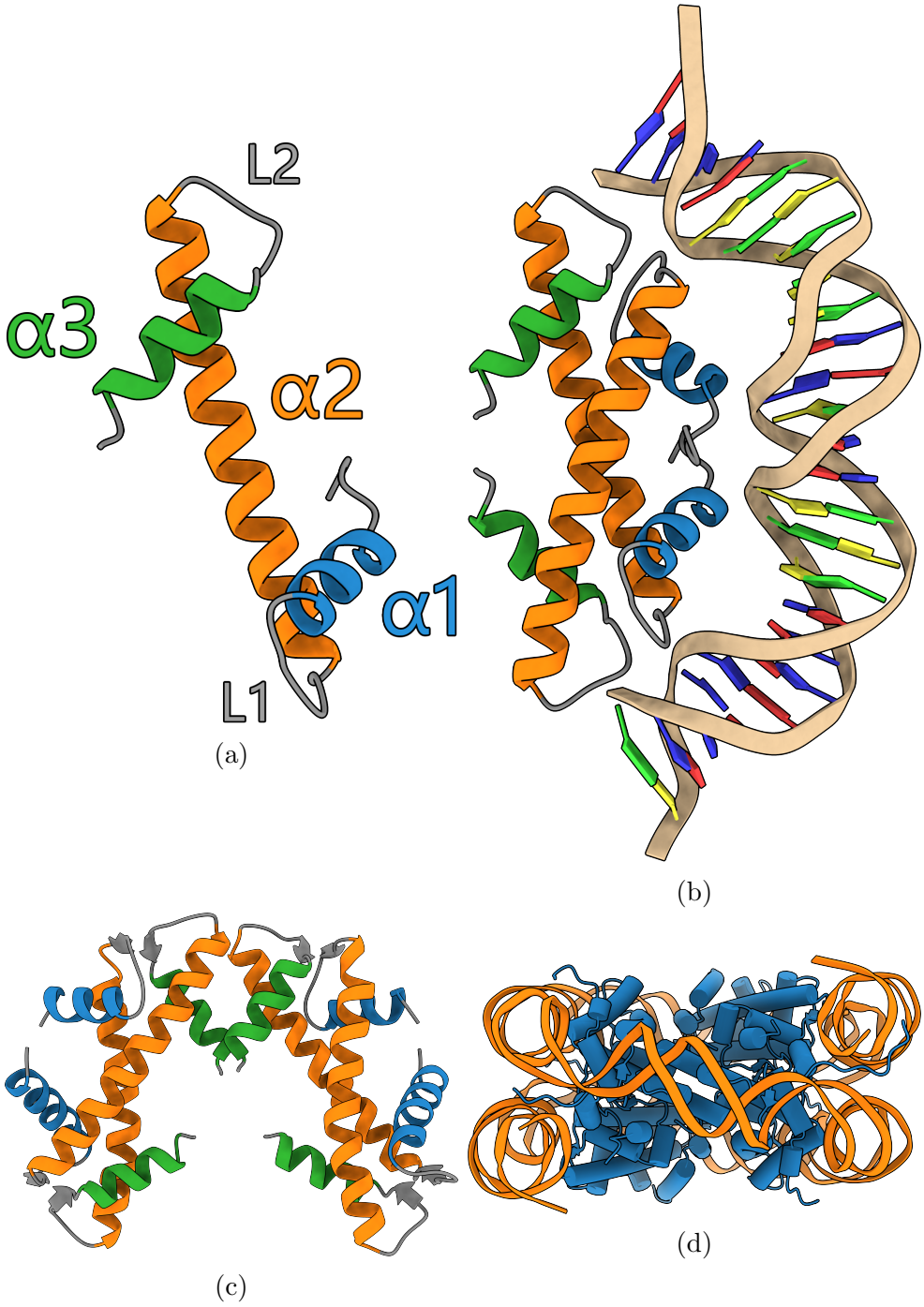
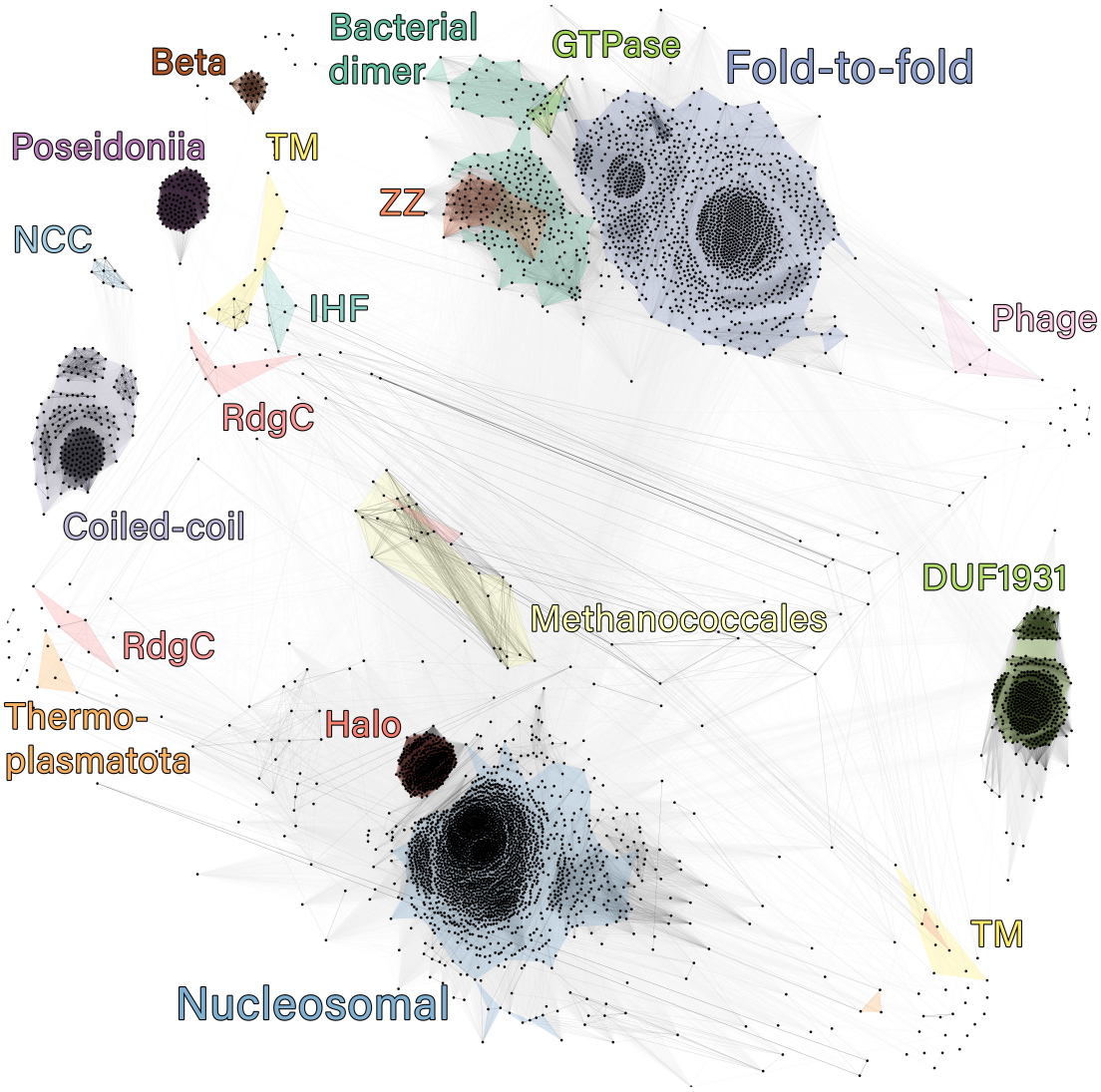
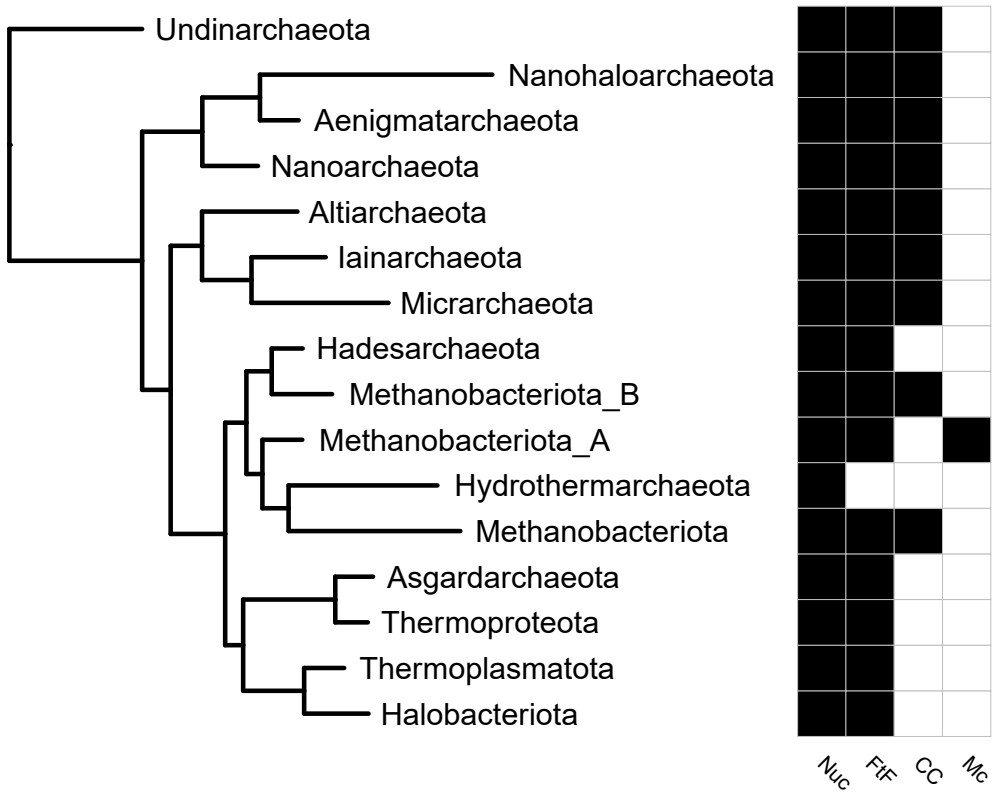


Figure 1: (a) The histone fold (PDB: 1a7w [8]). The N-terminal $\alpha 1$, central $\alpha 2$, and C-terminal $\alpha 3$ helices are coloured in blue, orange, and green respectively. Linkers L1 and L2 connect the $\alpha 1$ and $\alpha 3$ helices to the central $\alpha 2$ helix. (b) The histone dimer binds DNA at its $\alpha 1$ face (PDB: 5t5k [11]). (c) The eukaryotic H3-H4 tetramer (PDB: 1AOI [2]). (d) The eukaryotic octameric nucleosome (PDB: 1AOI [2]).


Bacterial
dimer
GTPase
Fold-to-fold
Beta
Poseidoniia
TM
ZZ
NCC
IHF
Phage
RdgC
Coiled-coil
DUF1931
Methanococcales
RdgC
Thermo-
plasmatota
Halo
TM
Nucleosomal

(a)



(b)

Figure 2: (a) Clustered sequence space of prokaryotic histones. Clustering was performed with CLANS. The colour of each line indicates the sequence similarity between the two sequences; sequences that are connected by darker lines are more similar than those connected by lighter lines. Clusters are coloured based on the histone category to which they belong as determined by the AlphaFold predictions. For a short description of each histone category, see supplementary table S1. (b) Cladogram of Archaea showing the distribution of nucleosomal (Nuc), fold-to-fold (FtF), coiled-coil (CC), and Methanococcales (Mc) histones across different phyla. The cladogram is based on GTDB version 207. For reference, Methanobacteriota A contains the Methanopyri and Methanococci classes; Methanobacteriota B contains the Thermococci and the Methanofastidiosa classes.

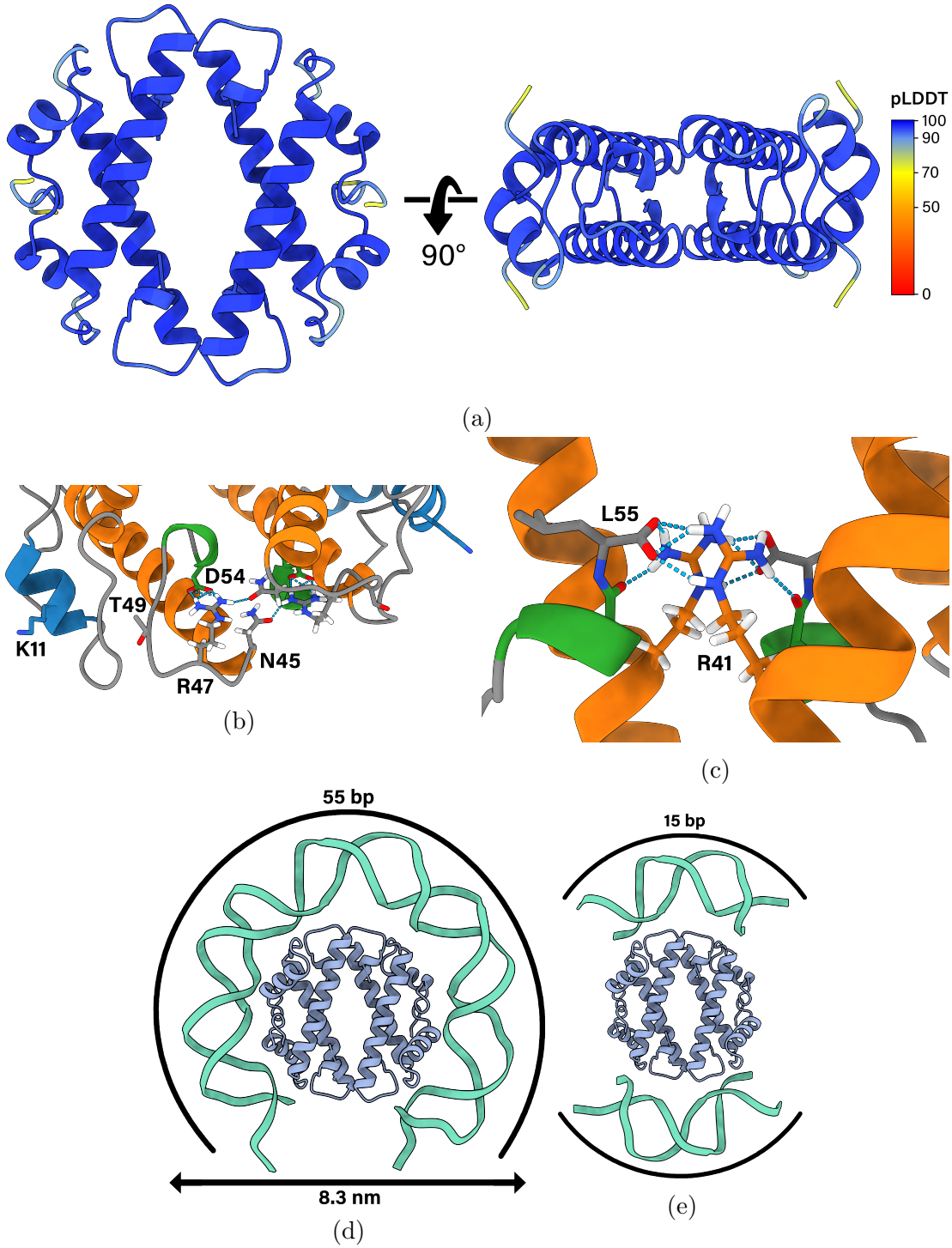


Figure 3: (a) The homotetramer of FtF histone D4GZE0 from Haloferax volcanii as predicted by AlphaFold2. Each residue is coloured by its pLDDT value. (b) The conserved RxTxxxxD motif, DNA binding residues K11, and tetramerisation residue N45 of FtF histone D4GZE0. K11, N54, R47, T49, and D54 relate to K16, N52, R54, T56, and D61 in the FtF HMM logo (Supplementary Fig. S5) (c) The 'back' of the dimer-dimer interface of FtF histone D4GZE0. R41 relates to R48 in the FtF HMM logo (Supplementary Fig. S5) (d,e) The DNA wrapping and bridging models for how FtF histones bind DNA.

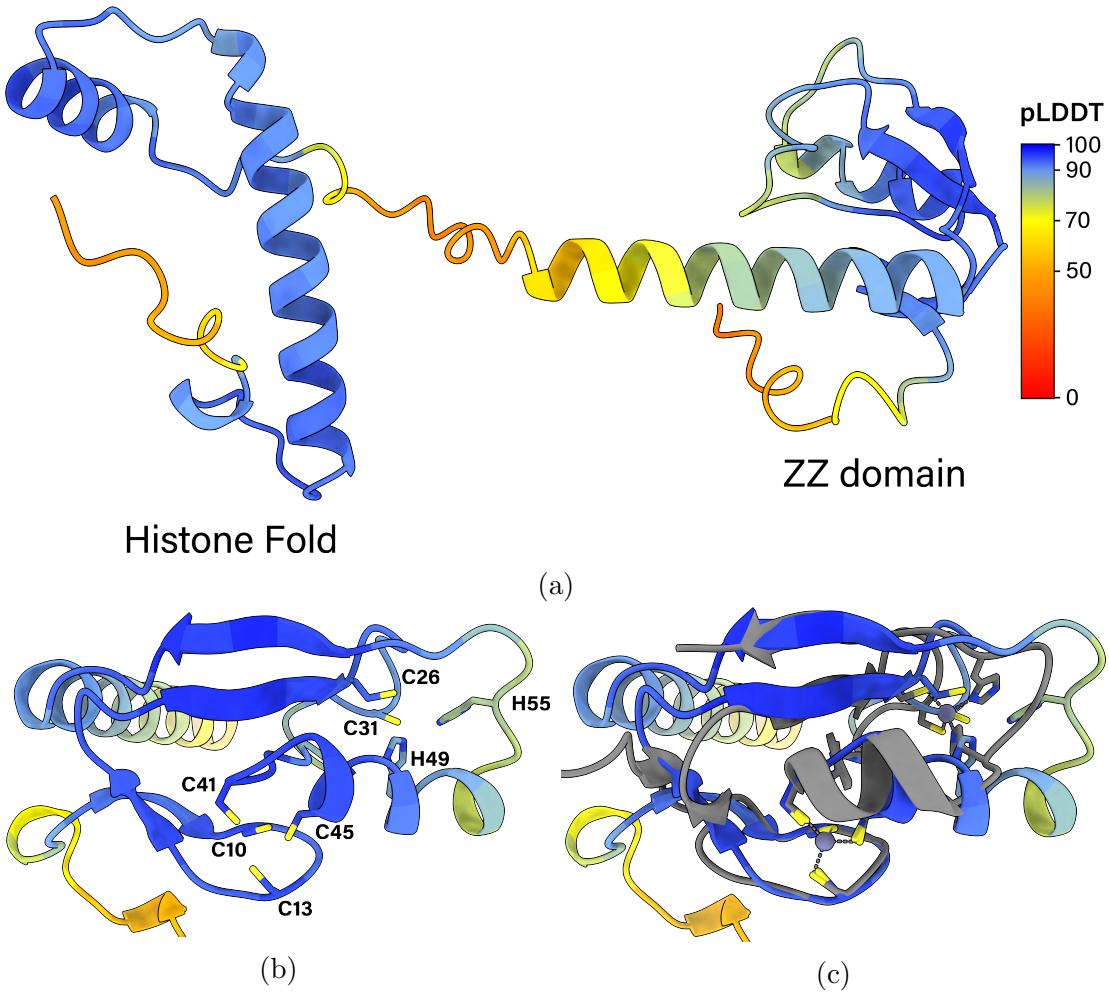
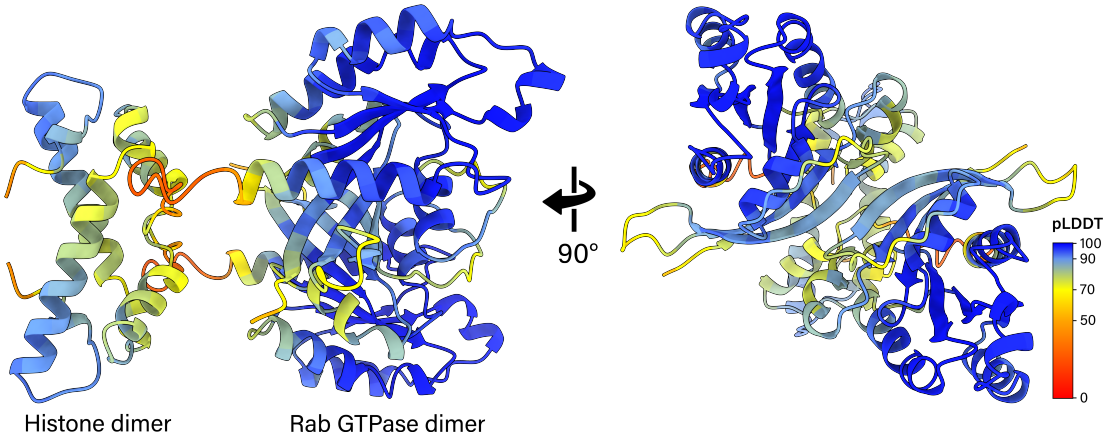
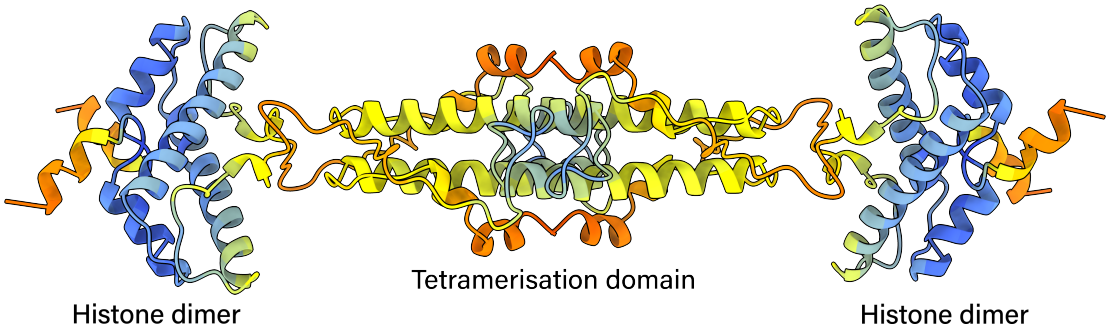


Figure 4: (a) The ZZ-histone D0LYE7 from *Haliangium ochraceum SMP-2* as predicted by AlphaFold2. ZZ-histones contain a ZZ-type zinc finger domain at the N-terminus and an $\alpha 3$ truncated histone fold at the C-terminus. Each residue is coloured by its pLDDT value. (b) The ZZ domain of D0LYE7. This domain contains two zinc binding motifs, one C4 motif (C20,C23,C51,C53) and a C2H2 motif (C36,C41,H59,H65). Each residue is coloured by its pLDDT value. (c) The ZZ domain of HERC2 (grey, PDB: 6ww4 [44]) aligned to D0LYE7.

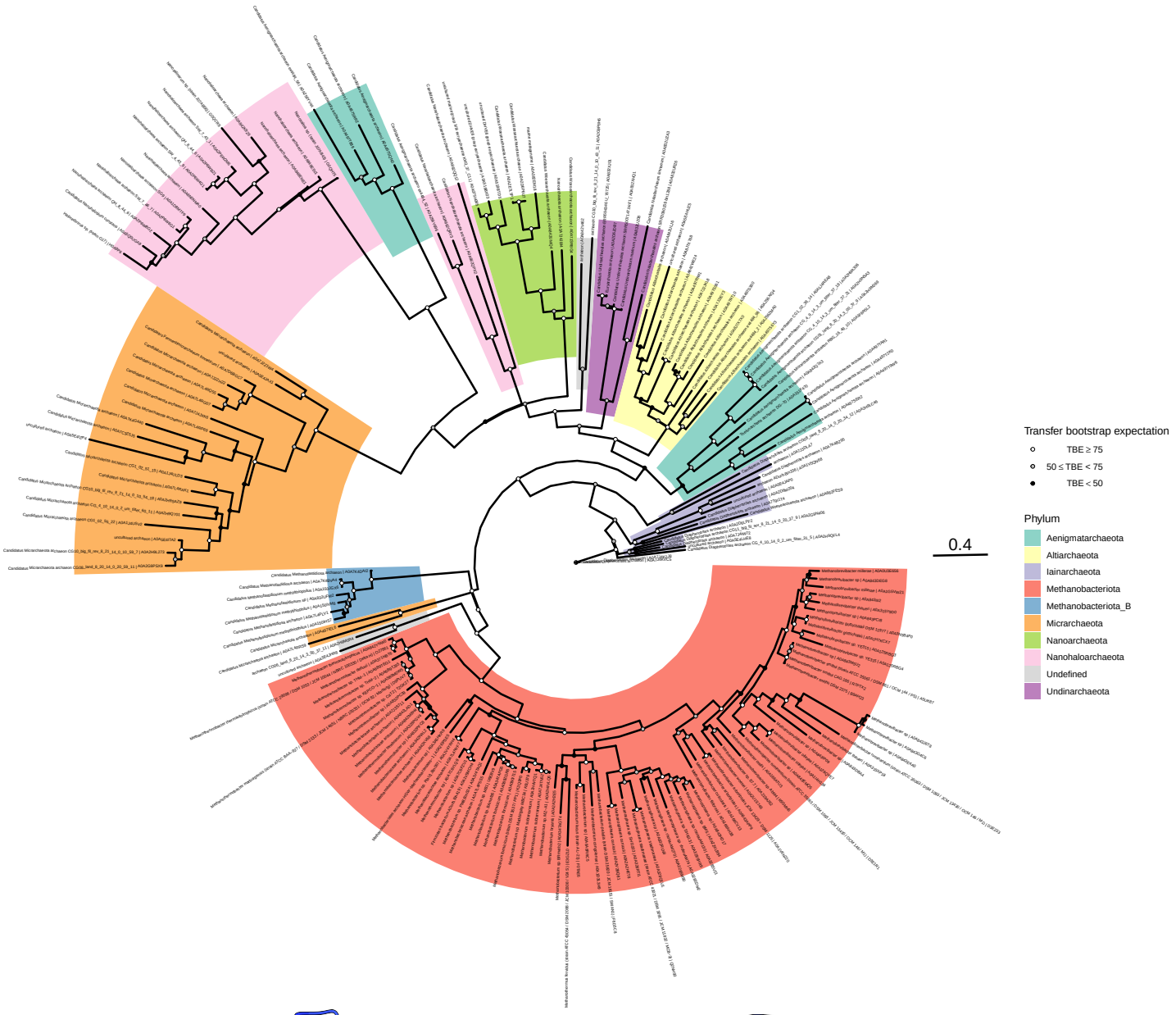


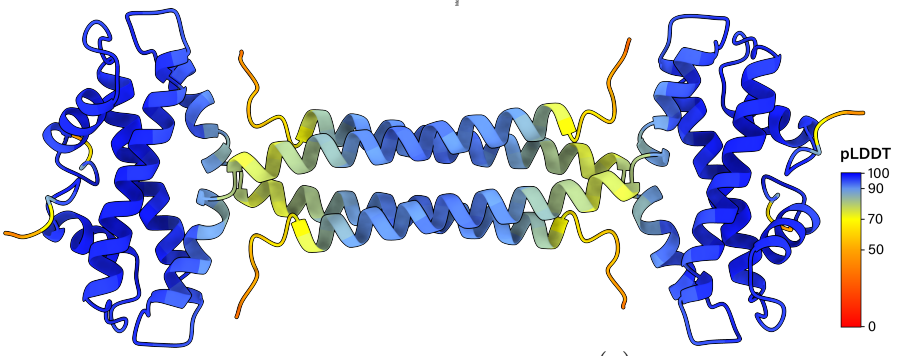
(a)



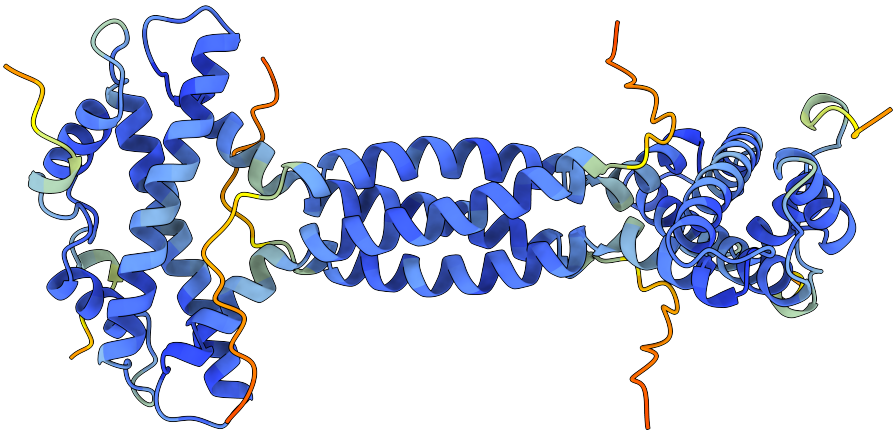
(b)

Figure 5: (a) The homodimer of Rab GTPase histone A0A0F8XJF6 as predicted by AlphaFold2. Rab GTPase histones contain a small Rab GTPase domain on the N-terminus and an FtF-like histone fold on the C-terminus. Each residue is coloured by its pLDDT value. (b) The homotetramer of phage histone A0A2E7QIQ9 as predicted by AlphaFold2. Phage histones contain an $\alpha 3$ truncated histone fold on the N-terminus and an α-helix which functions as the tetramerisation domain on the C-terminus.


Transfer bootstrap expectation
TBE ≥ 75
50 ≤ TBE < 75
TBE < 50
Phylum
Aenigmatarchaeota
Altiarchaeota
Iainarchaeota
Methanobacteriota
Methanobacteriota_B
Micrarchaeota
Nanoarchaeota
Nanohaloarchaeota
Undefined
Undinarchaeota
0.4


pLDDT
100
90
70
50
0

(b) (a)



(c)

Figure 6: (a) Phylogenetic tree of coiled-coil histones. Clades are coloured by phylum. The tree was generated with RAxML-NG. 1000 bootstraps were performed and used to calculate the transfer bootstrap expectation values (TBE). (b) The homotetramer of coiled-coil histone E3GZL0 from *Methanothermus fervidus* as predicted by AlphaFold2. Coiled-coil histones contain a long α-helix on the C-terminus. Each residue is coloured by its pLDDT value. (c) The homotetramer of RgdC histone D4GVY1 from *Haloferax volcanii* as predicted by AlphaFold2. RgdC histones form very similar tetramer structures to coiled-coil histones despite low sequence identity. Each residue is coloured by its pLDDT value.

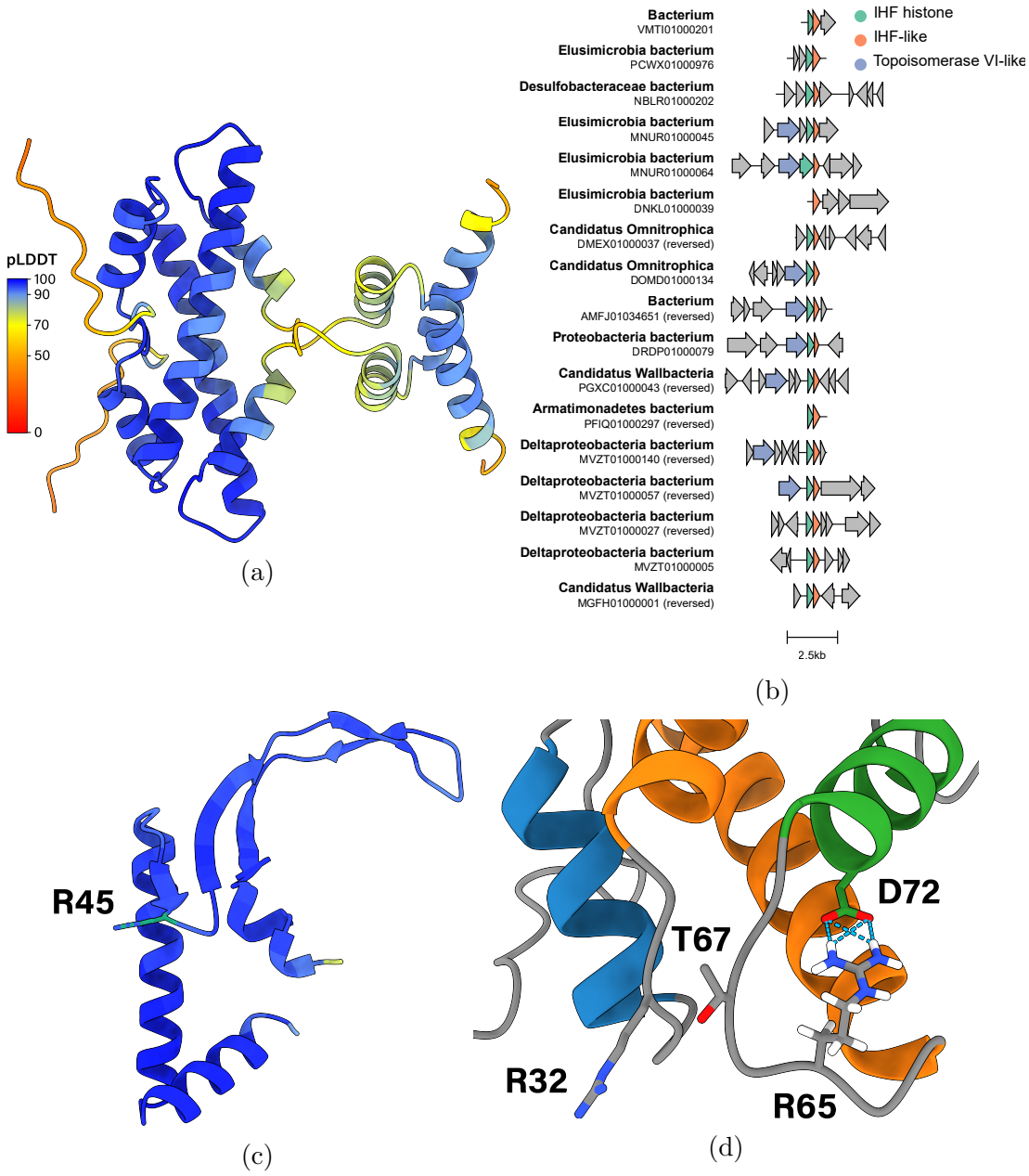


Figure 7: (a) The homodimer of IHF histone A0A358AGI2 from *Candidatus Omnitrophica* as predicted by AlphaFold2. Each residue is coloured by its pLDDT value. (b) Gene cluster comparison of bacterial metagenomes which contain the IHF histone. The organism and its genome ID are noted on the left. The IHF histone, IHF-like, and topoisomerase VI-like genes are coloured green, orange, and blue respectively. (c) The monomer of IHF-like A0A358AGI6 from *Candidatus Omnitrophica* as predicted by AlphaFold2. Each residue is coloured by its pLDDT value. Residue R45 is highlighted in green. (d) The RxTxxxxD motif and the "sprocket" R32 of IHF histone A0A358AGI2. R32, R65, T67, and D72 relate to R24, R57, T59, and D64 in the IHF histone HMM logo (Supplementary Fig. S29)