Evolutionary Biology

Coevolution of the CDCA7-HELLS ICF-related nucleosome remodeling complex and DNA methyltransferases

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
September 11, 2023 (this version)

Reviewed preprint version 1
June 5, 2023

Posted to bioRxiv
March 9, 2023

Sent for peer review
March 8, 2023

Hironori Funabiki ✉, **Isabel E. Wassing, Qingyuan Jia, Ji-Dung Luo, Thomas Carroll**

Laboratory of Chromosome and Cell Biology, The Rockefeller University, New York, NY 10065 • Bioinformatics Resource Center, The Rockefeller University, New York, NY 10065

Abstract

Summary

5-Methylcytosine (5mC) and DNA methyltransferases (DNMTs) are broadly conserved in eukaryotes but are also frequently lost during evolution. The mammalian SNF2 family ATPase HELLS and its plant ortholog DDM1 are critical for the maintenance of 5mC. Mutations in HELLS, its activator subunit CDCA7, and the *de novo* DNA methyltransferase DNMT3B, cause immunodeficiency-centromeric instability-facial anomalies (ICF) syndrome, a genetic disorder associated with the loss of DNA methylation. We here examine the coevolution of CDCA7, HELLS and DNMTs. While DNMT3, the maintenance DNA methyltransferase DNMT1, HELLS, and CDCA7 are all highly conserved in vertebrates and green plants, they are frequently co-lost in other evolutionary clades. The presence-absence patterns of these genes are not random; almost all CDCA7 harboring eukaryote species also have HELLS and DNMT1 (or another maintenance methyltransferase, DNMT5), whereas species that maintain DNMT1 or HELLS without CDCA7 are identified in several clades, such as Fungi and Ecdysozoa. Coevolution of presence-absence patterns (CoPAP) analysis in Ecdysozoa further indicates coevolutionary linkages among CDCA7, HELLS, DNMT1 and its activator UHRF1. We hypothesize that CDCA7 becomes dispensable in species that lost HELLS or DNA methylation, and/or the loss of CDCA7 triggers the replacement of DNA methylation by other chromatin regulation mechanisms. Our study suggests that a unique specialized role of CDCA7 in HELLS-dependent DNA methylation maintenance is broadly inherited from the last eukaryotic common ancestor.

eLife

eLife assessment

This **important** manuscript reveals signatures of co-evolution of two nucleosome remodeling factors, Lsh/HELLS and CDCA7, which are involved in the regulation of eukaryotic DNA methylation. The results suggest that the roles for the two factors in DNA methylation maintenance pathways can be traced back to the last eukaryotic common ancestor and that the CDC7A-HELLS-DNMT axis shaped the evolutionary retention of DNA methylation in eukaryotes. The **solid** evolutionary analyses form an interesting basis for experimental follow-up studies. The work should be of interest to colleagues in the fields of evolutionary biology, chromatin biology and genome biology.

Introduction

DNA methylation, particularly 5-methylcytosine (5mC) at CpG sequences, is widely conserved in eukaryotes. Along with its role in silencing transposable elements and suppressing aberrant intragenic transcription (Choi et al., 2020; Deniz et al., 2019; Neri et al., 2017), DNA methylation plays critical roles in developmental control, genome stability and the development of diseases such as cancers and immunodeficiencies (Greenberg and Bourc'his, 2019; Lyko, 2018; Nishiyama and Nakanishi, 2021; Robertson, 2005). Despite its versatility as an epigenetic control mechanism, DNA methyltransferases (DNMTs) are lost in multiple evolutionary lineages (Bewick et al., 2017b; Huff and Zilberman, 2014; Kyger et al., 2021; Ponger and Li, 2005; Zemach et al., 2010). While the evolutionary preservation and loss of DNMTs and other proteins involved in 5mC metabolism has been studied (Bewick et al., 2017b; de Mendoza et al., 2018; Dumesic et al., 2020; Engelhardt et al., 2022; Huff and Zilberman, 2014; Iyer et al., 2011; Lewis et al., 2020; Mondo et al., 2017; Mulholland et al., 2020; Nai et al., 2020; Tirot et al., 2021; Zemach et al., 2010), it remains unclear if there is any common process or event that leads to the loss of DNA methylation systems in certain evolutionary lineages. Since eukaryotic genomes are compacted by nucleosomes, DNA methylation systems must deal with this structural impediment (Felle et al., 2011). Could the emergence or loss of a specific nucleosome regulator affect the evolution of DNA methylation as an epigenetic mechanism?

DNMTs are largely subdivided into maintenance DNMTs and *de novo* DNMTs (Lyko, 2018; Ponger and Li, 2005). Maintenance DNMTs (directly or indirectly) recognize hemimethylated CpGs and restore symmetric methylation at these sites to prevent the passive loss of 5mC upon DNA replication. Conversely, methylation by *de novo* DNMTs does not require methylated DNA templates. In animals, 5mC is maintained during DNA replication by DNMT1 together with UHRF1, which directly recognizes hemimethylated cytosine via the SRA domain and stimulates activity of DNMT1 in a manner dependent on its ubiquitin-ligase activity (Nishiyama and Nakanishi, 2021). *De novo* DNA methylation is primarily carried out by DNMT3 in animals, while plants encode the closely related *de novo* methyltransferase DRM (Cao et al., 2000). Some species, such as the fungus *Cryptococcus neoformans*, lack *de novo* DNA methylation activity (Catania et al., 2020; Dumesic et al., 2020; Huff and Zilberman, 2014). In *C. neoformans*, DNA methylation is maintained by DNMT5, which has a SNF2-family ATPase domain that is critical for its methyltransferase activity (Dumesic et al., 2020; Huff and Zilberman, 2014). Although DNMT5 orthologs cannot be found in land plants and animals, their broad existence in Stramenopiles, Chlorophyta and Fungi suggests that DNMT5, perhaps together with DNMT1, coexisted in the last eukaryotic common ancestor (LECA) (Huff and Zilberman, 2014).

The loss of DNA methylation is a hallmark of immunodeficiency–centromeric instability–facial anomalies (ICF) syndrome, a rare genetic disorder which causes severe immune defects (Ehrlich, 2003; Ehrlich et al., 2006; Vukic and Daxinger, 2019). Activated lymphocytes of ICF patients display a characteristic cytogenetic abnormality at the juxtacentromeric heterochromatin of chromosome 1 and 16, where the satellite 2 repetitive element is highly enriched. In about 50 % of ICF patients, classified as ICF1, the disease is caused by mutations in the *de novo* DNA methyltransferase DNMT3B (Hansen et al., 1999; Okano et al., 1999). The rarer genotypes known as ICF2, ICF3 and ICF4, are caused by mutations in ZBTB24, CDCA7, and HELLS (Helicase, Lymphoid Specific, also known as LSH, Lymphoid-Specific Helicase), respectively (de Greef et al., 2011; Thijssen et al., 2015; Unoki, 2021). While hypomethylation at satellite II repeats is common to all ICF genotypes, ICF2-4 patient-derived cells, but not ICF1 patient cells, additionally exhibit hypomethylation at centromeric alpha-satellite repeats (Jiang et al., 2005; Thijssen et al., 2015; Velasco et al., 2018). Knock out of ICF genes in human HEK293 cells reproduces the DNA methylation profile observed in patient cells (Unoki et al., 2019). In mice, ZBTB24, HELLS and CDCA7, but not DNMT3B, are required for methylation at centromeric minor satellite repeats (Dennis et al., 2001; Hardikar et al., 2020; Ren et al., 2015; Thijssen et al., 2015). Therefore, although all ICF proteins promote DNA methylation, the ZBTB24-CDCA7-HELLS axis may target additional loci such as alpha satellites for DNA methylation in a DNMT3-independent manner. Indeed, the importance of HELLS in DNA methylation maintenance by DNMT1 has been reported (Han et al., 2020; Ming et al., 2020; Unoki, 2021; Unoki et al., 2020).

HELLS belongs to one of ~25 subclasses of the SNF2-like ATPase family (Flaus et al., 2006). Among these diverse SNF2 family proteins, HELLS appears to have a specialized role in DNA methylation. Reduced genomic DNA methylation was observed in HELLS (LSH) knockout mice (Dennis et al., 2001), transformed mouse fibroblasts (Dunican et al., 2013), mouse embryonic fibroblasts (Myant et al., 2011; Yu et al., 2014), and zebrafish and *Xenopus* embryos (Dunican et al., 2015). In fact, this function of HELLS in DNA methylation was originally inferred from studies in *Arabidopsis*, where mutations in the HELLS ortholog DDM1 (decrease in DNA methylation) cause drastic reduction of 5mC in transposable and repetitive elements (Miura et al., 2001; Vongs et al., 1993). Like HELLS, DDM1 is a SNF2 ATPase with demonstrable *in vitro* nucleosome remodeling activity (Brzeski et al., 2003; Jenness et al., 2018). Since DNA methylation defects in *ddm1* mutants can be rescued by the loss of histone H1, it has been proposed that DDM1-mediated remodeling of H1-bound nucleosomes is important for DNA methylation (Zemach et al., 2013).

CDCA7 (also known as JPO1) was originally identified as one of eight CDCA genes that exhibited cell division cycle-associated gene expression profiles (Walker, 2001). A putative 4CXXC zinc finger binding domain (zf-4CXXC_R1) is conserved among CDCA7 homologs, including its paralog CDCA7L (also known as JPO2 and R1) (Chen et al., 2005; Ou et al., 2006). Multiple lines of evidence support the idea that CDCA7 functions as a direct activator of the nucleosome remodeling enzyme HELLS. First, in *Xenopus* egg extracts, HELLS and CDCA7e (the sole CDCA7 paralog present in *Xenopus* eggs) both preferentially interact with nucleosomes rather than nucleosome-free DNA, and binding of HELLS to chromatin depends on CDCA7e (Jenness et al., 2018). Second, HELLS alone exhibits little nucleosome sliding activity, but CDCA7e greatly stimulates it (Jenness et al., 2018). Third, HELLS directly binds to CDCA7e (as well as CDCA7 and CDCA7L), even in the absence of DNA (Jenness et al., 2018). Fourth, HELLS and CDCA7 interact in human cells (Unoki et al., 2019), where chromatin binding of HELLS also depends on CDCA7 (Jenness et al., 2018). ICF disease mutations located in the conserved zf-4CXXC_R1 domain (R274C, R274H and R304H in human CDCA7) inhibited chromatin binding of CDCA7 and HELLS without interfering with CDCA7-HELLS interaction (Jenness et al., 2018), such that the zf-4CXXC_R1 domain likely serves as a chromatin binding module. Therefore, we proposed that CDCA7 and HELLS form a bipartite nucleosome remodeling complex, termed CHIRRC (CDCA7-HELLS ICF-related nucleosome remodeling complex) (Jenness et al., 2018).

We previously suggested that ZBTB24, CDCA7, and HELLS form a linear pathway to support DNA methylation (Jenness et al., 2018). ZBTB24 is a transcription factor which binds the promoter region of CDCA7 and is required for its expression (Wu et al., 2016). As CDCA7 binds HELLS to form the CHIRRC, we proposed that its ATP-dependent nucleosome sliding activity exposes DNA that was previously wrapped around the histone octamer and makes it accessible for DNA methylation (Jenness et al., 2018). Indeed, DNMT3A and DNMT3B cannot methylate DNA within a nucleosome (Felle et al., 2011), and the importance of HELLS and DDM1 for DNA methylation at nucleosomal DNA has been reported in mouse embryonic fibroblasts and *Arabidopsis*, respectively (Lyons and Zilberman, 2017). Given the frequent loss of 5mC as an epigenetic mark in multiple evolutionary lineages, it is striking that the role for HELLS/DDM1 in DNA methylation appears to be conserved in evolutionarily distant mammals and plants. Importantly, this would suggest that the promotion of DNA methylation through nucleosome sliding is specific to HELLS and cannot be substituted by other SNF2 family nucleosome remodelers, such as SNF2 (SMARCA2/4), INO80, and ISWI (SMARCA1/5). If the specific function of HELLS and CDCA7 in DNA methylation is indeed derived from the last eukaryotic common ancestor (LECA), we hypothesize that HELLS and CDCA7 coevolved with other DNA methylation machineries. We here test this hypothesis and discuss the potential sequence of events that led to the loss of DNA methylation in some species.

Results

CDCA7 is absent from the classic model organisms that lack genomic 5mC

CDCA7 is characterized by the unique zf-4CXXC_R1 domain (Pfam PF10497 or Conserved Domain Database [CDD] cl20401) (Lu et al., 2020; Mistry et al., 2021). Conducting a BLAST search using human CDCA7 as a query sequence against the Genbank protein database, we realized that no zf-4CXXC_R1-containing proteins are identified in the classic model organisms *Drosophila melanogaster*, *Caenorhabditis elegans*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, which are all also known to lack any DNMTs and genomic 5mC (Zemach et al., 2010). However, we identified a protein with a zf-4CXXC_R1 motif in the bumblebee *Bombus terrestris* and the thale cress *A. thaliana*, which have both maintenance and *de novo* DNMTs (Bewick et al., 2017b; Chan et al., 2005; Li et al., 2018; Zemach et al., 2010) (**Fig. 1**). Based on the reciprocal best hits (RBH) criterion using human HELLS as a query sequence (see Methods) (Ward and Moreno-Hagelsieb, 2014), HELLS orthologs were identified in *A. thaliana* (DDM1) and *S. cerevisiae* (Irc5), which is in line with previous reports (Litwin et al., 2017), as well as a putative HELLS ortholog in *B. terrestris*. No clear HELLS orthologs were identified in *D. melanogaster*, *C. elegans*, or *S. pombe*. This pilot-scale analysis based on the RBH criterion led us to hypothesize that the evolutionary maintenance of CDCA7 is linked to that of HELLS and DNMTs. In order to statistically validate this hypothesis, we set out to systematically define orthologs of CDCA7, HELLS and DNMTs in a broad range of evolutionary lineages.

CDCA7 family proteins in vertebrates

We first characterized evolutionary conservation of CDCA7 family proteins in vertebrates, where 5mC, DNMT1 and DNMT3 are highly conserved. A BLAST search against the Genbank protein database identified two zf-4CXXC_R1 domain-containing proteins, CDCA7/JPO1 and CDCA7L/R1/JPO2, throughout Gnathostomata (jawed vertebrates) (**Fig. 2A, B**). In frogs (such as *Xenopus*, but not all amphibians), and some fishes (such as *Astyanax mexicanus* and *Takifugu rubripes*), a third paralog CDCA7e exists (**Fig. 2A, B**, **Fig. S1**). CDCA7e is the only CDCA7-like protein that can be detected in *Xenopus* eggs (Jenness et al., 2018), and thus likely represents a form specific to oocytes and early embryos in these species. Among twelve conserved cysteine

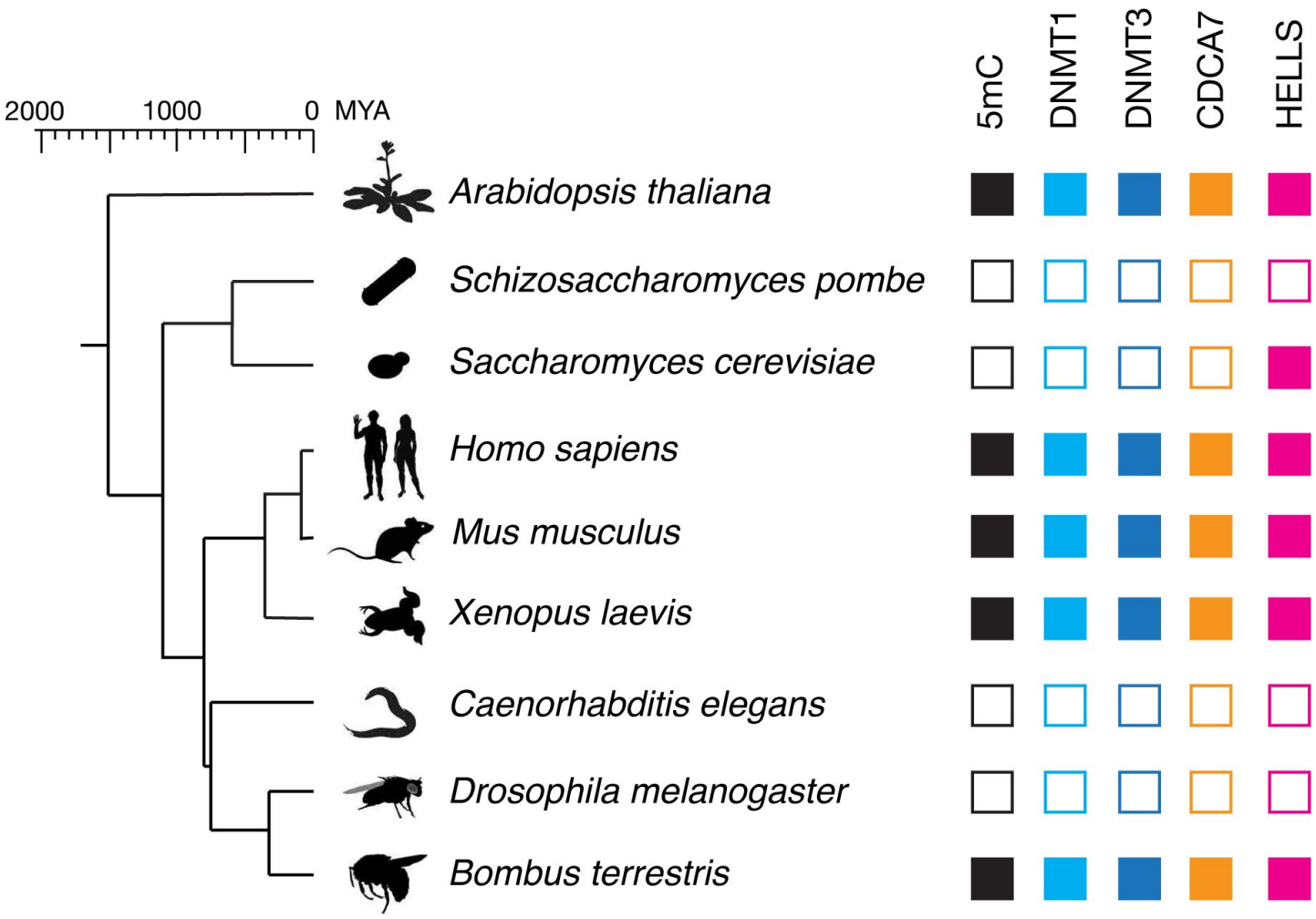

Figure 1.

CDCA7 is absent from model organisms with undetectable genomic 5mC

Filled squares and open squares indicate presence and absence of an orthologous protein(s), respectively. CDCA7 homologs are absent from model organisms where DNMT1, DNMT3 and 5mC on genomic are absent. See Table S1 for supporting evidence.

residues originally reported in the zf-4CXXC_R1 domain (Ou et al., 2006), the 12th cysteine residue is not conserved in *Rhincodon typus* (whale shark) CDCA7 and in *Xenopus laevis* CDCA7e. In general, the 12th cysteine residue of the zf-4CXXC_R1 domain is least conserved among CDCA7 homologs within and outside of vertebrates, such that we do not consider it a key component of the zf-4CXXC_R1 domain (see below). Considering that the jawless fish *Petromyzon marinus* (sea lamprey) and other invertebrates commonly possess only one CDCA7 family gene (**Fig. 2**, and see below), the CDCA7 paralogs may have emerged in jawed vertebrates. Overall, the zf-4CXXC_R1 sequence is highly conserved among vertebrate CDCA7 homologs, including three residues that are mutated in ICF3 patients (R274, G294 and R304 in human CDCA7) (**Fig. 2B**) (Thijssen et al., 2015). Note that these amino acid positions of the ICF3 mutations in human CDCA7 are based on the previously reported sequence of an NP_665809 (Thijssen et al., 2015), which is annotated as isoform 2 (371 amino acids), whereas we list the isoform 1 (NP_114148) with 450 amino acids in this study (**Fig. 2A**).

While the presence of four CXXC motifs in the zf-4CXXC_R1 domain is reminiscent of a classic zinc finger-CXXC domain (zf-CXXC, Pfam PF02008), which binds to nonmethylated CpG (Long et al., 2013), their cysteine arrangement is distinctly different, perhaps reflecting the capacity of zf-4CXXC_R1 domain to recognize nucleosomes (Jenness et al., 2018) and potentially specific epigenetic marks or DNA sequences. In vertebrate CDCA7 paralogs, eleven conserved cysteines are arranged as $\underline{\text{CXXC}}X_{10}CX_4CX_7\underline{\text{CXXC}}X_{19}\underline{\text{CXXC}}X_3C\underline{\text{CXXC}}$. In contrast, in the classic zf-CXXC domain eight cysteines are arranged as $\underline{\text{CXXCXXC}}X_{4\text{-}5}\underline{\text{CXXCXXC}}X_{8\text{-}14}CX_4C$ (Long et al., 2013). Apart from the zf-4CXXC_R1 domain, vertebrate CDCA7-like proteins often, but not always, contain one or two Lens epithelium-derived growth factor (LEDGF)-binding motif(s) (**Fig. 2A and C**), defined as ([E/D]XEXFXGF) (Tesina et al., 2015). It has been reported that human CDCA7L and CDCA7 both interact with c-Myc but apparently via different regions, a leucine zipper and a less-defined adjacent segment that overlaps with a bipartite nuclear localization signal, respectively (Gill et al., 2013; Huang et al., 2005). This leucine zipper sequence is highly conserved among vertebrate CDCA7 family proteins (**Fig. 2A and D**). In contrast to zf-CXXC domain-containing proteins such as KDM2A/B, DNMT1, MLL1/2, and TET1/3, the vertebrate CDCA7 proteins do not contain any predicted enzymatic domains (Huang et al., 2005; Maertens et al., 2006; Tesina et al., 2015).

Plant homologs of CDCA7

A BLAST sequence homology search identified three classes of zf-4CXXC_R1 domain-containing proteins in *Arabidopsis thaliana* (**Fig. 3**). Class I proteins have one zf-4CXXC_R1 domain with no other detectable domains listed in CDD or Pfam/InterPro (Paysan-Lafosse et al., 2023). Strikingly, all three ICF-associated residues (R274, G294 and R304 in human CDCA7) as well as all eleven characteristic cysteines of the zf-4CXXC_R1 domain are conserved, with the exception that the position of the fourth cysteine is shifted two residues toward the N terminus. We define the domain with the conservation of the signature cysteine residues and three ICF-associated residues as “class I zf-4CXXC_R1”, and proteins with class I zf-4CXXC_R1 as prototypical CDCA7 orthologs.

Class II proteins contain a zf-4CXXC_R1 domain, a DDT domain and a WHIM1 domain. These proteins were previously identified as DDR1-3 (Dong et al., 2013). DDT and WHIM1 domains are commonly found in proteins that interact with SNF2h/ISWI (Aravind and Iyer, 2012; Li et al., 2017; Yamada et al., 2011). Indeed, it was reported that *Arabidopsis* DDR1 and DDR3 interact with the ISWI orthologs CHR11 and CHR17 (Tan et al., 2020). Among the eleven cysteine residues in the zf-4CXXC_R1 domain of these proteins, the position of the fourth residue is shifted towards the C-terminus. The ICF-associated glycine residue (G294 in human CDCA7, mutated to valine in ICF3 patients) is replaced by isoleucine. We define the zf-4CXXC_R1 containing a substitution at this glycine residue as class II zf-4CXXC_R1.

Class III proteins are longer (~1000 amino acid) and contain an N-terminal zf-4CXXC_R1 domain and a C-terminal JmjC domain (Pfam, PF02373), which is predicted to possess demethylase activity against histone H3K9me2/3 (Saze et al., 2008). While all eleven cysteine residues can be

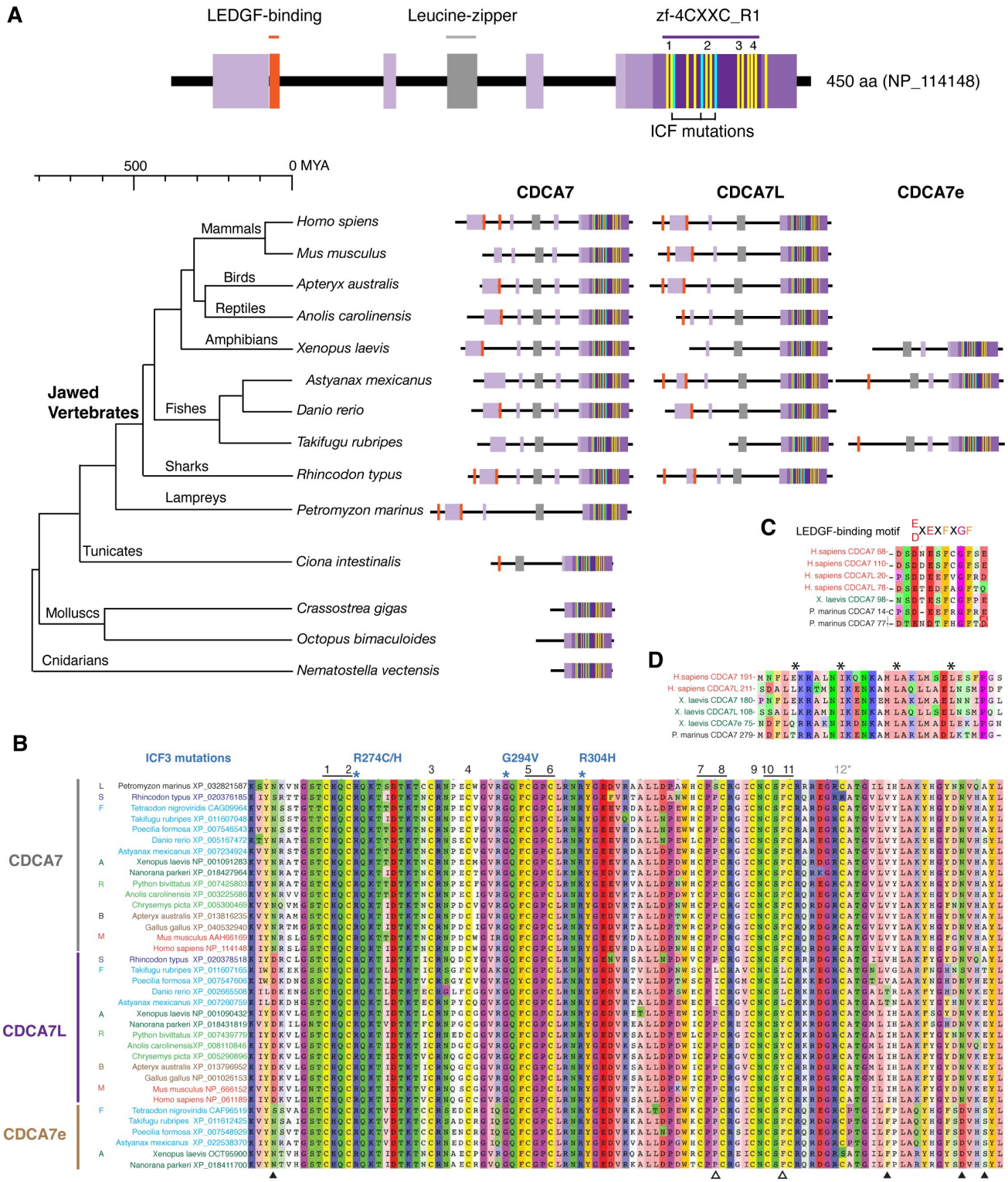

Figure 2.

CDCA7 paralogs in vertebrates

A. Schematics of vertebrate CDCA7 primary sequence composition, based on NP_114148. Yellow lines and light blue lines indicate positions of evolutionary conserved cysteine residues and residues that are mutated in ICF patients, respectively. B. Sequence alignment of the zf-4CXXC_R1 domain of vertebrate CDCA7-family proteins. White arrowheads; amino residues unique in fish CDCA7L. Black arrowheads; residues that distinguish CDCA7L and CDCA7e from CDCA7. C. Sequence alignment of LEDGF-binding motifs. D. Sequence alignment of the conserved leucine-zipper.

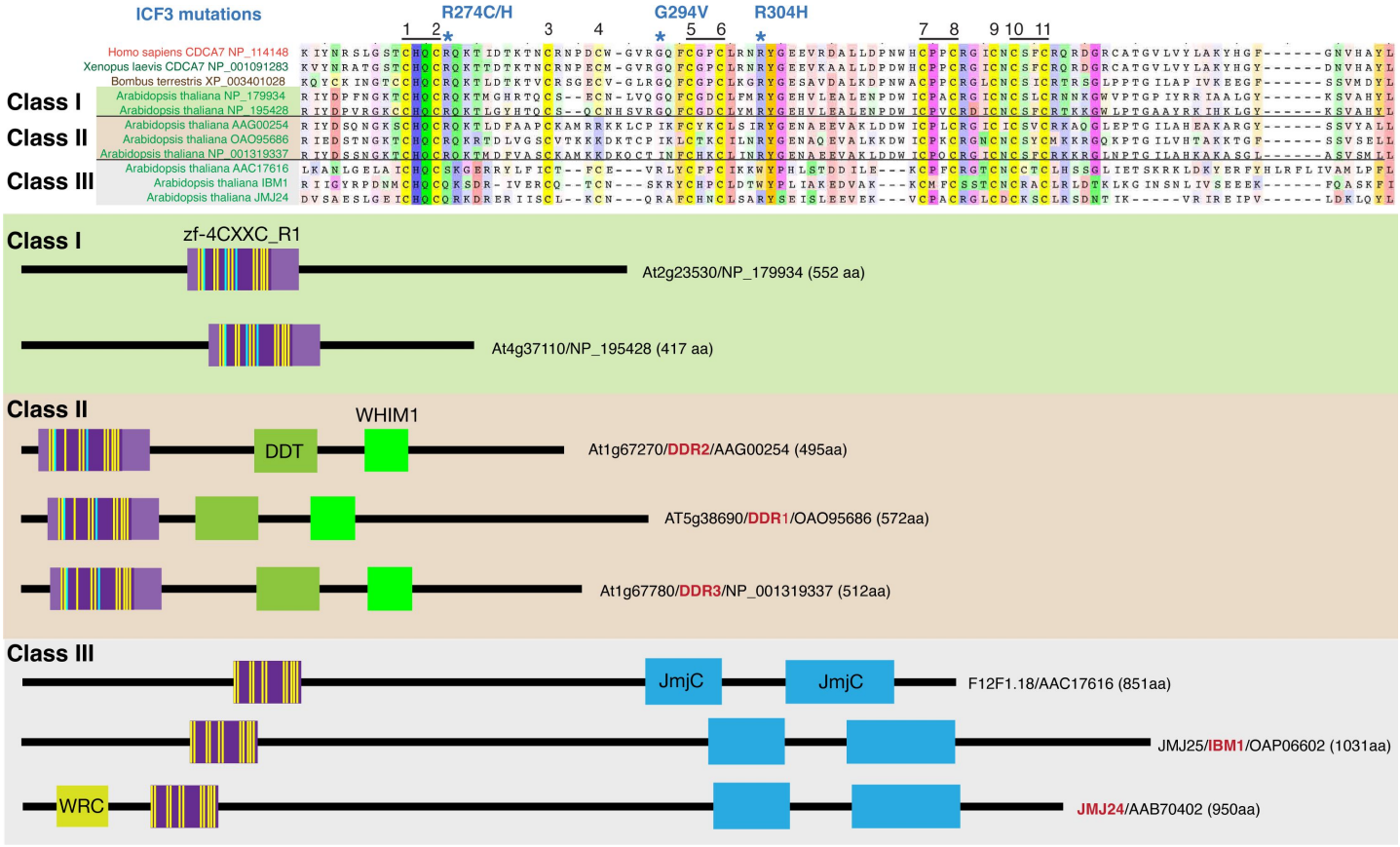

Figure 3.

CDCA7 homologs and other zf-4CXXC_R1-containing proteins in Arabidopsis

Top; alignments of the zf-4CXXC_R1 domain found in Arabidopsis thaliana. Bottom; domain structure of the three classes of zf-4CXXC_R1-containing proteins in Arabidopsis.

identified, there are deletions between the 4th and 5th cysteine and 6th and 7th cysteine residues. None of the ICF-associated residues are conserved in the class III. One of these class III proteins is IBM1 (increase in bonsai mutation 1), whose mutation causes the dwarf "bonsai" phenotype (Saze et al., 2008), which is accompanied with increased H3K9me2 and DNA methylation levels at the *BONSAI* (*APC13*) locus. Double mutants of *ddm1* and *ibm1* exacerbate the bonsai phenotype, indicating that DDM1 and IBM1 act independently to regulate DNA methylation (Saze et al., 2008). Another class III protein is JMJ24, which harbors a RING finger domain in addition to the 4CXXC and JmjC domains. This RING finger domain promotes ubiquitin-mediated degradation of the DNA methyltransferase CMT3, and thus opposes DNA methylation (Deng et al., 2016).

Orthologs of these three classes of CDCA7 proteins found in *Arabidopsis* are widely identified in green plants (Viridiplantae), including Streptophyta (e.g., rice, maize, moss, fern) and Chlorophyta (green algae). Other variants of zf-4CXXC_R1 are also found in Viridiplantae. In contrast to green plants, in which the combined presence of HELLS/DDM1-, CDCA7- and DNMT-orthologs is broadly conserved, no zf-4CXXC_R1-containing proteins can be identified in red algae (Rhodophyta) (Table S1, see **Fig.5**, **Fig. S2**).

Zf-4CXXC_R1-containing proteins in Fungi

Although *S. pombe* and *S. cerevisiae* genomes do not encode any CDCA7-like proteins, a BLAST search identified various fungal protein(s) with a zf-4CXXC_R1 domain. Among the zf-4CXXC_R1-containing proteins in fungi, 10 species (*Kwoniella mangroviensis*, *Coprinopsis cinere*, *Agaricus bisporus*, *Taphrina deformans*, *Gonapodya prolifera*, *Basidiobolus meristosporus*, *Coemansia reversa*, *Linderina pennispora*, *Rhizophagus irregularis*, *Podila verticillate***)** harbor a class II zf-4CXXC_R1 domain with two notable deviations (**Fig. 4A**). First, the space between the third and fourth cysteine residues is variable. Second, the fifth cysteine is replaced by aspartate in Zoopagomycota (*Basidiobolus meristosporus* ORX82853, *Coemansia reversa* PIA14937, *Linderina pennispora* ORX71196) (**Fig. 4A and B**). As these proteins do not contain any additional CDD-annotated domain like vertebrate and plant CDCA7 orthologs (except for *Taphrina deformans* CCG81269, which also has a CHROMO domain), we define this fungal protein family as CDCA7F, which forms a distinct clade in our phylogenetic analysis of zf-4CXXC_R1 sequence alignment (**Fig. S1**).

Besides CDCA7F, several fungal species encode a protein with a diverged zf-4CXXC_R1 domain, including those with a JmjC domain at the N-terminus (**Figure 4B**, Table S1). This composition mimics the plant class III proteins, for which the JmjC domain is located at the C-terminus. Among these proteins, it was suggested that *Neurospora crassa* DMM-1 does not directly regulate DNA methylation or demethylation but rather controls the deposition of histone H2A.Z and/or H3K56 acetylation, which inhibit spreading of heterochromatin segments with methylated DNA and H3K9me3 (Honda et al., 2010; Zhang et al., 2022). Thus, the known roles of these divergent variants of zf-4CXXC_R1 are not directly associated with DNA methylation.

Systematic identification of CDCA7 and HELLS homologs in eukaryotes

To systematically identify CDCA7 and HELLS orthologs in the major eukaryotic supergroups, we conducted a BLAST search against the NCBI protein database using human CDCA7 and HELLS protein sequences. To omit species with a high risk of false negative identification, we selected species containing at least 6 distinct proteins with compelling homology to the SNF2 ATPase domain of HELLS, based on the assumption that each eukaryotic species is expected to have 6-20 SNF2 family ATPases (Flaus et al., 2006). Indeed, even the microsporidial pathogen *Encephalitozoon cuniculi*, whose genome size is a mere 2.9 Mb, contains six SNF2 family ATPases (Flaus et al., 2006). As such, we generated a panel of 180 species encompassing all major

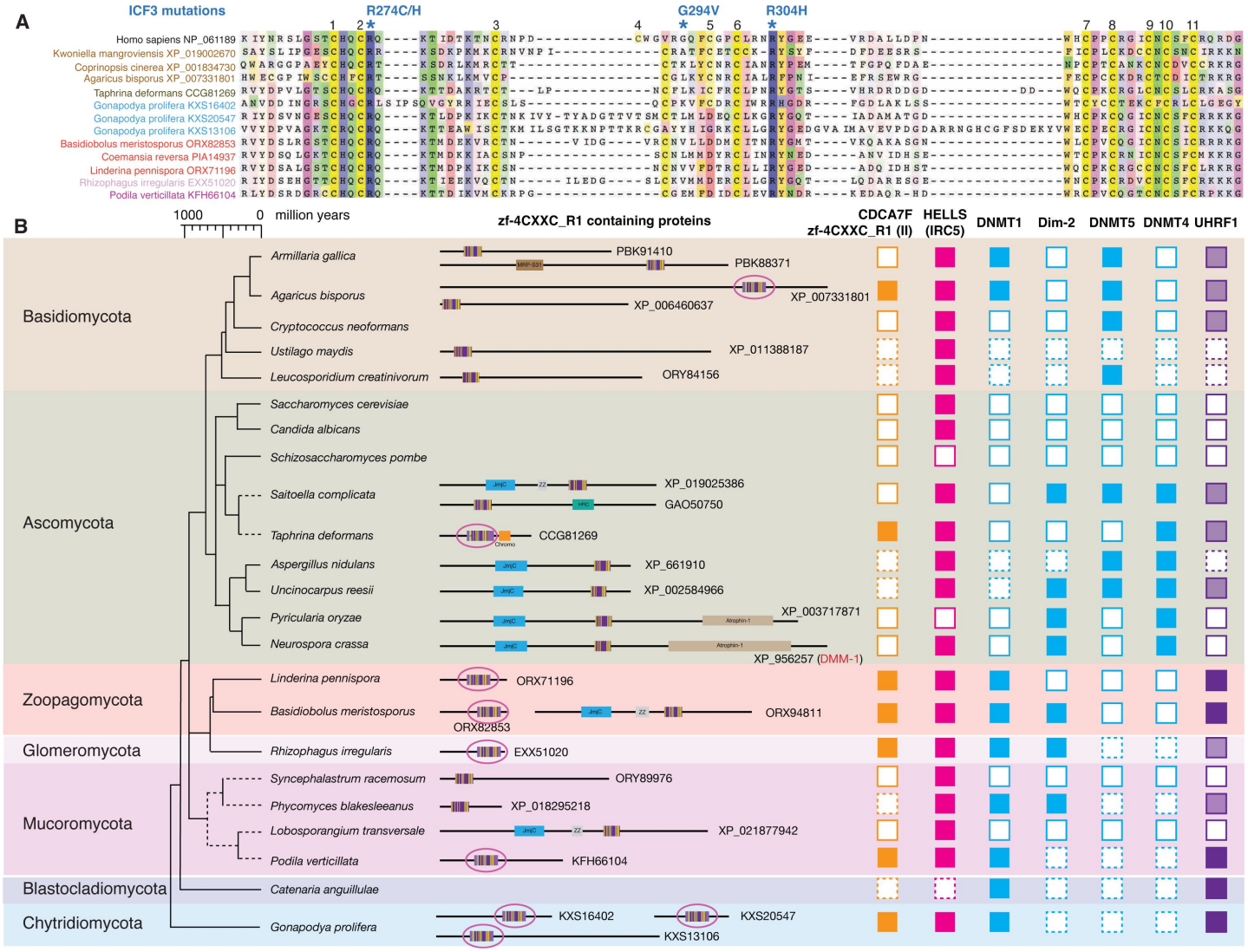

Figure 4.

Evolutionary conservation of CDCA7F, HELLS and DNMTs in fungi

A. Sequence alignment of fungi-specific CDCA7F with class II zf-4CXXC_R1 sequences. B. Domain architectures of zf-4CXXC_R1-containg proteins in fungi. The class II zf-4CXXC_R1 domain is indicated with purple circles. Squares with dotted lines indicate preliminary genome assemblies. Opaque boxes of UHRF1 indicate homologs that harbor the SRA domain but not the RING-finger domain.

eukaryote supergroups (5 Excavata, 18 SAR [2 Rhizaria, 6 Alveolata, 10 Stramenopiles]), 1 Haptista, 1 Cryptista, 15 Archaeplastida [3 Rhodophyta and 12 Viridiplantae], 4 Amoebozoa, 136 Opisthokonts [34 Fungi, 3 Holozoa, and 99 Metazoa]) (**Fig. S2**, Table S1).

HELLS orthologs were initially identified if they satisfied the RBH criterion. To further validate the annotation of HELLS orthologs, a phylogenetic tree was constructed from a multiple sequence alignment of the putative HELLS orthologs alongside other SNF2-family proteins of *H. sapiens*, *D. melanogaster*, *S. cerevisiae*, and *A. thaliana*. If HELLS orthologs are correctly identified (i.e. without erroneously including orthologs of another SNF2-subfamily) they should cluster together in a single clade. However, the sequence alignment using the full-length protein sequence failed to cluster HELLS and DDM1 in the same clade (**Fig. S3**). Since HELLS and other SNF2-family proteins have variable insertions within the SNF2 ATPase domain, multiple sequence alignment of the SNF2 domains was then conducted after removing the insertion regions, as previously reported (Flaus et al., 2006). By this SNF2 domain-only alignment method, all HELLS orthologs formed a clade, separated from CHD1, ISWI, SMARCA2/4, SRCAP, and INO80 (**Fig. S4**). An independent phylogenetic tree construction with a maximum-likelihood based method and 1000 bootstrap replicates confirmed these assignments (see Methods).

A BLAST search with the human CDCA7 sequence across the panel of 180 species identified a variety of proteins containing the zf-4CXXC_R1 domain, which is prevalent in all major supergroups (**Fig. 5**, **Fig. S1**, Table S1). Each of these identified proteins contains only one zf-4CXXC_R1 domain. The resulting list of CDCA7 BLAST hits were further classified as prototypical CDCA7 orthologs if they preserve the criteria of the class I zf-4CXXC_R1 (signature eleven cysteine residues *and* the three ICF-associated residues) (**Fig. S5**). A phylogenetic tree analysis of zf-4CXXC_R1 domains from diverse species confirmed that these CDCA7 orthologs with the class I zf-4CXXC_R1 domain are clustered under the same clade (**Fig. S1**). CDCA7 orthologs are broadly found in the three supergroup lineages (Archaeplastida, Amoebozoa, Opisthokonta) (**Fig.5**, **Fig. S1**, **Fig. S2**, **Fig. S5**, Table S1). In Excavata, the amoeboflagellate *Naegleria gruberi* encodes a protein that is a likely ortholog of CDCA7 with an apparent C-terminal truncation (XP_002678720), possibly due to a sequencing error (**Fig. S5**). In SAR, CDCA7 orthologs are absent from all available genomes, except stramenopile *Tribonema minus*, which encodes a distantly related likely ortholog of CDCA7 (KAG5177154). These conservations suggest that the class I zf-4CXXC_R1 domain in CDCA7 was inherited from the LECA.

In addition to the class I zf-4CXXC_R1 domain, we also identified divergent zf-4CXXC_R1 domains across eukaryotes, although metazoan species only contain CDCA7 orthologs (and their paralogs such as CDCA7L and CDCA7e) with the exception of the sponge *Amphimedon queenslandica*, which also encodes a protein (XP_019849176) with the class II zf-4CXXC_R1 domain, DDT, and WHIM1 domain, the composition of which is similar to plant DDR1-3 (**Fig. 3**, **Fig. 5**, **Fig. S2**, **Fig. S5**, and Table S1). The Amoebozoa *Acanthamoeba castellanii* encodes a protein (XP_004340890) containing the class II zf-4CXXC_R1 domain as well as the CHROMO domain (cd00024) (**Fig. S5**). This domain combination is also found in the *Taphrina deformans* CDCA7F, CCG81269 (**Fig. 4**). Proteins with a combinatory presence of a diverged zf-4CXXC_R1 domain and the JmjC domain are found in plants, fungi, Amoebozoa, and *Naegleria gruberi* (Table S1).

Despite the prevalence of the zf-4CXXC_R1 domain and its variants in eukaryotes, no zf-4CXXC_R1 domain was found in prokaryotes and Archaea. This is in contrast to SNF2 family proteins and DNA methyltransferases, which can be identified in prokaryotes and Archaea (Colot and Rossignol, 1999; Flaus et al., 2006; Huff and Zilberman, 2014; Ponger and Li, 2005), pointing toward a possibility that the zf-4CXXC_R1 domain emerged to deal with unique requirement of eukaryotic chromatin.

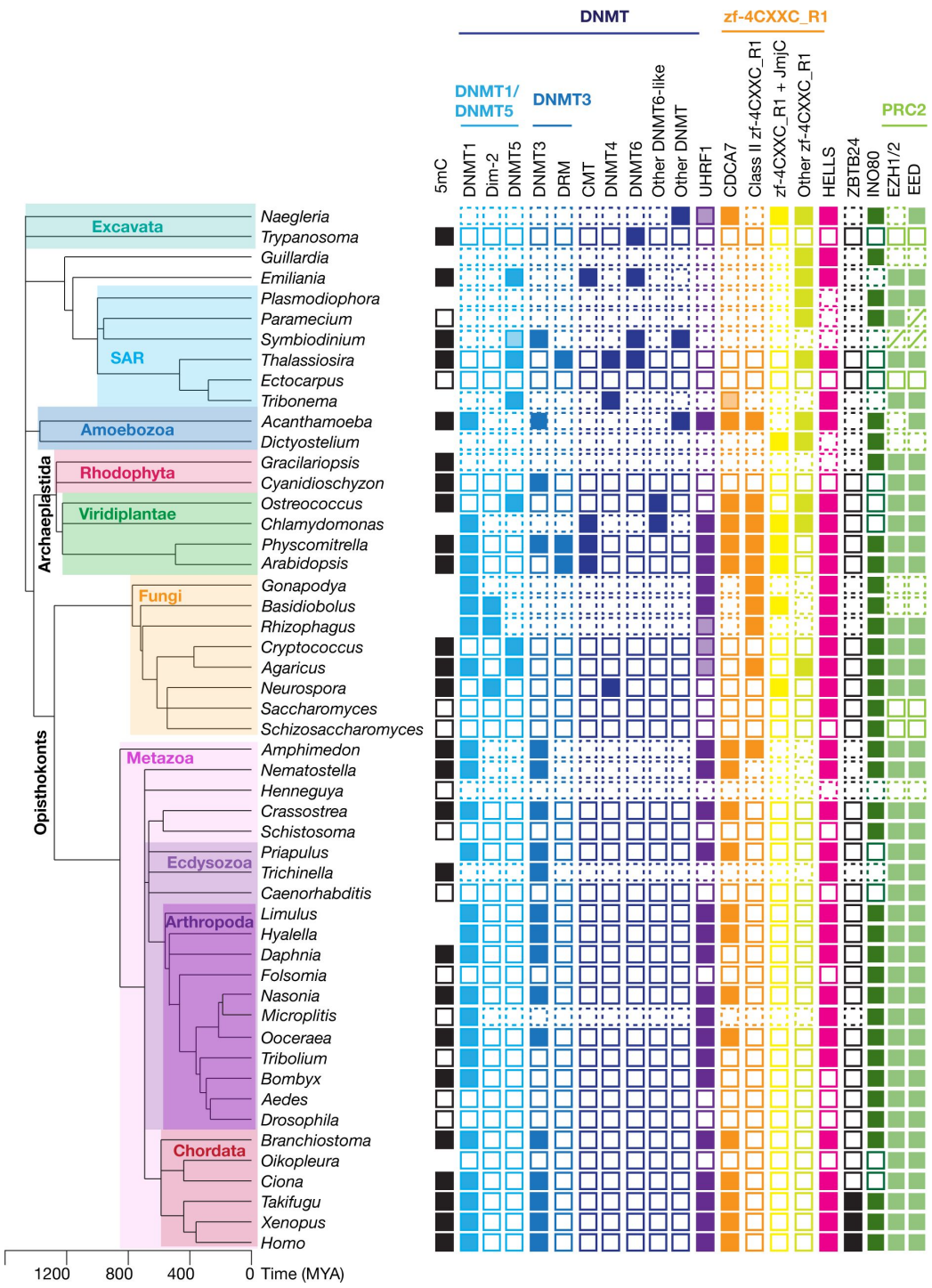

Figure 5.

Evolutionary conservation of CDCA7, HELLS and DNMTs

The phylogenetic tree was generated based on timetree.org. Filled squares and open squares indicate presence and absence of an orthologous protein(s), respectively. Squares with dotted lines imply preliminary-level genome assemblies. Squares with a diagonal line; *Paramecium* EED was functionally identified (Miro-Pina et al., 2022), but not by the sequence-based search in this study; homologs of EZH1/2 and EED were identified in *Symbiodinium sp. KB8* but not in *Symbiodinium microadriaticum* (Table S1). An opaque box of DNMT5 in *Symbiodinium* indicates a homolog that does not contain the ATPase domain, which is commonly found in DNMT5 family proteins. Opaque boxes of UHRF1 indicate homologs that harbor the SRA domain but not the RING-finger domain. Full set of analysis on the panel of 180 eukaryote species is shown in **Figure S1** and Table S1. Genbank accession numbers of each protein and PMID numbers of published papers that report presence or absence of 5mC are reported in Table S1.

Classification of DNMTs in eukaryotes

A simple RBH approach is not practical to classify eukaryotic DNMT proteins due to the presence of diverse lineage-specific DNMTs (Huff and Zilberman, 2014). Therefore, we collected proteins with a DNMT domain within the panel of 180 eukaryote species, and then extracted the DNMT domains from each sequence (based on an NCBI conserved domains, Dcm [COG0270], Dcm super family [cl43082], AdoMet_MTases superfamily [cl17173]). Generating a phylogenetic tree based on the multisequence alignment of the DNMT domains, we were able to classify the majority of all identified DNMTs as previously characterized DNMT subtypes according to their sequence similarity. These DNMT subtypes include DNMT1; DNMT3; the plant specific *de novo* DNA methyltransferases DRM1-3; the "true" plant DNMT3 orthologs (Yaari et al., 2019); the plant specific CMT (Bewick et al., 2017a); the fungi-specific maintenance methyltransferase Dim-2 and *de novo* methyltransferase DNMT4 (Bewick et al., 2019a; Nai et al., 2020); the SNF2 domain-containing maintenance methyltransferase DNMT5 (Dumesic et al., 2020; Huff and Zilberman, 2014); DNMT6 (a poorly characterized putative DNMT identified in Stramenopiles, Haptista and Chlorophyta) (Huff and Zilberman, 2014), and the tRNA methyltransferase TRDMT1 (also known as DNMT2) (**Fig. S6**). We also identified other DNMTs, which did not cluster into these classes. For example, although it has been reported that DNMT6 is identified in *Micromonas* but not in other Chlorophyta species such as *Bathycoccus* and *Ostreococcus* (Huff and Zilberman, 2014), we identified Chlorophyta-specific DNMTs that form a distinct clade, which seems to be diverged from DNMT6 and DNMT3. We temporarily called this class Chlorophyta DNMT6-like (**Fig. S6**). Other orphan DNMTs include the *de novo* DNA methyltransferase DNMTX in fungus *Kwoniella mangroviensis* (Catania et al., 2020), and an uncharacterized DNMT in *N. gruberi* (XP_002682263) (**Fig. S6**).

Coevolution of CDCA7, HELLS and DNMTs

The identification of protein orthologs across the panel of 180 eukaryotic species reveals that homologs of CDCA7, HELLS and DNMTs are conserved across the major eukaryote supergroups, but they are also dynamically lost (**Fig. 5**, **Fig. S2**). We found 40 species encompassing Excavata, SAR, Amoebozoa, and Opisthokonta that lack CDCA7 (or CDCA7F), HELLS and DNMT1. Species that encode the set of DNMT1, UHRF1, CDCA7 and HELLS are particularly enriched in Viridiplantae and Metazoa (**Fig. 5**, **Fig. S2**, Table S1). A clear exception is *Acanthamoeba castellanii*, whose genome also encodes these four genes and is reported to have methylated cytosines (Moon et al., 2017) (**Fig. 5**, Table S1, **Fig. S2**). *N. gruberi* is an exceptional example among Excavata species, encoding a CDCA7 ortholog (XP_002678720) as well as HELLS; an orphan DNMT protein (XP_002682263), and a UHRF1-like protein (**Fig. 5**, **Fig. S2**, Table S1). DNMT1, DNMT3, HELLS and CDCA7 seem to be absent in other Excavata lineages, although Euglenozoa variants of DNMT6 and cytosine methylation are identified in *Trypanosoma brucei* and *Leishmania major* (Huff and Zilberman, 2014; Militello et al., 2008). The yellow-green algae *Tribonema minus* is the only SAR species that encodes a CDCA7 ortholog (KAG5177154), along with DNMT5 and HELLS.

Among the panel of 180 eukaryote species, we found 82 species that encode CDCA7 (including fungal CDCA7F) (**Fig. S2**, Table S1). Strikingly, all 82 species containing CDCA7 (or CDCA7F) also harbor HELLS. Almost all CDCA7 encoding species also possess DNMT1. Exceptions are: the yellow-green algae *T. minus* and the Mamiellophyceae lineage of the green algae/Chlorophyta (*Bathycoccus prasinos*, *Ostreococcus lucimarinus*, *Micromonas pusilla*), which lost DNMT1 but possess DNMT5; the amoeboflagellate *N. gruberi*, which encodes an orphan DNMT and UHRF1, and the fungus *T. deformans*, which encodes UHRF1 and DNMT4. In contrast, 20 species (e.g., *S. cerevisiae*) possess only HELLS, while 10 species (e.g. the silk moth *Bombyx mori*) retains only DNMT1 among the set of DNMT1, CDCA7 and HELLS. These observations indicate that the function of CDCA7 is strongly linked to HELLS and DNMT1, such that the presence of CDCA7 depends on HELLS and DNMT1, while CDCA7 is easier to lose than DNMT1 and HELLS. Compared to the apparent HELLS/DNMT1-dependent existence of CDCA7, DNMT3 seems to be more dispensable;

besides Fungi, which lack DNMT3, 16 species (e.g., the bed bug *Cimex lectularius* and the red paper wasp *Polistes canadensis*) possess the set of CDCA7, HELLS and DNMT1 but not DNMT3/DNMT3-like proteins (**Fig. S2** and Table S2).

To quantitatively assess coevolution of DNMTs, CDCA7 and HELLS, we performed CoPAP analysis on the panel of 180 eukaryote species (**Fig. S7**)(Cohen et al., 2013). The analysis was complicated due to the lineage-specific diverse DNMT classes (e.g., Dim2, DNMT5, DNMT6 and other plant specific DNMT variants) and divergent variants of zf-4CXXC_R1. Considering this caveat, we conducted CoPAP analysis of five DNMTs (DNMT1, Dim-2, DNMT3, DNMT5, DNMT6), UHRF1, HELLS, CDCA7, proteins with class II zf-4CXXC_R1, and proteins with zf-4CXXC_R1 and JmjC. As fungi-specific CDCA7F contains class II zf-4CXXC_R1, and all the other class II zf-4CXXC_R1 containing proteins were identified in species that also possess CDCA7, we conducted CoPAP against two separate lists; in the first list (**Fig. S7A**) CDCA7F was included in the CDCA7 category (i.e. considered as a prototypical CDCA7 ortholog in fungi), whereas in the second list (**Fig. S7B**) CDCA7F was included in the class II zf-4CXXC_R1 category. As positive and negative controls for the CoPAP analysis, we also included subunits of the PRC2 complex (EZH1/2, EED and Suz12), and other SNF2 family proteins SMARCA2/SMARCA4, INO80 and RAD54L, which have no direct role related to DNA methylation, respectively. As expected for proteins that act in concert within the same biological pathway, both CoPAP results showed significant coevolution between DNMT1 and UHRF1, as well as between the PRC2 subunits EZH1 and EED. Suz12 did not show a significant linkage to other PRC2 subunits by this analysis, most likely due to a failure in identifying diverged Suz12 orthologs, such as those in *Neurospora* and *Paramecium* (Jamieson et al., 2013; Miro-Pina et al., 2022). The evolutionary linkage between DNMT5 and DNMT6 were also observed (**Fig. 6**, **Fig. S2**, Table S1). In the result of CoPAP analysis against the first list, where CDCA7F was included in the CDCA7 category, CDCA7 exhibited a coevolutionary linkage to DNMT1 (**Fig. S7A**). In the second CoPAP, where CDCA7F was included in the class II zf-4CXXC_R1 category, CDCA7 exhibited linkage to UHRF1 and HELLS (**Fig. S7B**). Notably, none of these proteins show an evolutionary association with the PRC2 proteins or other SNF2 family proteins, while the coevolutionary linkage between CDCA7 and the DNMT1-UHRF1 cluster was reproducible.

We next conducted the CoPAP analysis against a panel of 50 Ecdysozoa species, where DNA methylation system is dynamically lost in multiple lineages (**Fig. 5** and **Fig. 6A**) (Bewick et al., 2017b; Engelhardt et al., 2022), yet the annotation of DNMTs, UHRF1, CDCA7 and HELLS is unambiguous. As a negative control, we included INO80, which is dynamically lost in several Ecdysozoa lineages, such as *C. elegans* (**Fig. 5** and **6A**, **Fig. S2**, Table S1). As expected, CoPAP analysis showed a highly significant coevolutionary interaction between DNMT1 and UHRF1 (**Fig. 6B**). In addition, HELLS interacts with DNMT1, UHRF1 and CDCA7. In contrast, no linkage from INO80 or DNMT3 was determined. Altogether, these CoPAP analyses reproducibly found coevolutionary linkages between DNMT1-UHRF1 and CDCA7-HELLS, although the exact interaction topology is sensitive to the selection of species and protein classification methods.

Loss of CDCA7 in braconid wasps together with DNMT1 or DNMT3

CoPAP analysis detected the coevolutionary linkage between CDCA7 and DNMT1-UHRF1, rather than DNMT3. We were therefore intrigued by the apparent absence of CDCA7 and DNMT3 in two insect species, the red flour beetle *Tribolium castaneum* and the braconid wasp *Microplitis demolitor*, whose genomic DNA does not have any detectable 5mC despite the presence of DNMT1 (Bewick et al., 2017b; Schulz et al., 2018; Zemach et al., 2010) (**Fig. 5**, **Fig. S2**, Table S1). To further validate the co-loss of CDCA7 and DNMT1/DNMT3 in insects, we focused on the Hymenoptera clade (including *M. demolitor*), for which genome synteny has been reported (Li et al., 2021). Indeed, a striking synteny is observed in the genome region surrounding CDCA7 among the parasitic wood wasp (*Orussus abietinus*) and Aculeata species (bees *Bombus terrestris* and *Habropoda laborlosa*, and the eusocial wasp *Polistes canadensis*), which diverged ~250 MYA (Li et al., 2021; Peters et al., 2017) (**Fig. 7**, Table S2). In these species, CDCA7 is located between Methyltransferase-like protein 25 homolog (MET25, E) and Ornithine aminotransferase homolog

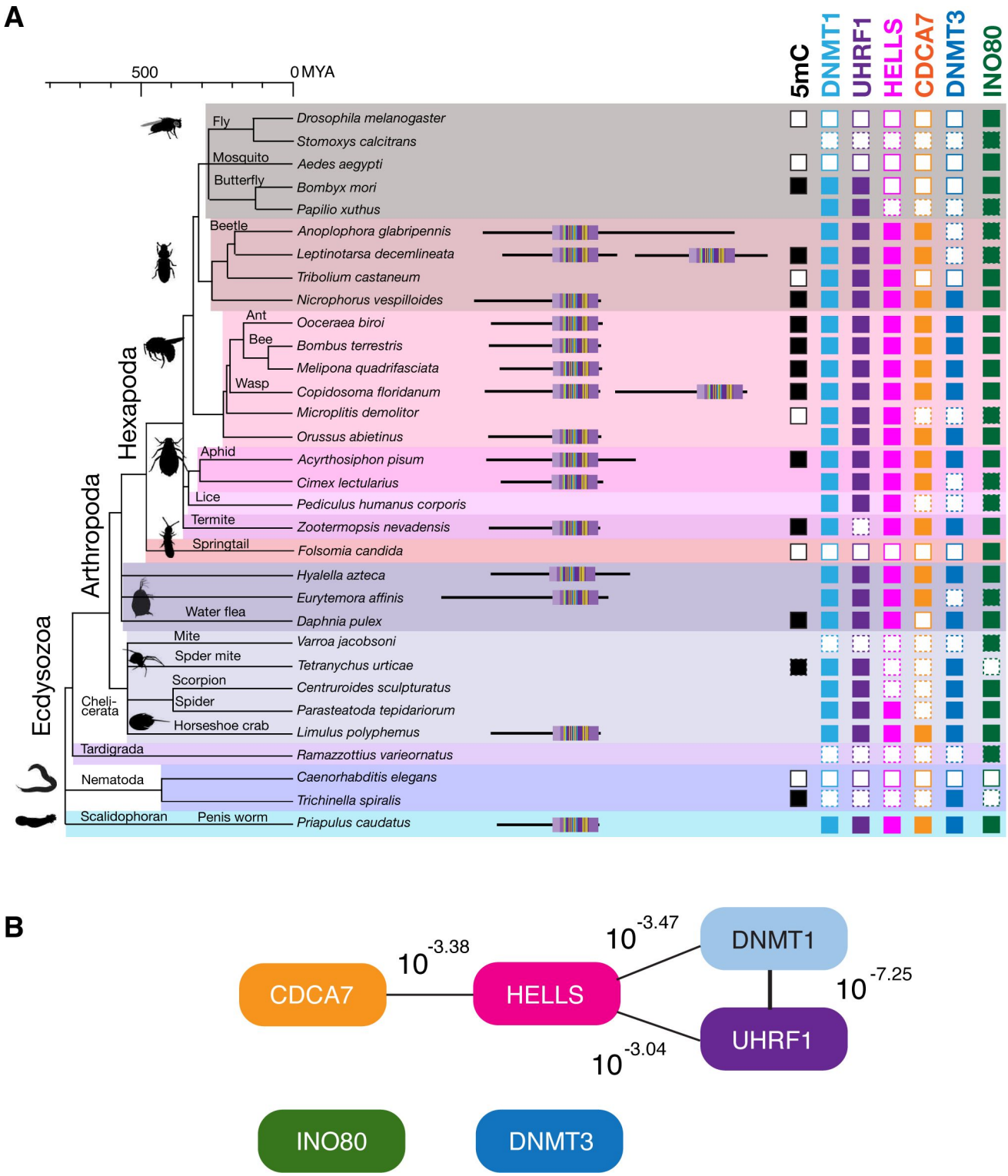

Figure 6.

Coevolution of CDCA7, HELLS, UHRF1 and DNMT1 in Ecdysozoa

A. Presence (filled squares) /absence (open squares) patterns of indicated proteins and genomic 5mC in selected Ecdysozoa species. Squares with dotted lines imply preliminary-level genome assemblies. Domain architectures of CDCA7 proteins with a zf-4CXXC_R1 domain are also shown. B. CoPAP analysis of 50 Ecdysozoa species. Presence/absence patterns of indicated proteins during evolution were analyzed. List of species are shown in Table S1. Phylogenetic tree was generated by amino acid sequences of all proteins shown in Table S1. The number indicates the *p*-values.

(OAT, F). In fact, the gene cluster containing LTO1 homolog (D), MET25 (E), OAT (F), and Zinc finger protein 808 (ZN808, G) is highly conserved in all the analyzed Hymenoptera species, but not outside of Hymenoptera (e.g., *Drosophila*).

However, CDCA7 is absent from this gene cluster in parasitoid wasps, including Ichneumonoidea wasps (braconid wasps [*M. demolitor*, *Cotesia glomerata, Aphidius gifuensis*, *Fopius arisanus*] and ichneumon wasp [*Venturia canescent*]) and chalcid wasps (*Copidosoma floridanum*, *Nasonia vitripennis*) (**Figure 7**, Table S3). Among Ichneumonoidea wasps, CDCA7 is undetectable in the genome of braconid wasps, while CDCA7 is located on a different chromosome in the ichneumon wasp *Venturia canescens*. In chalcid wasps, CDCA7 is present at a genome segment between Artemis (I) and Chromatin accessibility complex protein 1 (CHRC1, J). Supported by the chromosome-level genome assembly for *Cotesia glomerata* and *Aphidius gifuensis* (Feng et al., 2020; Pinto et al., 2021), the synteny analysis strongly indicates that CDCA7 is lost from braconid wasps. Intriguingly, braconid wasps co-lost CDCA7 either with DNMT3 (in *M. demolitor*, *Cotesia glomerata*, *Cotesia typhae* and *Chelonus insularis*) or with DNMT1 and UHRF1 (in *Fopius arisanus*, *Diachasma alloeum* and *Aphidius gifuensis*) (**Fig. 7** and Table S3). Notably, in addition to *M. demolitor* (which lacks CDCA7 and DNMT3), it has been reported that little or no 5mC can be detected in *Aphidius ervi* (which lacks CDCA7, DNMT1 and UHRF1, Table S3) (Bewick et al., 2017b). In contrast, all Hymenoptera species that are known to have detectable 5mC possess CDCA7 along with DNMT1, UHRF1, DNMT3, and HELLS, except for *Polistes canadensis*, which has all but DNMT3 (**Fig. 7** and Table S3). This co-loss of CDCA7 and DNA methylation (together with either DNMT1-UHRF1or DNMT3) in braconid wasps suggests that evolutionary preservation of CDCA7 is more sensitive to DNA methylation status *per se* than to the presence or absence of a particular DNMT subtype.

Discussion

Although DNA methylation is prevalent across eukaryotes, DNA methyltransferases are missing from a variety of lineages. Our study reveals that the nucleosome remodeling complex CHIRRC, composed of CDCA7 and HELLS, is frequently lost in conjunction with DNA methylation status. More specifically, evolutionary preservation of CDCA7 is tightly coupled to the presence of HELLS and DNMT1-UHRF1. The conservation of CDCA7's signature cysteine residues alongside three ICF-associated residues across diverse eukaryote lineages suggests a unique evolutionary conserved role in DNA methylation. Our co-evolution analysis suggests that DNA methylation-related functionalities of CDCA7 and HELLS are inherited from LECA.

The evolutionary coupling of CDCA7, HELLS and DNMT1 is consistent with a proposed role of HELLS in replication-uncoupled DNA methylation maintenance (Ming et al., 2020). Commonly, DNA methylation maintenance occurs directly behind the DNA replication fork. Replication-uncoupled DNA methylation maintenance is distinct from this process (Nishiyama et al., 2020), and HELLS and CDCA7 may be important for the maintenance of DNA methylation long after the completion of DNA replication, particularly at heterochromatin where chromatin has restricted accessibility (Ming et al., 2020). The tighter evolutionary coupling of CDCA7-HELLS to DNMT1 rather than to DNMT3 may also reflect a potential capacity of CDCA7 in sensing DNA methylation, similar to the way that the zf-CXXC domain is sensitive to CpG methylation (Long et al., 2013), but in a replication-coupled manner. However, this does not necessarily suggest that the role of CDCA7 is always coupled to maintenance DNA methylation.

The loss of CDCA7 is not always coupled to the loss of DNMT1 or HELLS. In the Hymenoptera clade, CDCA7 loss in the braconid wasps is accompanied with the loss of DNMT1/UHRF1 or the loss of DNMT3. Among these species, it was reported that 5mC DNA methylation is undetectable in *M. demolitor*, which harbors DNMT1, UHRF1 and HELLS but lost DNMT3 and CDCA7 (Bewick et al., 2017b). Similarly, in the Coleoptera clade, the red flour beetle *T. castaneum* possesses DNMT1

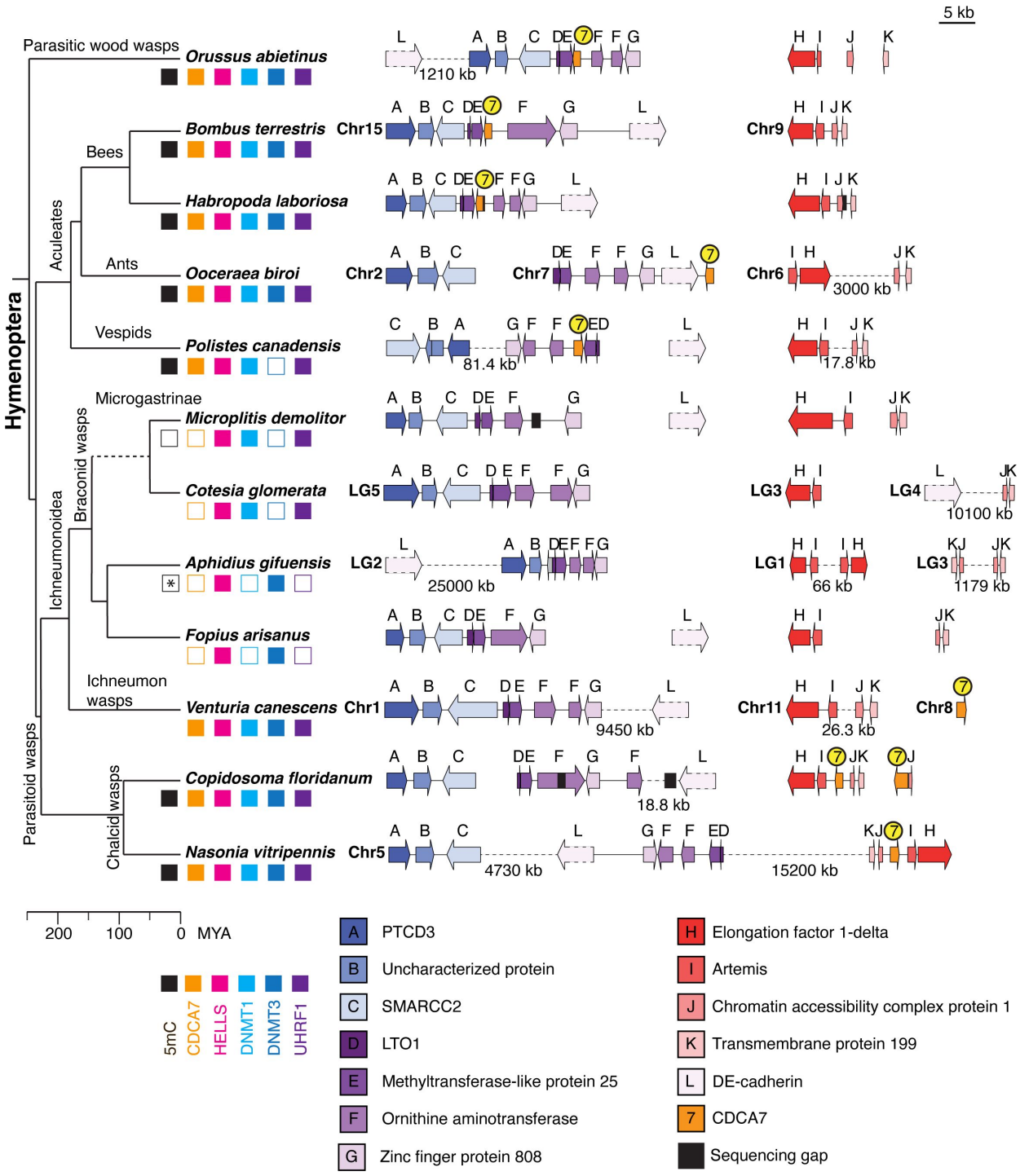

Figure 7.

Synteny of Hymenoptera genomes adjacent to CDCA7 genes

Genome compositions around CDCA7 genes in Hymenoptera insects are shown. For genome with annotated chromosomes, chromosome numbers (Chr) or linkage group numbers (LG) are indicated at each gene cluster. Gene clusters without chromosome annotation indicate that they are within a same scaffold or contig. Dash lines indicate the long linkages not proportionally scaled in the figure. Due to their extraordinarily long sizes, DE-cadherin genes (L) are not scaled proportionally. Presence and absence of 5mC, CDCA7, HELLS, DNMT1, DNMT3, and UHRF1 in each genome is indicated by filled and open boxes, respectively. Absence of 5mC in *Aphidius gifuensis* (marked with an asterisk) is deduced from the study in *Aphidius ervi* (Bewick et al., 2017b), which has an identical presence/absence pattern of the listed genes (Table S3).The phylogenetic tree is drawn based on published analysis (Li et al., 2021; Peters et al., 2017) and TimeTree.

and HELLS, but lost DNMT3 and CDCA7 (**Fig. 5**, **Fig. S2**, Table S1). Since DNMT1 is essential for the embryonic development of *T. castaneum* and CpG DNA methylation is undetectable in this organism (Schulz et al., 2018), it has been predicted that DNMT1 has a function independent of DNA methylation in this species. Indeed, species that preserves DNMTs and/or HELLS in the absence of CDCA7 emerge repeatedly during eukaryote evolution, whereas CDCA7 appears to be immediately dispensable in species that have a dampened requirement for DNA methylation. In other words, there is no evolutionary advantage to retain CDCA7 in the absence of DNA methylation, and CDCA7 is almost never maintained in the absence of any DNMTs. Alternatively, could the loss of CDCA7 precede the loss of DNA methylation? If CDCA7 is important to promote DNA methylation maintenance, the loss of CDCA7 may exacerbate an impaired epigenetic environment and stimulate the adaptation of the organism towards DNA methylation-independent epigenetic mechanisms, thereby decreasing the necessity to maintain the DNA methylation system. In this way, the loss of CDCA7 could trigger the subsequent loss of DNMTs (and HELLS), unless these proteins acquired important DNA methylation-independent roles. In line with this scenario, insects, whose DNA methylation is largely limited to gene bodies, possess robust DNA methylation-independent mechanisms (such as piwi-RNA and H3K9me3) to silence transposons (Bonasio et al., 2012; Feng et al., 2010; Libbrecht et al., 2016)(Libbrecht et al., 2016; Wang et al., 2013; Zemach et al., 2010). The loss of CDCA7 is more frequently observed in insects than plants and vertebrates, where DNA methylation and HELLS/DDM1 play major roles in silencing transposons. (Czech et al., 2018; Dunican et al., 2013; Huang et al., 2004; Kato et al., 2003; Miura et al., 2001; Onishi et al., 2021; Osakabe et al., 2021; Vongs et al., 1993). Thus, lowering the demand of DNA methylation at transposable elements might reduce the essentiality of CDCA7 (and then perhaps that of DNMTs), though it is difficult to deduce which evolutionary change occurs first.

Considering the importance of HELLS/DDM1 in silencing transposable elements, it is intriguing that CDCA7, HELLS, and DNMT1 are conserved in many insects, in which transposable elements are generally hypomethylated (Bonasio et al., 2012; Feng et al., 2010; Libbrecht et al., 2016; Wang et al., 2013; Zemach et al., 2010). DNMT1 knockout in the clonal raider ant, *Ooceraea biroi*, which has a full set of DNMT1, UHRF1, DNMT3, CDCA7 and HELLS (**Fig. 5**, **7**), does not cause major developmental defects but leads to failure in reproductive oogenesis and compromised longevity (Ivasyk et al., 2023). Similarly, DNMT1 is essential for meiosis (but not for repression of transposable elements) in the milkweed bug *Oncopeltus fasciatus* (Bewick et al., 2019b; Washington et al., 2021). HELLS and DNMT1 are also important for meiotic progression in mice (Baumann et al., 2020; Spruce et al., 2020; Takada et al., 2021; Zeng et al., 2011). In *Neurospora*, DNA methylation is a critical component for trans-sensing homologous chromosomes during meiosis (Pratt et al., 2004). Since 5mC is highly mutagenic and DNA methylation patterns are altered during aging (Lowe et al., 2018; Wang et al., 2020), it is tempting to speculate that the precise maintenance of DNA methylation patterns may act as a hallmark to distinguish between young/healthy DNA and old/mutated (or competitive/pathogenic) DNA during meiosis, perhaps in a way analogous to the role of methylated DNA for mismatch repair in bacteria (Schofield and Hsieh, 2003).

The observation that some species retain HELLS but lose CDCA7 (while the reverse is never true) suggests that HELLS can evolve a CDCA7-independent function. Indeed, it has been suggested that the sequence-specific DNA-binding protein PRDM9 recruits HELLS to meiotic chromatin to promote DNA double-strand breaks and recombination (Imai et al., 2020; Spruce et al., 2020). Unlike CDCA7, clear PRDM9 orthologs are found only in metazoans, and are even lost in some vertebrates such as *Xenopus laevis* and *Gallus gallus* (Birtle and Ponting, 2006). Thus, it is plausible that HELLS acquires a species-specific CDCA7-independent role during evolution through acquiring a new interaction partner. Another example of this may be found in *S. cerevisiae*, where DNA methylation and CDCA7 are absent and the HELLS homolog Irc5 interacts with cohesin to facilitate its chromatin loading (Litwin et al., 2017). In *Neurospora*, where

genomic DNA is methylated, the HELLS homolog MUS-30 is not required for DNA methylation, but plays a role in DNA damage responses (Basenko et al., 2016). We speculate that the role of CDCA7 is evolutionarily more tightly coupled to DNA methylation than HELLS.

Recently, the role of HELLS in the deposition of the histone variant macroH2A, which compacts chromatin, has been reported in mice (Ni and Muegge, 2021; Ni et al., 2020; Xu et al., 2021). Similarly, in *Arabidopsis*, DDM1 is critical for deposition of H2A.W, which is enriched on heterochromatin, in a manner independent of DNA methylation (Osakabe et al., 2021). The role of CDCA7 in the deposition of these H2A variants remains to be tested. HELLS and DDM1 can directly interact with macroH2A and H2A.W, respectively, even in the absence of CDCA7 (Ni and Muegge, 2021; Osakabe et al., 2021). It is thus possible that HELLS/DDM1 family proteins have an evolutionary conserved function in H2A variant deposition independent of CDCA7 and DNA methylation. However, there is no clear indication of coevolution of HELLS and macroH2A, as macroH2A is largely missing from insects and the chelicerata *Centruroides sculpturatus* has macroH2A (XP_023217082, XP_023212717) but lost HELLS and CDCA7.

Whereas the class I CDCA7-like proteins are evolutionarily coupled to HELLS and DNMT1-UHRF1, other variants of zf-4CXXC_R1 are widespread in eukaryotes except for metazoans, which encode only CDCA7 orthologs and their close paralogs (with the exception of the sponge *A. queenslandica*). Proteins containing a diverged variant of zf-4CXXC_R1 and the JmjC domain, such as IBM1 in Arabidopsis and DMM-1 in Neurospora, are broadly found in Viridiplantae and Fungi (Honda et al., 2010; Saze et al., 2008; Zhang et al., 2022). Functional studies of IBM1 and DMM-1 suggested that they contribute to DNA methylation regulation via indirect mechanisms. As IBM1 and DMM-1 do not preserve ICF-associated residues, which are critical for nucleosome binding in CDCA7 (Jenness et al., 2018), it is likely that these variants of zf-4CXXC_R1 are adapted to recognize different structural features of the genome and no longer preserve the DNA methylation function of CDCA7 orthologs.

Considering the broad conservation of DNA methylation in vertebrates (Hemmi et al., 2000; Kondilis-Mangum and Wade, 2013), plants (Deleris et al., 2016), prokaryotes (Beaulaurier et al., 2019; Casadesus and Low, 2006; Dimitriu et al., 2020; Vasu and Nagaraja, 2013) and Archaea (Grogan, 2003; Hayashi et al., 2021; Ishikawa et al., 2005; Prangishvili et al., 1985), along with the existence of SNF2-like proteins (SSO1653) in prokaryotes and Archaea (Flaus et al., 2006), we hypothesize that the evolutionary advent of zf-4CXXC_R1-containing CDCA7 was a key step to transmit the DNA methylation system from the last universal common ancestor (LUCA) to the eukaryotic ancestor with nucleosome-containing genomes.

Methods

Building a curated list of 180 species for analysis of evolutionary co-selection

A list of 180 eukaryote species was manually generated to encompass broad eukaryote evolutionary clades (Table S1). Species were included in this list based on two criteria: (i) the identification UBA1 and PCNA homologs, two highly conserved and essential proteins for cell proliferation; and (ii) the identification of more than 6 distinct SNF2 family sequences. Homologs of CDCA7, HELLS, UBA1 and PCNA were identified by BLAST search against the Genbank eukaryote protein database available at National Center for Biotechnology Information using the human protein sequence as a query (NCBI). Homologs of human UHRF1, ZBTB24, SMARCA2/SMARCA4, INO80, RAD54L, EZH2, EED or Suz12 were also identified based on the RBH criterion. To get a sense of genome assembly level of each genome sequence, we divided “Total Sequence Length” by “Contig N50” (“length such that sequence contigs of this length or longer include half the bases of the assembly”; *https://www.ncbi.nlm.nih.gov/assembly/help/*). In the

species whose genome assembly level is labeled as “complete”, this value is close to the total number of chromosomes or linkage groups. As such, as a rule of thumb, we arbitrarily defined the genome assembly “preliminary”, if this value is larger than 100. In **Figure 5**, these species with preliminary-level genome assembly were noted as boxes with dotted outlines.

CDCA7 homolog identification and annotation

BLAST search was conducted using human CDCA7 (NP_114148) as the search query against NCBI protein database. The obtained list of CDCA7 homologs was classified based on the conservation of eleven cysteine and three ICF-associated residues in the zf-4CXXC_R1 domain, as described in Results. This classification was further validated based on their clustering in a phylogenetic tree built from the CLUSTALW alignment of the zf-4CXXC_R1 domain identified by NCBI conserved domains search (Higgins and Sharp, 1988; Lu et al., 2020; Thompson et al., 1994) (**Fig S1**), using MacVector (MacVector, Inc.). The cluster of class I zf-4CXXC_R1 domain-containing proteins (where all three ICF-associated residues are conserved) was segregated from other variants of zf-4CXXC_R1-containing proteins except for the moss Physcomitrium XP_024393821 and green algae Bathycoccus XP_007512509 and Chlamydomonas GAX81623, which has class II zf-4CXXC-R1 (where the ICF-associated glycine residue is substituted). To further assess this validation, another phylogenetic tree was built by the optimal maximum likelihood-based model selected by IQ-TREE 2.2.2.6 (Minh et al., 2020), using the zf-4CXXC_R1 domain alignment generated and selected by Muscle v5 (Dataset S1)(Edgar, 2022). A consensus tree was then constructed from 1000 bootstrap trees using UFBoot2 (Dataset S2) (Hoang et al., 2018). The phylogenetic tree built by this second method deviated all class II and other zf-4CXXC_R1 variants from CDCA7 orthologs with class I zf-4CXXC_R1 (Dataset S2).

HELLS homolog identification and annotation

Putative HELLS orthologs were first identified according to the RBH criterion. Briefly, a BLAST search was conducted using human HELLS as the query sequence, after which protein sequences of obtained top hits (or secondary hits, if necessary) in each search were used as a query sequence to conduct reciprocal BLAST search against the *Homo sapiens* protein database. If the top hit in the reciprocal search returned the original input sequence (i.e. human HELLS), it was temporarily annotated as an orthologous protein. If HELLS showed up as a next best hit, it is temporarily listed as a “potential HELLS ortholog”. To further validate the identified HELLS orthologs, full length amino acid sequences of these proteins were aligned using CLUSTALW in MacVector with homologs of the Snf2 family proteins SMARCA2/SMARCA4, CHD1/CHD3/CHD7, ISWI, RAD54L, ATRX, HLTF, TTF2, SHPRH, INO80, SMARCAD1, SWR1, MOT1, ERCC6, and SMARCAL1, which were also identified and temporarily annotated with a similar reciprocal BLAST search methods. The phylogenetic tree generated by this full-length alignment separated the clade containing HELLS, DDM1, INO80, SMARCA2/SMARCA4, CHD1/CDH3/CHD7, and ISWI from other Snf2 family proteins (**Fig. S3**). This alignment was used to define the conserved SNF2 domain and variable linker regions in the putative HELLS orthologs and other homologous SNF2-family proteins (INO80, SMARCA2/SMARCA4, CHD1/CDH3/CHD7, and ISWI). The variable linker regions were then removed from selected proteins for each kingdom/phylum to conduct the secondary CLUSTALW alignment in MacVector, from which a phylogenetic tree was generated (**Fig. S4**). A distinct cluster of HELLS and other SNF2 family proteins can be identified from the phylogenetic tree, hence confirming that the annotation of HELLS orthologs based on the reciprocal BLAST search method is reasonable (**Fig. S4**). Exceptions are *Leucosporidium creatinivorum* ORY88017 and ORY88018, which did not cluster within the HELLS clade of the phylogenetic tree. However, we decided to annotate *L. creatinivorum* ORY88017 and ORY88018 as HELLS orthologs since among other *L. creatinivorum* SNF2-family proteins in this species these two proteins are most similar to human HELLS while clear orthologs of other *L. creatinivorum* SNF2 family proteins, CHD1 (ORY55731), ISWI (ORY89162), SMARCA2/4 (ORY76015), SRCAP/SWR1 (ORY90750) and INO80 (ORY91599), can be identified. To further validate the phylogenetic tree generated in **Fig. S4**, another phylogenetic tree was built by the optimal maximum likelihood-based model selected by

IQ-TREE 2.2.2.6 (Minh et al., 2020), using the alignment generated and selected by Muscle v5 (Dataset S3) (Edgar, 2022). A consensus tree was then constructed from 1000 bootstrap trees using UFBoot2 (Dataset S4) (Hoang et al., 2018). The topology of the phylogenetic tree build by this second method was consistent with the original tree shown in **Fig. S4**.

DNMT homolog identification and annotation

Proteins with a DNA methyltransferase domain were identified with BLAST searches using human DNMT1 and DNMT3A. Additional BLAST searches were conducted using human DNMT2, *C. neoformans* DNMT5, *N. crassa* Dim-2 and DNMT4, *Thalassiosira pseudonana* XP_002287631 (Huff and Zilberman, 2014), and Kwoniella mangroviensis DNMTX (XP_019001759) (Catania et al., 2020). DNMT domains were extracted from selected proteins that represent each kingdom/phylum and DNMT type (including ambiguous classes) based on a NCBI conserved domains search, and were aligned with CLUSTALW to build a phylogenetic tree in MacVector. To further assess this validation, another phylogenetic tree was built by the optimal maximum likelihood-based model selected by IQ-TREE 2.0.3 (Minh et al., 2020), using the alignment generated and selected by Muscle v5 (Dataset S5) (Edgar, 2022). A consensus tree was then constructed from 1000 bootstrap trees using UFBoot2 (Dataset S6) (Hoang et al., 2018). The topology of the phylogenetic tree built by this second method was largely consistent with the original tree except for placement of orphan DNMTs. In both trees, *Vitrella brassicaformis* CEM15752, *Fragilariopsis cylindrus* OEU08290, OEU15938, *Thalassiosira pseudonana* XP_002295472, and *Tribonema minus* KAG5185060 form a clade, which, only in the IQ-TREE generated tree, splits from the canonical DNMT4 clade. As proteins in this clade have been previously annotated as DNMT4 (Huff and Zilberman, 2014), we followed this classification and present this IQ-TREE-based phylogenetic tree of DNMTs in **Fig. S6**, which was illustrated using the ETE3 toolkit.

CoPAP

The published method was used (Cohen et al., 2013). The curated list of orthologous proteins listed in Table S1 was first used to generate a presence-absence FASTA file. Next, a phylogenetic species tree was generated from all orthologous protein sequences listed in Table S1 using the ETE3 toolkit. For this, protein sequences were retrieved using the rentrez Bioconductor package and exported to a FASTA file alongside a COG file containing gene to orthologous group mappings. ETE3 was used with the parameters *-w clustalo_default-trimal01-none-none* and *-m cog_all-alg_concat_default-fasttree_default* and the resulting tree exported in Newark format. CoPAP was run using default parameters and results visualised using Cytoscape. Code and files required for CoPAP input generation as well CoPAP parameters and output results can be found in our Github repository (*https://github.com/RockefellerUniversity/Copap_Analysis*). A method to calculate *p*-value for CoPAP was described previously (Cohen et al., 2012). Briefly, for each pair of tested genes, Pearson's correlation coefficient was computed. Parametric bootstrapping was used to compute a *p*-value by comparing it with a simulated correlation coefficient calculated based on a null distribution of independently evolving pairs with a comparable exchangeability (a value reporting the likelihood of gene gain and loss events across the tree)

As negative and positive controls for the CoPAP analysis, we identified several well-conserved protein orthologs across the panel of 180 eukaryotic species, including Snf2-like proteins SMARCA2/SMARCA4, INO80, and RAD54L (Flaus et al., 2006), as well as subunits of the polycomb repressive complex 2 (PRC2), which plays an evolutionary conserved role in gene repression via deposition of the H3K27me3 mark. PRC2 is conserved in species where DNMTs are absent (including in *D. melanogaster* and *C. elegans*) but is frequently lost particularly in several lineages of SAR and Fungi (Sharaf et al., 2022). Among the four core subunits of PRC2, we focused on the catalytic subunit EZH1/2, EED, and SUZ12, since the fourth subunit RbAp46/48 has a PRC2-independent role (Margueron and Reinberg, 2011). We are aware that the reciprocal BLAST search missed previously reported highly divergent functional orthologs of SUZ12 in *Neurospora*

As negative and positive controls for the CoPAP analysis, we identified several well-conserved protein orthologs across the panel of 180 eukaryotic species, including Snf2-like proteins SMARCA2/SMARCA4, INO80, and RAD54L (Flaus et al., 2006), as well as subunits of the polycomb repressive complex 2 (PRC2), which plays an evolutionary conserved role in gene repression via deposition of the H3K27me3 mark. PRC2 is conserved in species where DNMTs are absent (including in *D. melanogaster* and *C. elegans*) but is frequently lost particularly in several lineages of SAR and Fungi (Sharaf et al., 2022). Among the four core subunits of PRC2, we focused on the catalytic subunit EZH1/2, EED, and SUZ12, since the fourth subunit RbAp46/48 has a PRC2-independent role (Margueron and Reinberg, 2011). We are aware that the reciprocal BLAST search missed previously reported highly divergent functional orthologs of SUZ12 in *Neurospora* (Jamieson et al., 2013), and EED and Suz12 in *Paramecium* (Miro-Pina et al., 2022). However, we did not attempt to use these divergent homologs of EED and SUZ12 as baits to expand our search in order to consistently apply our homology-based definition of orthologs.

Hymenoptera synteny analysis

The mapping of gene loci is based on the information available on the Genome Data Viewer (*https://www.ncbi.nlm.nih.gov/genome/gdv*). Genome positions of listed genes are summarized in Table S2.

Artworks

Artworks of species images were obtained from PhyloPic.com, of which images of Daphnia, Platyhelminthes, Tribolium and Volvox were generated by Mathilde Cordellier, Christopher Laumer/T. Michael Keesey, Gregor Bucher/Max Farnworth and Matt-Crook, respectively.

Acknowledgements

We thank Daniel Kronauer and Rochelle Shih for critical reading of the manuscript, and

D. Kronauer, Li Zhao, Junhui Peng and Erick Jarvis for helpful discussion. We thank Yasuhiro Arimura and Hideaki Konishi for their assistance in multiple sequence alignment and phylogenetic tree generation. The research by Q. J. was in part executed through Chemers Neustein Summer Undergraduate Research Fellowship (SURF) Program at the Rockefeller University.

Funding

This work was supported by National Institutes of Health Grants R35GM132111 to H.F., and the Women & Science Postdoctoral Fellowship Program at The Rockefeller University to I.E.W.

Competing interests

H.F. is affiliated with Graduate School of Medical Sciences, Weill Cornell Medicine, and Cell Biology Program, the Sloan Kettering Institute. The authors declare that no competing financial interests exist.

Authors' Contribution

H. F., Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing; I.E.W., Funding acquisition, Investigation, Supervision, Writing – review & editing; Q.J., Data curation, Investigation, Visualization; J.L., Data curation, Methodology, Software; T.C., Data curation, Methodology, Software.

Supplemental Figure Legends

Table S1. Lists of proteins and species used in this study

Tab1, Full list. The list contains species names, their taxonomies, Genbank accession numbers of proteins, PMID of references supporting the 5mC status, and genome sequence assembly statistics. ND; not detected. DNMT5 proteins shown in red lack the Snf2-like ATPase domain. UHRF1 proteins shown in red lack the Ring-finger E3 ubiquitin-ligase domain. CDCA7 proteins shown in red indicate ambiguous annotation as described in the main text. CDCA7 orthologs that contain additional conserved domains found by NCBI CD-search were shown in light blue.

Tab2, Full list 2. The list is used to make presence (1) or absence (0) list.

Tab3 Ecdysozoa CoPAP. List of presence/absence annotations for Ecdysozoa species used for CO-PAP analysis.

Tab4 Full CoPAP1. List of presence/absence data annotations for the panel of all 180 species used for CO-PAP analysis. Fungal CDCA7F proteins with class II zf-4CXXC_R1 are included in CDCA7.

Tab5 Full CoPAP2. List of presence/absence data annotations for the panel of all 180 species used for CO-PAP analysis. Fungal CDCA7F proteins are included in class II zf-4CXXC_R1.

Tab6 Full clustering. Table used for clustering analysis

Tab7 Metazoan invertebrates. Table used for clustering analysis for metazoan invertebrates.

Table S2. Summary of Hymenoptera genome location supporting Figure 7

Table S3. Lists of proteins in Hymenoptera species supporting Figure 7

Dataset S1. Multiple sequence alignment of zf-4CXXC_R1 domains. Alignment was done by MUSCLE v5.

Dataset S2. An IQ-TREE result of the consensus phylogenetic tree generation of zf-4CXXC_R1 containing proteins using Dataset S1.

Dataset S3. Multiple sequence alignment of the SNF2 ATPase domains of HELLS orthologs and other close homologs after removing the variable linker regions. Alignment was done by MUSCLE v5.

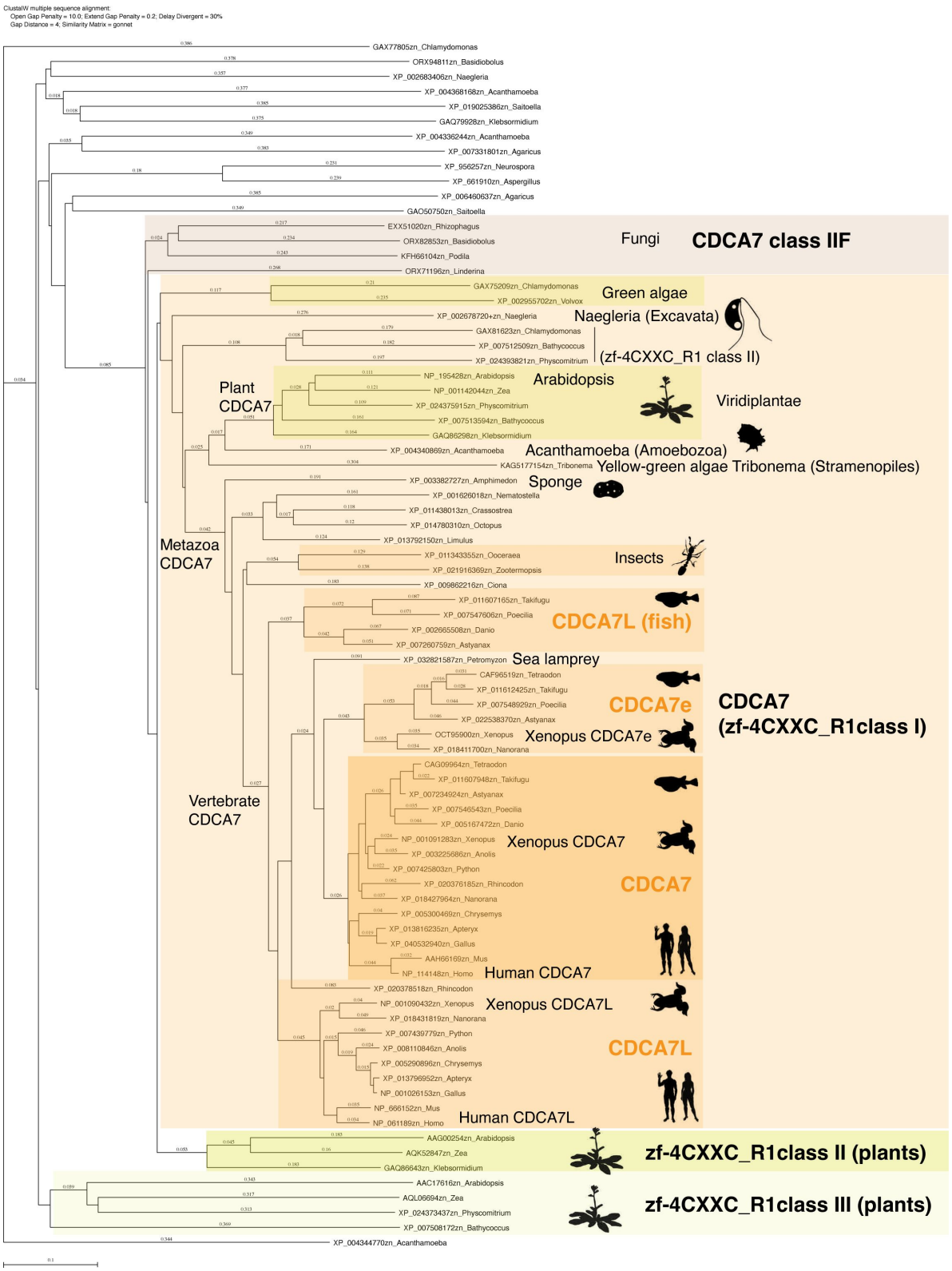

Figure S1.

Evolutionary conservation of CDCA7-family proteins and other zf-4CXXC_R1-containig proteins

Amino acid sequences of zf-4CXXC_R1 domain from indicated species were aligned with CLUSTALW. A phylogenetic tree of this alignment is shown. Genbank accession numbers of analyzed sequences are indicated.

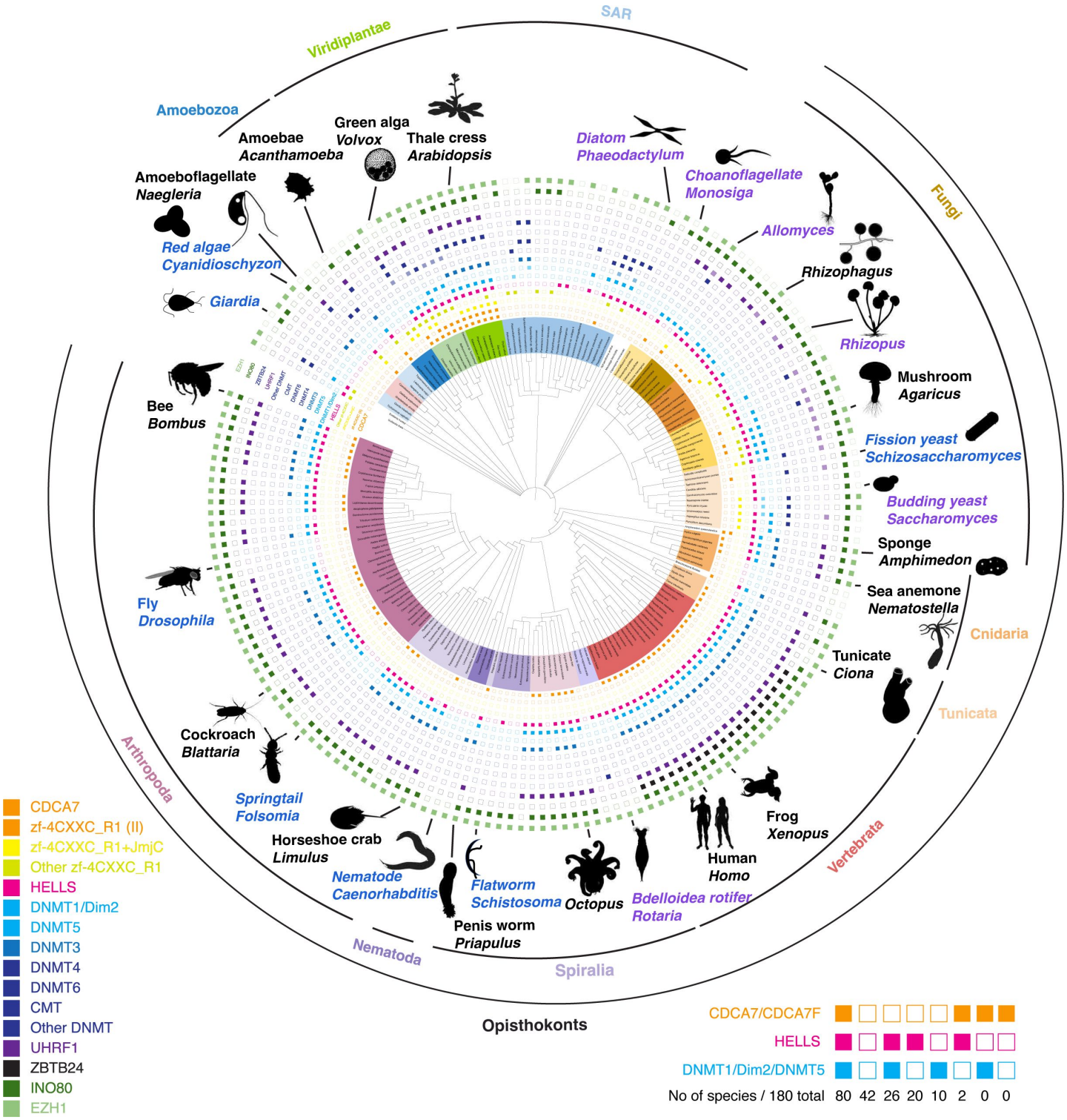

Figure S2.

Evolutionary conservation of CDCA7, HELLS and DNMTs

Presence and absence of each annotated proteins in the panel of 180 eukaryote species is marked as filled and blank boxes. The phylogenetic tree was generated based on NCBI taxonomy by phyloT. Bottom right; summary of combinatory presence or absence of CDCA7 (including fungal CDCA7F containing class II zf-4CXXC_R1), HELLS, and maintenance DNA methyltransferases DNMT1/Dim-2/DNMT5.

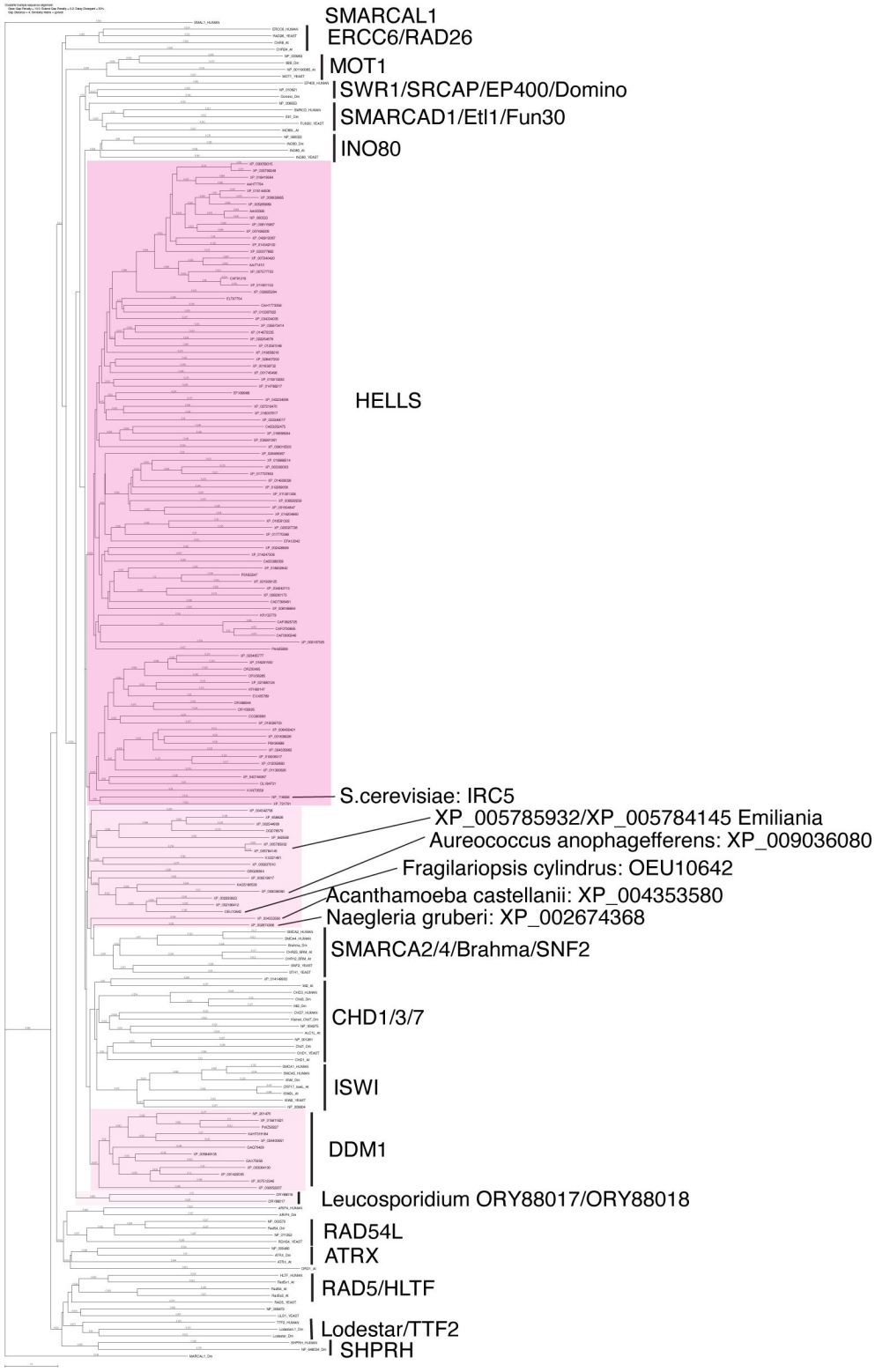

Figure S3.

Phylogenetic tree of HELLS and other SNF2 family proteins

Amino acid sequences of full-length HELLS proteins from the panel of 180 eukaryote species listed in Table S1 were aligned with full length sequences of other SNF2 family proteins with CLUSTALW. A phylogenetic tree of this alignment is shown. Genbank accession numbers of analyzed sequences are indicated.

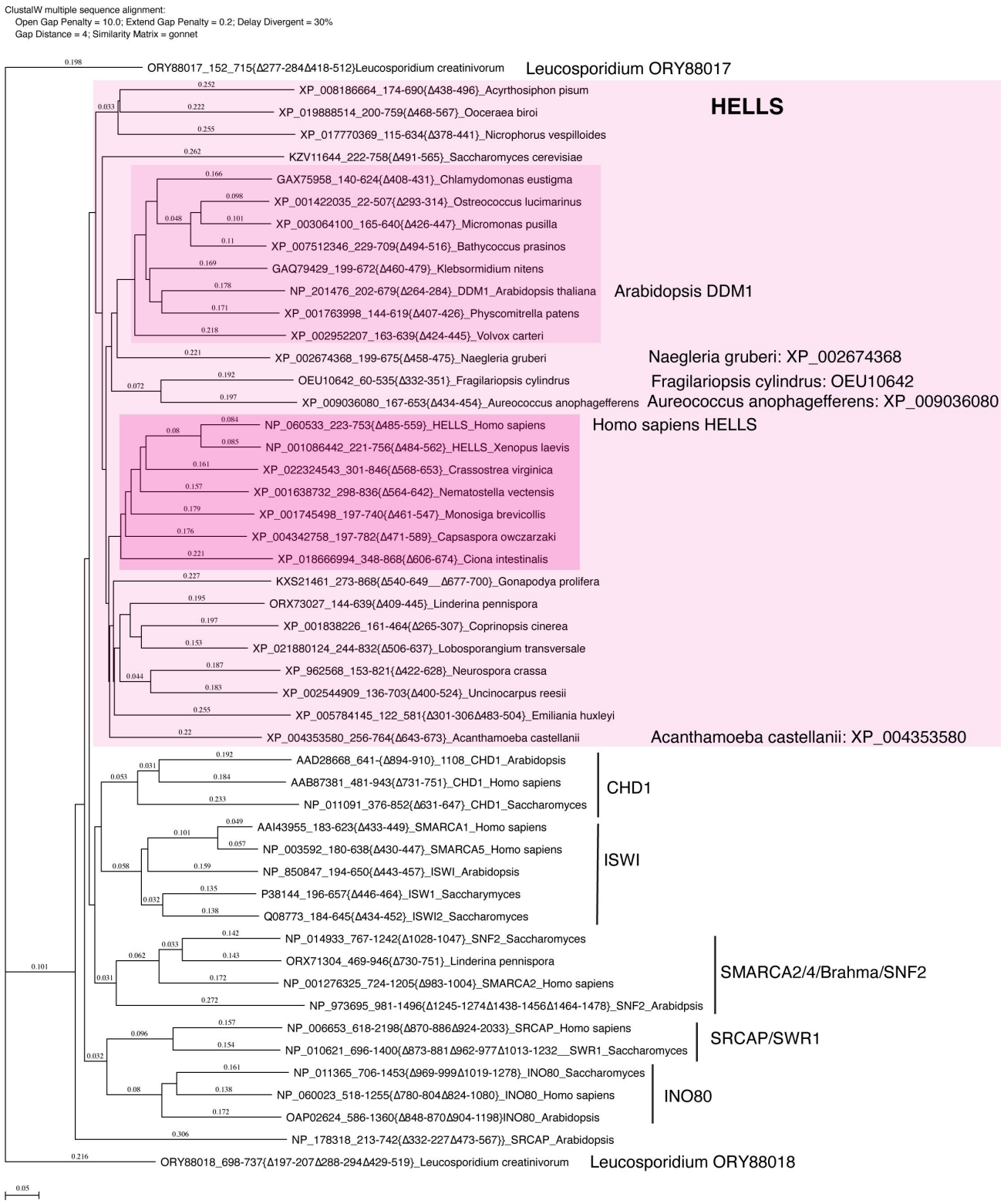

Figure S4.

Phylogenetic tree of the SNF2-domain

Amino acid sequences of SNF2-doman without variable insertions from representative HELLS and DDM1-like proteins from **Figure S3** were aligned with the corresponding domain of other SNF2 family proteins with CLUSTALW. A phylogenetic tree of this alignment is shown. Genbank accession numbers of analyzed sequences are indicated.

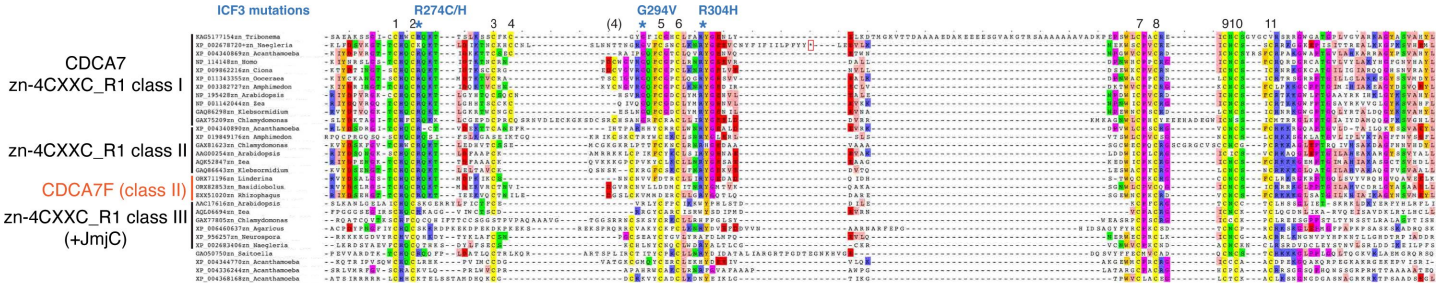

Figure S5.

Sequence alignment and classification of zf-4CXXC_R1 domains across eukaryotes

CDCA7 orthologs are characterized by the class I zf-4CXXC_R1 domain, where eleven cysteine residues and three residues mutated in ICF patients are conserved. Class II zf-4CXXC_R1 domain is similar to class I except that ICF-associated glycine (G294 in human) is substituted. Class III is zf-4CXXC_R domain with more substitutions at the ICF-associated residues (R274 and/or G294). Proteins that also contain JmjC domain (sequence not shown here) are also indicated. Note that codon frame after the stop codon (an asterisk in a magenta box) of Naegleria XP_002678720 encodes a peptide sequence that aligns well with human CDCA7, indicating that the apparent premature termination of XP_002678720 is likely caused by a sequencing or annotation error.

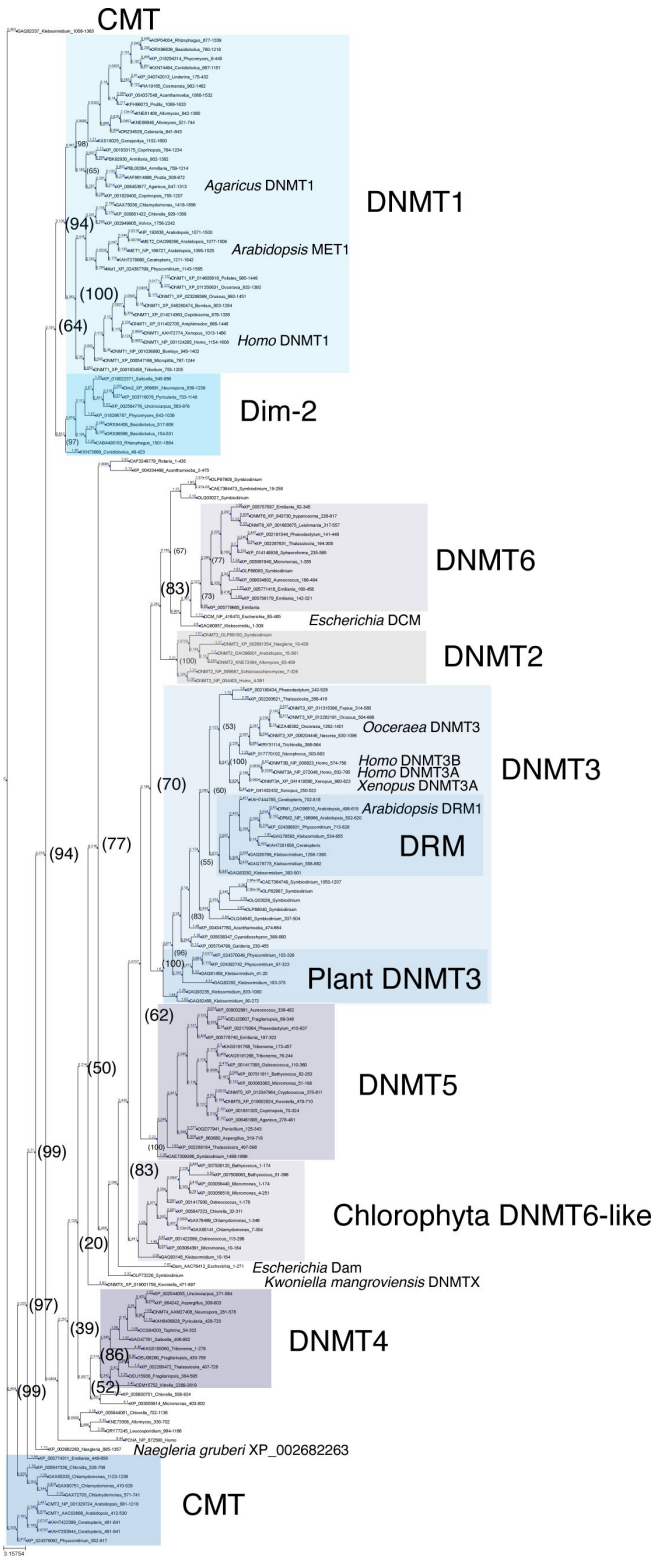

Figure S6.

Phylogenetic tree of DNMT proteins

DNA methyltransferase domain of DNMT proteins across eukaryotes (Table S1, excluding majority of those from Metazoa), the Escherichia coli DNA methylases DCM and Dam, and Homo sapiens PCNA as an outlier sequence, were aligned with CLUSTALW. A phylogenetic tree classification of this alignment is shown.

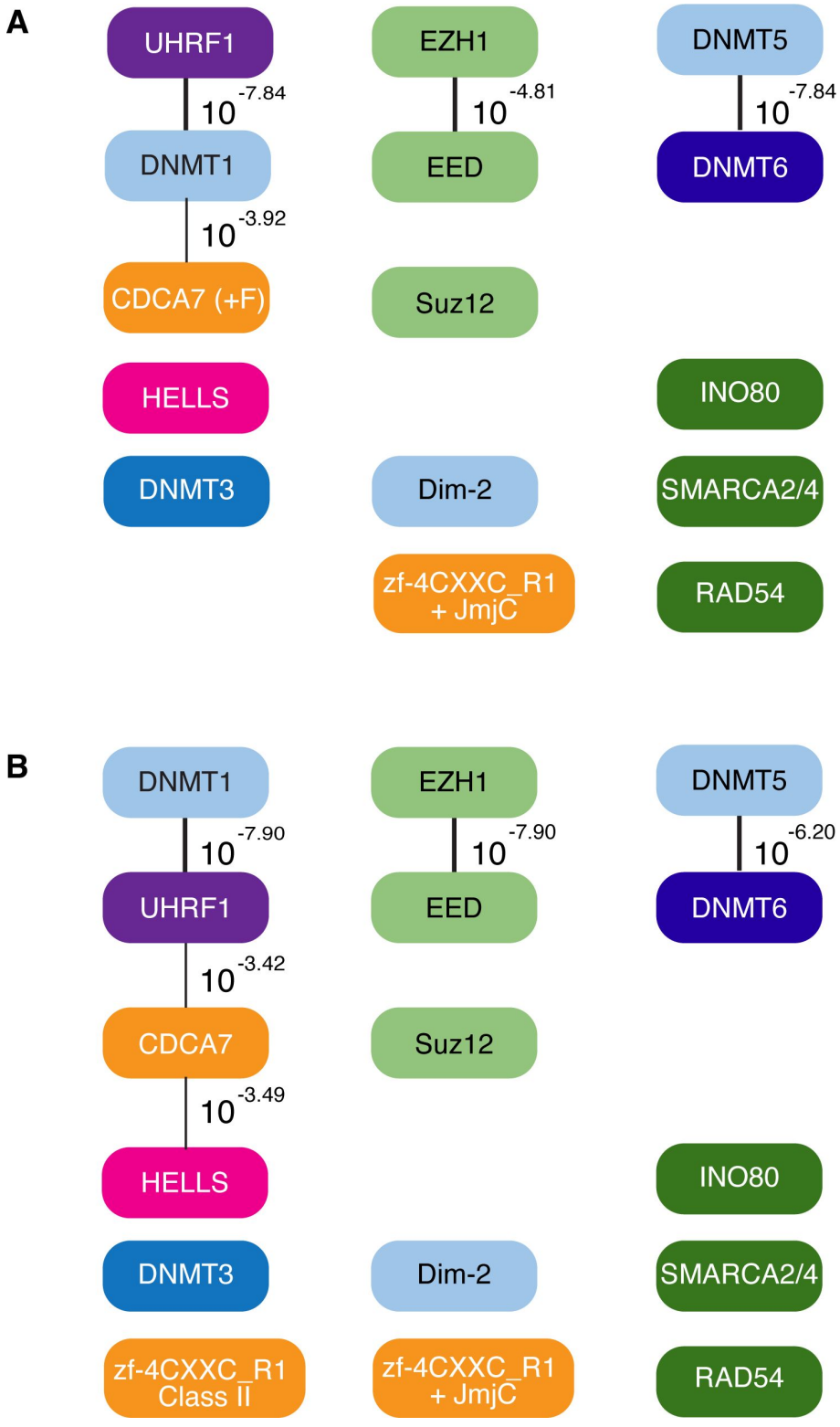

Figure S7.

CoPAP analysis of CDCA7, HELLS and DNMTs in eukaryotes

CoPAP analysis of 180 eukaryote species. Presence and absence patterns of indicated proteins during evolution were analyzed. List of species are shown in Table S1 (**A**, Full CoPAP1; **B**, Full CoPAP2). Fungal CDCA7F proteins are included in CDCA7 and zf-4CXXC_R1 class II in **A** and **B**, respectively. Phylogenetic tree was generated by amino acid sequences of all proteins shown in Table S1. The number indicates the *p*-values.

Dataset S4. An IQ-TREE result of the consensus phylogenetic tree generation of HELLS orthologs and other close homologs using Dataset S3.

Dataset S5. Multiple sequence alignment of DNA methyltransferase domains. Alignment was done by MUSCLE v5.

Dataset S6. An IQ-TREE result of the consensus phylogenetic tree generation of DNMTs using Dataset S5.

References

Aravind L., Iyer L.M (2012) **The HARE-HTH and associated domains: novel modules in the coordination of epigenetic DNA and protein modifications** *Cell Cycle* **11**:119–131

Basenko E.Y., Kamei M., Ji L., Schmitz R.J., Lewis Z.A (2016) **The LSH/DDM1 Homolog MUS-30 Is Required for Genome Stability, but Not for DNA Methylation in Neurospora crassa** *PLoS Genet* **12**

Baumann C., Ma W., Wang X., Kandasamy M.K., Viveiros M.M., De La Fuente R. (2020) **Helicase LSH/Hells regulates kinetochore function, histone H3/Thr3 phosphorylation and centromere transcription during oocyte meiosis** *Nature communications* **11**

Beaulaurier J., Schadt E.E., Fang G (2019) **Deciphering bacterial epigenomes using modern sequencing technologies** *Nat Rev Genet* **20**:157–172

Bewick A.J., Hofmeister B.T., Powers R.A., Mondo S.J., Grigoriev I.V., James T.Y., Stajich J.E., Schmitz R.J (2019) **Diversity of cytosine methylation across the fungal tree of life** *Nat Ecol Evol* **3**:479–490

Bewick A.J., Niederhuth C.E., Ji L., Rohr N.A., Griffin P.T., Leebens-Mack J., Schmitz R.J (2017) **The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants** *Genome Biol* **18**

Bewick A.J., Sanchez Z., McKinney E.C., Moore A.J., Moore P.J., Schmitz R.J (2019) **Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, Oncopeltus fasciatus** *Epigenetics Chromatin* **12**

Bewick A.J., Vogel K.J., Moore A.J., Schmitz R.J (2017) **Evolution of DNA Methylation across Insects** *Mol Biol Evol* **34**:654–665

Birtle Z., Ponting C.P (2006) **Meisetz and the birth of the KRAB motif** *Bioinformatics* **22**:2841–2845

Bonasio R., Li Q., Lian J., Mutti N.S., Jin L., Zhao H., Zhang P., Wen P., Xiang H., Ding Y., et al. (2012) **Genome-wide and caste-specific DNA methylomes of the ants Camponotus floridanus and Harpegnathos saltator** *Current biology : CB* **22**:1755–1764

Cao X., Springer N.M., Muszynski M.G., Phillips R.L., Kaeppler S., Jacobsen S.E. (2000) **Conserved plant genes with similarity to mammalian de novo DNA methyltransferases** *Proc Natl Acad Sci U S A* **97**:4979–4984

Casadesus J., Low D (2006) **Epigenetic gene regulation in the bacterial world** *Microbiol Mol Biol Rev* **70**:830–856

Catania S., Dumesic P.A., Pimentel H., Nasif A., Stoddard C.I., Burke J.E., Diedrich J.K., Cooke S., Shea T., Gienger E., et al. (2020) **Evolutionary Persistence of DNA Methylation for Millions of Years after Ancient Loss of a De Novo Methyltransferase** *Cell* **180**

Chan S.W., Henderson I.R., Jacobsen S.E (2005) **Gardening the genome: DNA methylation in Arabidopsis thaliana** *Nat Rev Genet* **6**:351–360

Chen K., Ou X.M., Chen G., Choi S.H., Shih J.C (2005) **R1, a novel repressor of the human monoamine oxidase A** *The Journal of biological chemistry* **280**:11552–11559

Choi J., Lyons D.B., Kim M.Y., Moore J.D., Zilberman D (2020) **DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts** *Mol Cell* **77**:310–323

Cohen O., Ashkenazy H., Burstein D., Pupko T (2012) **Uncovering the co-evolutionary network among prokaryotic genes** *Bioinformatics* **28**:i389–i394

Cohen O., Ashkenazy H., Levy Karin E., Burstein D., Pupko T (2013) **CoPAP: Coevolution of presence-absence patterns** *Nucleic Acids Res* **41**:W232–237

Colot V., Rossignol J.L (1999) **Eukaryotic DNA methylation as an evolutionary device** *Bioessays* **21**:402–411

Czech B., Munafo M., Ciabrelli F., Eastwood E.L., Fabry M.H., Kneuss E., Hannon G.J (2018) **piRNA-Guided Genome Defense: From Biogenesis to Silencing** *Annu Rev Genet* **52**:131–157

de Greef J.C., Wang J., Balog J., den Dunnen J.T., Frants R.R., Straasheijm K.R., Aytekin C., van der Burg M., Duprez L., Ferster A., et al. (2011) **Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2** *Am J Hum Genet* **88**:796–804

de Mendoza A., Bonnet A., Vargas-Landin D.B., Ji N., Li H., Yang F., Li L., Hori K., Pflueger J., Buckberry S., et al. (2018) **Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons** *Nature communications* **9**

Deleris A., Halter T., Navarro L. (2016) **DNA Methylation and Demethylation in Plant Immunity** *Annu Rev Phytopathol* **54**:579–603

Deng S., Jang I.C., Su L., Xu J., Chua N.H (2016) **JMJ24 targets CHROMOMETHYLASE3 for proteasomal degradation in Arabidopsis** *Genes Dev* **30**:251–256

Deniz O., Frost J.M., Branco M.R (2019) **Regulation of transposable elements by DNA modifications** *Nat Rev Genet* **20**:417–431

Dennis K., Fan T., Geiman T., Yan Q., Muegge K (2001) **Lsh, a member of the SNF2 family, is required for genome-wide methylation** *Genes Dev* **15**:2940–2944

Dimitriu T., Szczelkun M.D., Westra E.R (2020) **Evolutionary Ecology and Interplay of Prokaryotic Innate and Adaptive Immune Systems** *Current biology : CB* **30**:R1189–R1202

Dong J., Gao Z., Liu S., Li G., Yang Z., Huang H., Xu L (2013) **SLIDE, the protein interacting domain of Imitation Switch remodelers, binds DDT-domain proteins of different subfamilies in chromatin remodeling complexes** *J Integr Plant Biol* **55**:928–937

Dumesic P.A., Stoddard C.I., Catania S., Narlikar G.J., Madhani H.D. (2020) **ATP Hydrolysis by the SNF2 Domain of Dnmt5 Is Coupled to Both Specific Recognition and Modification of Hemimethylated DNA** *Mol Cell* **79**:127–139

Dunican D.S., Cruickshanks H.A., Suzuki M., Semple C.A., Davey T., Arceci R.J., Greally J., Adams I.R., Meehan R.R (2013) **Lsh regulates LTR retrotransposon repression independently of Dnmt3b function** *Genome Biol* **14**

Dunican D.S., Pennings S., Meehan R.R (2015) **Lsh Is Essential for Maintaining Global DNA Methylation Levels in Amphibia and Fish and Interacts Directly with Dnmt1** *Biomed Res Int* **2015**

Edgar R.C. (2022) **High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny** *bioRxiv*

Ehrlich M (2003) **The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease** *Clin Immunol* **109**:17–28

Ehrlich M., Jackson K., Weemaes C (2006) **Immunodeficiency, centromeric region instability, facial anomalies syndrome (ICF)** *Orphanet J Rare Dis* **1**

Engelhardt J., Scheer O., Stadler P.F., Prohaska S.J (2022) **Evolution of DNA Methylation Across Ecdysozoa** *J Mol Evol* **90**:56–72

Felle M., Hoffmeister H., Rothammer J., Fuchs A., Exler J.H., Langst G (2011) **Nucleosomes protect DNA from DNA methylation in vivo and in vitro** *Nucleic Acids Res* **39**:6956–6969

Feng S., Cokus S.J., Zhang X., Chen P.Y., Bostick M., Goll M.G., Hetzel J., Jain J., Strauss S.H., Halpern M.E., et al. (2010) **Conservation and divergence of methylation patterning in plants and animals** *Proc Natl Acad Sci U S A* **107**:8689–8694

Feng Z., Wu Y., Yang C., Gu X., Wilson J.J., Li H., Cai W., Yang H., Song F (2020) **Evolution of tRNA gene rearrangement in the mitochondrial genome of ichneumonoid wasps (Hymenoptera: Ichneumonoidea)** *Int J Biol Macromol* **164**:540–547

Flaus A., Martin D.M., Barton G.J., Owen-Hughes T (2006) **Identification of multiple distinct Snf2 subfamilies with conserved structural motifs** *Nucleic Acids Res* **34**:2887–2905

Gill R.M., Gabor T.V., Couzens A.L., Scheid M.P (2013) **The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation** *Mol Cell Biol* **33**:498–513

Greenberg M.V.C., Bourc'his D (2019) **The diverse roles of DNA methylation in mammalian development and disease** *Nat Rev Mol Cell Biol* **20**:590–607

Grogan D.W (2003) **Cytosine methylation by the SuaI restriction-modification system: implications for genetic fidelity in a hyperthermophilic archaeon** *J Bacteriol* **185**:4657–4661

Han M., Li J., Cao Y., Huang Y., Li W., Zhu H., Zhao Q., Han J.J., Wu Q., Li J., et al. (2020) **A role for LSH in facilitating DNA methylation by DNMT1 through enhancing UHRF1 chromatin association** *Nucleic Acids Res* **48**:12116–12134

Hansen R.S., Wijmenga C., Luo P., Stanek A.M., Canfield T.K., Weemaes C.M., Gartler S.M (1999) **The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome** *Proc Natl Acad Sci U S A* **96**:14412–14417

Hardikar S., Ying Z., Zeng Y., Zhao H., Liu B., Veland N., McBride K., Cheng X., Chen T (2020) **The ZBTB24-CDCA7 axis regulates HELLS enrichment at centromeric satellite repeats to facilitate DNA methylation** *Protein Cell* **11**:214–218

Hayashi M., Sugahara K., Yamamura A., Iida Y. (2021) **Evaluation of the Properties of the DNA Methyltransferase from Aeropyrum pernix K1** *Microbiol Spectr* **9**

Hemmi H., Takeuchi O., Kawai T., Kaisho T., Sato S., Sanjo H., Matsumoto M., Hoshino K., Wagner H., Takeda K., et al. (2000) **A Toll-like receptor recognizes bacterial DNA** *Nature* **408**:740–745

Higgins D.G., Sharp P.M (1988) **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer** *Gene* **73**:237–244

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. (2018) **UFBoot2: Improving the Ultrafast Bootstrap Approximation** *Mol Biol Evol* **35**:518–522

Honda S., Lewis Z.A., Huarte M., Cho L.Y., David L.L., Shi Y., Selker E.U. (2010) **The DMM complex prevents spreading of DNA methylation from transposons to nearby genes in Neurospora crassa** *Genes Dev* **24**:443–454

Huang A., Ho C.S., Ponzielli R., Barsyte-Lovejoy D., Bouffet E., Picard D., Hawkins C.E., Penn L.Z (2005) **Identification of a novel c-Myc protein interactor, JPO2, with transforming activity in medulloblastoma cells** *Cancer Res* **65**:5607–5619

Huang J., Fan T., Yan Q., Zhu H., Fox S., Issaq H.J., Best L., Gangi L., Munroe D., Muegge K (2004) **Lsh, an epigenetic guardian of repetitive elements** *Nucleic Acids Res* **32**:5019–5028

Huff J.T., Zilberman D (2014) **Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes** *Cell* **156**:1286–1297

Imai Y., Biot M., Clement J.A., Teragaki M., Urbach S., Robert T., Baudat F., Grey C., de Massy B. (2020) **PRDM9 activity depends on HELLS and promotes local 5-hydroxymethylcytosine enrichment** *Elife* **9**

Ishikawa K., Watanabe M., Kuroita T., Uchiyama I., Bujnicki J.M., Kawakami B., Tanokura M., Kobayashi I (2005) **Discovery of a novel restriction endonuclease by genome comparison and application of a wheat-germ-based cell-free translation assay: PabI (5'-GTA/C) from the hyperthermophilic archaeon Pyrococcus abyssi** *Nucleic Acids Res* **33**

Ivasyk I., Olivos-Cisneros L., Valdes-Rodriguez S., Droual M., Jang H., Schmitz R.J., Kronauer D.J.C (2023) **DNMT1 mutant ants develop normally but have disrupted oogenesis** *Nature communications* **14**

Iyer L.M., Abhiman S., Aravind L (2011) **Natural history of eukaryotic DNA methylation systems** *Prog Mol Biol Transl Sci* **101**:25–104

Jamieson K., Rountree M.R., Lewis Z.A., Stajich J.E., Selker E.U (2013) **Regional control of histone H3 lysine 27 methylation in Neurospora** *Proc Natl Acad Sci U S A* **110**:6027–6032

Jenness C., Giunta S., Muller M.M., Kimura H., Muir T.W., Funabiki H (2018) **HELLS and CDCA7 comprise a bipartite nucleosome remodeling complex defective in ICF syndrome** *Proc Natl Acad Sci U S A* **115**:E876–E885

Jiang Y.L., Rigolet M., Bourc'his D., Nigon F., Bokesoy I., Fryns J.P., Hulten M., Jonveaux P., Maraschio P., Megarbane A., et al. (2005) **DNMT3B mutations and DNA methylation defect define two types of ICF syndrome** *Hum Mutat* **25**:56–63

Kato M., Miura A., Bender J., Jacobsen S.E., Kakutani T (2003) **Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis** *Current biology : CB* **13**:421–426

Kondilis-Mangum H.D., Wade P.A (2013) **Epigenetics and the adaptive immune response** *Mol Aspects Med* **34**:813–825

Kyger R., Luzuriaga-Neira A., Layman T., Milkewitz Sandberg T.O., Singh D., Huchon D., Peri S., Atkinson S.D., Bartholomew J.L., Yi S.V., et al. (2021) **Myxosporea (Myxozoa, Cnidaria) Lack DNA Cytosine Methylation** *Mol Biol Evol* **38**:393–404

Lewis S.H., Ross L., Bain S.A., Pahita E., Smith S.A., Cordaux R., Miska E.A., Lenhard B., Jiggins F.M., Sarkies P. (2020) **Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods** *PLoS Genet* **16**

Li B., Du Z., Tian L., Zhang L., Huang Z., Wei S., Song F., Cai W., Yu Y., Yang H., et al. (2021) **Chromosome-level genome assembly of the aphid parasitoid Aphidius gifuensis using Oxford Nanopore sequencing and Hi-C technology** *Mol Ecol Resour* **21**:941–954

Li B., Hou L., Zhu D., Xu X., An S., Wang X (2018) **Identification and caste-dependent expression patterns of DNA methylation associated genes in Bombus terrestris** *Sci Rep* **8**

Li D., Liu J., Liu W., Li G., Yang Z., Qin P., Xu L (2017) **The ISWI remodeler in plants: protein complexes, biochemical functions, and developmental roles** *Chromosoma* **126**:365–373

Libbrecht R., Oxley P.R., Keller L., Kronauer D.J (2016) **Robust DNA Methylation in the Clonal Raider Ant Brain** *Current biology : CB* **26**:391–395

Litwin I., Bakowski T., Maciaszczyk-Dziubinska E., Wysocki R. (2017) **The LSH/HELLS homolog Irc5 contributes to cohesin association with chromatin in yeast** *Nucleic Acids Res* :6404–6416

Long H.K., Blackledge N.P., Klose R.J (2013) **ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection** *Biochem Soc Trans* **41**:727–740

Lowe R., Barton C., Jenkins C.A., Ernst C., Forman O., Fernandez-Twinn D.S., Bock C., Rossiter S.J., Faulkes C.G., Ozanne S.E., et al. (2018) **Ageing-associated DNA methylation dynamics are a molecular readout of lifespan variation among mammalian species** *Genome Biol* **19**

Lu S., Wang J., Chitsaz F., Derbyshire M.K., Geer R.C., Gonzales N.R., Gwadz M., Hurwitz D.I., Marchler G.H., Song J.S., et al. (2020) **CDD/SPARCLE: the conserved domain database in 2020** *Nucleic Acids Res* **48**:D265–D268

Lyko F (2018) **The DNA methyltransferase family: a versatile toolkit for epigenetic regulation** *Nat Rev Genet* **19**:81–92

Lyons D.B., Zilberman D (2017) **DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes** *Elife* **6**

Maertens G.N., Cherepanov P., Engelman A (2006) **Transcriptional co-activator p75 binds and tethers the Myc-interacting protein JPO2 to chromatin** *J Cell Sci* **119**:2563–2571

Margueron R., Reinberg D (2011) **The Polycomb complex PRC2 and its mark in life** *Nature* **469**:343–349

Militello K.T., Wang P., Jayakar S.K., Pietrasik R.L., Dupont C.D., Dodd K., King A.M., Valenti P.R (2008) **African trypanosomes contain 5-methylcytosine in nuclear DNA** *Eukaryot Cell* **7**:2012–2016

Ming X., Zhang Z., Zou Z., Lv C., Dong Q., He Q., Yi Y., Li Y., Wang H., Zhu B (2020) **Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration** *Cell Res* **30**:980–996

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. (2020) **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era** *Mol Biol Evol* **37**:1530–1534

Miro-Pina C., Charmant O., Kawaguchi T., Holoch D., Michaud A., Cohen I., Humbert A., Jaszczyszyn Y., Chevreux G., Del Maestro L., et al. (2022) **Paramecium Polycomb repressive complex 2 physically interacts with the small RNA-binding PIWI protein to repress transposable elements** *Developmental cell* **57**:1037–1052

Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., et al. (2021) **Pfam: The protein families database in 2021** *Nucleic Acids Res* **49**:D412–D419

Miura A., Yonebayashi S., Watanabe K., Toyama T., Shimada H., Kakutani T (2001) **Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis** *Nature* **411**:212–214

Mondo S.J., Dannebaum R.O., Kuo R.C., Louie K.B., Bewick A.J., LaButti K., Haridas S., Kuo A., Salamov A., Ahrendt S.R., et al. (2017) **Widespread adenine N6-methylation of active genes in fungi** *Nat Genet* **49**:964–968

Moon E.K., Hong Y., Lee H.A., Quan F.S., Kong H.H (2017) **DNA Methylation of Gene Expression in Acanthamoeba castellanii Encystation** *Korean J Parasitol* **55**:115–120

Mulholland C.B., Nishiyama A., Ryan J., Nakamura R., Yigit M., Gluck I.M., Trummer C., Qin W., Bartoschek M.D., Traube F.R., et al. (2020) **Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals** *Nature communications* **11**

Myant K., Termanis A., Sundaram A.Y., Boe T., Li C., Merusi C., Burrage J., de Las Heras J.I., Stancheva I. (2011) **LSH and G9a/GLP complex are required for developmentally programmed DNA methylation** *Genome Res* **21**:83–94

Nai Y.S., Huang Y.C., Yen M.R., Chen P.Y (2020) **Diversity of Fungal DNA Methyltransferases and Their Association With DNA Methylation Patterns** *Front Microbiol* **11**

Neri F., Rapelli S., Krepelova A., Incarnato D., Parlato C., Basile G., Maldotti M., Anselmi F., Oliviero S (2017) **Intragenic DNA methylation prevents spurious transcription initiation** *Nature* **543**:72–77

Ni K., Muegge K (2021) **LSH catalyzes ATP-driven exchange of histone variants macroH2A1 and macroH2A2** *Nucleic Acids Res* **49**:8024–8036

Ni K., Ren J., Xu X., He Y., Finney R., Braun S.M.G., Hathaway N.A., Crabtree G.R., Muegge K (2020) **LSH mediates gene repression through macroH2A deposition** *Nature communications* **11**

Nishiyama A., Mulholland C.B., Bultmann S., Kori S., Endo A., Saeki Y., Qin W., Trummer C., Chiba Y., Yokoyama H., et al. (2020) **Two distinct modes of DNMT1 recruitment ensure stable maintenance DNA methylation** *Nature communications* **11**

Nishiyama A., Nakanishi M (2021) **Navigating the DNA methylation landscape of cancer** *Trends Genet* **37**:1012–1027

Okano M., Bell D.W., Haber D.A., Li E (1999) **DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development** *Cell* **99**:247–257

Onishi R., Yamanaka S., Siomi M.C (2021) **piRNA- and siRNA-mediated transcriptional repression in Drosophila, mice, and yeast: new insights and biodiversity** *EMBO Rep* **22**

Osakabe A., Jamge B., Axelsson E., Montgomery S.A., Akimcheva S., Kuehn A.L., Pisupati R., Lorkovic Z.J., Yelagandula R., Kakutani T., et al. (2021) **The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W** *Nat Cell Biol* **23**:391–400

Ou X.M., Chen K., Shih J.C (2006) **Glucocorticoid and androgen activation of monoamine oxidase A is regulated differently by R1 and Sp1** *The Journal of biological chemistry* **281**:21512–21525

Paysan-Lafosse T., Blum M., Chuguransky S., Grego T., Pinto B.L., Salazar G.A., Bileschi M.L., Bork P., Bridge A., Colwell L., et al. (2023) **InterPro in 2022** *Nucleic Acids Res* **51**:D418–D427

Peters R.S., Krogmann L., Mayer C., Donath A., Gunkel S., Meusemann K., Kozlov A., Podsiadlowski L., Petersen M., Lanfear R., et al. (2017) **Evolutionary History of the Hymenoptera** *Current biology : CB* **27**:1013–1018

Pinto B.J., Weis J.J., Gamble T., Ode P.J., Paul R., Zaspel J.M (2021) **A Chromosome-Level Genome Assembly of the Parasitoid Wasp, Cotesia glomerata (Hymenoptera: Braconidae)** *J Hered* **112**:558–564

Ponger L., Li W.H (2005) **Evolutionary diversification of DNA methyltransferases in eukaryotic genomes** *Mol Biol Evol* **22**:1119–1128

Prangishvili D.A., Vashakidze R.P., Chelidze M.G., Gabriadze I (1985) **A restriction endonuclease SuaI from the thermoacidophilic archaebacterium Sulfolobus acidocaldarius** *FEBS Lett* **192**:57–60

Pratt R.J., Lee D.W., Aramayo R (2004) **DNA methylation affects meiotic trans-sensing, not meiotic silencing, in Neurospora** *Genetics* **168**:1925–1935

Ren J., Briones V., Barbour S., Yu W., Han Y., Terashima M., Muegge K (2015) **The ATP binding site of the chromatin remodeling homolog Lsh is required for nucleosome density and de novo DNA methylation at repeat sequences** *Nucleic Acids Res* **43**:1444–1455

Robertson K.D (2005) **DNA methylation and human disease** *Nat Rev Genet* **6**:597–610

Saze H., Shiraishi A., Miura A., Kakutani T (2008) **Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana** *Science* **319**:462–465

Schofield M.J., Hsieh P (2003) **DNA mismatch repair: molecular mechanisms and biological function** *Annu Rev Microbiol* **57**:579–608

Schulz N.K.E., Wagner C.I., Ebeling J., Raddatz G., Diddens-de Buhr M.F., Lyko F., Kurtz J. (2018) **Dnmt1 has an essential function despite the absence of CpG DNA methylation in the red flour beetle Tribolium castaneum** *Sci Rep* **8**

Sharaf A., Vijayanathan M., Obornik M., Mozgova I (2022) **Phylogenetic profiling resolves early emergence of PRC2 and illuminates its functional core** *Life Sci Alliance* **5**

Spruce C., Dlamini S., Ananda G., Bronkema N., Tian H., Paigen K., Carter G.W., Baker C.L. (2020) **HELLS and PRDM9 form a pioneer complex to open chromatin at meiotic recombination hot spots** *Genes Dev* **34**:398–412

Takada Y., Yaman-Deveci R., Shirakawa T., Sharif J., Tomizawa S.I., Miura F., Ito T., Ono M., Nakajima K., Koseki Y., et al. (2021) **Maintenance DNA methylation in pre-meiotic germ cells regulates meiotic prophase by facilitating homologous chromosome pairing** *Development* **148**

Tan L.M., Liu R., Gu B.W., Zhang C.J., Luo J., Guo J., Wang Y., Chen L., Du X., Li S., et al. (2020) **Dual Recognition of H3K4me3 and DNA by the ISWI Component ARID5 Regulates the Floral Transition in Arabidopsis** *Plant Cell* **32**:2178–2195

Tesina P., Cermakova K., Horejsi M., Prochazkova K., Fabry M., Sharma S., Christ F., Demeulemeester J., Debyser Z., Rijck J., et al. (2015) **Multiple cellular proteins interact with LEDGF/p75 through a conserved unstructured consensus motif** *Nature communications* **6**

Thijssen P.E., Ito Y., Grillo G., Wang J., Velasco G., Nitta H., Unoki M., Yoshihara M., Suyama M., Sun Y., et al. (2015) **Mutations in CDCA7 and HELLS cause immunodeficiency-centromeric instability-facial anomalies syndrome** *Nature communications* **6**

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

Tirot L., Jullien P.E., Ingouff M (2021) **Evolution of CG Methylation Maintenance Machinery in Plants** *Epigenomes* **5**

Unoki M (2021) **Chromatin remodeling in replication-uncoupled maintenance DNA methylation and chromosome stability: Insights from ICF syndrome studies** *Genes Cells* **26**:349–359

Unoki M., Funabiki H., Velasco G., Francastel C., Sasaki H (2019) **CDCA7 and HELLS mutations undermine nonhomologous end joining in centromeric instability syndrome** *J Clin Invest* **129**:78–92

Unoki M., Sharif J., Saito Y., Velasco G., Francastel C., Koseki H., Sasaki H (2020) **CDCA7 and HELLS suppress DNA:RNA hybrid-associated DNA damage at pericentromeric repeats** *Sci Rep* **10**

Vasu K., Nagaraja V (2013) **Diverse functions of restriction-modification systems in addition to cellular defense** *Microbiol Mol Biol Rev* **77**:53–72

Velasco G., Grillo G., Touleimat N., Ferry L., Ivkovic I., Ribierre F., Deleuze J.F., Chantalat S., Picard C., Francastel C (2018) **Comparative methylome analysis of ICF patients identifies heterochromatin loci that require ZBTB24, CDCA7 and HELLS for their methylated state** *Hum Mol Genet*

Vongs A., Kakutani T., Martienssen R.A., Richards E.J (1993) **Arabidopsis thaliana DNA methylation mutants** *Science* **260**:1926–1928

Vukic M., Daxinger L (2019) **DNA methylation in disease: Immunodeficiency, Centromeric instability, Facial anomalies syndrome** *Essays Biochem* **63**:773–783

Walker M.G. (2001) **Drug target discovery by gene expression analysis: cell cycle genes** *Curr Cancer Drug Targets* **1**:73–83

Wang T., Ma J., Hogan A.N., Fong S., Licon K., Tsui B., Kreisberg J.F., Adams P.D., Carvunis A.R., Bannasch D.L., et al. (2020) **Quantitative Translation of Dog-to-Human Aging by Conserved Remodeling of the DNA Methylome** *Cell Syst* **11**:176–185

Wang X., Wheeler D., Avery A., Rago A., Choi J.H., Colbourne J.K., Clark A.G., Werren J.H (2013) **Function and evolution of DNA methylation in Nasonia vitripennis** *PLoS Genet* **9**

Ward N., Moreno-Hagelsieb G (2014) **Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss?** *PloS one* **9**

Washington J.T., Cavender K.R., Amukamara A.U., McKinney E.C., Schmitz R.J., Moore P.J (2021) **The essential role of Dnmt1 in gametogenesis in the large milkweed bug Oncopeltus fasciatus** *Elife* **10**

Wu H., Thijssen P.E., de Klerk E., Vonk K.K., Wang J., den Hamer B., Aytekin C., van der Maarel S.M., Daxinger L. (2016) **Converging disease genes in ICF syndrome: ZBTB24 controls expression of CDCA7 in mammals** *Hum Mol Genet* **25**:4041–4051

Xu X., Ni K., He Y., Ren J., Sun C., Liu Y., Aladjem M.I., Burkett S., Finney R., Ding X., et al. (2021) **The epigenetic regulator LSH maintains fork protection and genomic stability via MacroH2A deposition and RAD51 filament formation** *Nature communications* **12**

Yaari R., Katz A., Domb K., Harris K.D., Zemach A., Ohad N (2019) **RdDM-independent de novo and heterochromatin DNA methylation by plant CMT and DNMT3 orthologs** *Nature communications* **10**

Yamada K., Frouws T.D., Angst B., Fitzgerald D.J., DeLuca C., Schimmele K., Sargent D.F., Richmond T.J (2011) **Structure and mechanism of the chromatin remodelling factor ISW1a** *Nature* **472**:448–453

Yu W., McIntosh C., Lister R., Zhu I., Han Y., Ren J., Landsman D., Lee E., Briones V., Terashima M., et al. (2014) **Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways** *Genome Res* **24**:1613–1623

Zemach A., Kim M.Y., Hsieh P.H., Coleman-Derr D., Eshed-Williams L., Thao K., Harmer S.L., Zilberman D (2013) **The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin** *Cell* **153**:193–205

Zemach A., McDaniel I.E., Silva P., Zilberman D (2010) **Genome-wide evolutionary analysis of eukaryotic DNA methylation** *Science* **328**:916–919

Zeng W., Baumann C., Schmidtmann A., Honaramooz A., Tang L., Bondareva A., Dores C., Fan T., Xi S., Geiman T., et al. (2011) **Lymphoid-specific helicase (HELLS) is essential for meiotic progression in mouse spermatocytes** *Biol Reprod* **84**:1235–1241

Zhang C., Tian Y., Song S., Zhang L., Dang Y., He Q. (2022) **H3K56 deacetylation and H2A.Z deposition are required for aberrant heterochromatin spreading** *Nucleic Acids Res* **50**:3852–3866

Author information

Hironori Funabiki
Laboratory of Chromosome and Cell Biology, The Rockefeller University, New York, NY 10065
For correspondence: funabih@rockefeller.edu
ORCID iD: 0000-0003-4831-4087

Isabel E. Wassing
Laboratory of Chromosome and Cell Biology, The Rockefeller University, New York, NY 10065
ORCID iD: 0000-0001-5937-7273

Qingyuan Jia
Laboratory of Chromosome and Cell Biology, The Rockefeller University, New York, NY 10065

Ji-Dung Luo
Bioinformatics Resource Center, The Rockefeller University, New York, NY 10065

Thomas Carroll
Bioinformatics Resource Center, The Rockefeller University, New York, NY 10065

Editors

Reviewing Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Germany

Senior Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Germany

Reviewer #1 (Public Review):

Funabiki et al, performed a co-evolutionary analysis of Lsh/HELLS and CDCA7, two factors with links to DNA methylation pathways in mammals, amphibia and fish. The authors suggest that conserved roles for the two factors in DNA methylation maintenance pathways can be traced back to the last eukaryotic common ancestor. Overall, the findings are important and the results could be useful for researchers studying DNA methylation pathways in many different organisms.

Comments on current version:

In the revised version of this manuscript the authors addressed all previously raised issues. I would like to thank them for that. The data is now clearly presented and interpreted and more experimental detail has been added. Thus, the manuscript is much improved and provides an interesting basis for experimental follow-up and further functional investigations.

Reviewer #2 (Public Review):

In this manuscript, Funabiki and colleagues investigated the co-evolution of DNA methylation and nucleosome remolding in eukaryotes. This study is motivated by several observations: (1) despite being ancestrally derived, many eukaryotes lost DNA methylation and/or DNA methyltransferases; (2) over many genomic loci, the establishment and maintenance of DNA methylation relies on a conserved nucleosome remodeling complex composed of CDCA7 and HELLS; (3) it remains unknown if/how this functional link influenced the evolution of DNA methylation. The authors hypothesize that if CDCA7-HELLS function was required for DNA methylation in the last eukaryote common ancestor, this should be accompanied by signatures of co-evolution during eukaryote radiation.

To test this hypothesis, they first set out to investigate the presence/absence of putative functional orthologs of CDCA7, HELLS and DNMTs across major eukaryotic clades. They succeed in identifying homologs of these genes in all clades spanning 180 species. To annotate putative functional orthologs, they use similarity over key functional domains and residues - such as ICF related mutations for CDCA7 and SNF2 domains for HELLS - as well as maximum likelihood phylogenetic analyses. Using established eukaryote phylogenies, the authors conclude that the CDCA7-HELLS-DNMT axis arose in the last common ancestor to all eukaryotes. Importantly, they found recurrent loss events of CDCA7-HELLS-DNMT in at least 40 eukaryotic species, most of them lacking DNA methylation.

Having identified these factors, they successfully identify signatures of co-evolution between DNMTs, CDCA7 and HELLS using CoPAP analysis - a probabilistic model inferring the likelihood of interactions between genes given a set of presence/absence patterns. As a control, such interactions are not detected with other remodelers or chromatin modifying pathways also found across eukaryotes. Expanding on this analysis, the authors found that CDCA7 was more likely to be lost in species without DNA methylation.

In conclusion, the authors suggest that the CDCA7-HELLS-DNMT axis is ancestral in eukaryotes and raise the hypothesis that CDCA7 becomes quickly dispensable upon the loss of DNA methylation and/or that CDCA7 might be the first step toward the switch from DNA methylation-based genome regulation to other modes.

The data and analyses reported are significant and solid. Overall, this work is a conceptual advance in our understanding of the evolutionary coupling between nucleosome remolding and DNA methylation. It also provides a useful resource to study the early origins of DNA methylation related molecular process. Finally, it brings forward the interesting hypothesis that since eukaryotes are faced with the challenge of performing DNA methylation in the context of nucleosome packed DNA, loosing factors such as CDCA7-HELLS likely led to recurrent innovations in chromatin-based genome regulation.

Strengths:
- The hypothesis linking nucleosome remodeling and the evolution of DNA methylation.
- Deep mapping of DNA methylation related process in eukaryotes.
- Identification and evolutionary trajectories of novel homologs/orthologs of CDCA7.
- Identification of CDCA7-HELLS-DNMT co-evolution across eukaryotes.

Author Response

The following is the authors' response to the original reviews.

eLife assessment

This important manuscript reveals signatures of co-evolution of two nucleosome remodeling factors, Lsh/HELLS and CDCA7, which are involved in the regulation of eukaryotic DNA methylation. The results suggest that the roles for the two factors in DNA methylation maintenance pathways can be traced back to the last eukaryotic common ancestor and that the CDC7A-HELLS-DNMT axis shaped the evolutionary retention of DNA methylation in eukaryotes. The evolutionary analyses are solid, although more refined phylogenetic approaches could have strengthened some of the claims. Overall, this study should be useful for researchers studying DNA methylation pathways in different organisms, and it should be of general interest to colleagues in the fields of evolutionary biology, chromatin biology and genome biology.

We sincerely appreciate constructive comments and suggestions by the reviewers and a fair and accurate summary by the monitoring editor. Below we made point-by-point responses to reviewers' comments.

Reviewer #1 (Public Review):

Overall, I find the work performed by the authors very interesting. However, the authors have not always included literature that seems relevant to their study. For instance, I do not understand why two papers Dunican et al 2013 and Dunican et al 2015, which provide important insight into Lsh/HELLS function in mouse, frog and fish were not cited. It is also important that the authors are specific about what is known and in particular about what is not known about CDCA7 function in DNA methylation regulation. Unless I am mistaken, there is currently only one study (Velasco et al 2018) investigating the effect of CDCA7 disruption on DNA methylation levels (in ICF3 patient lymphoblastoid cell lines) on a genome-wide scale (Illumina 450K arrays). Unoki et al 2019 report that CDCA7 and HELLS gene knockout in human HEK293T cells moderately and extremely reduces DNA methylation levels at pericentromeric satellite-2 and centromeric alpha-satellite repeats, respectively. No other loci were investigated, and it is therefore not known whether a CDCA7-associated maintenance methylation phenotype extends beyond (peri)centromeric satellites. Thijssen et al performed siRNA- mediated knockdown experiments in mouse embryonic fibroblasts (differentiated cells) and showed that lower levels of Zbtb24, Cdca7 and Hells protein correlate with reduced minor satellite repeat methylation, thereby implicating these factors in mouse minor satellite repeat DNA methylation maintenance. Furthermore, studies that demonstrate a HELLS-CDCA7 interaction are currently limited to Xenopus egg extract (Jenness et al 2018) and the human HEK293 cell line (Unoki et al 2019). Whether such an interaction exists in any other organism and is of relevance to DNA methylation mechanisms remains to be determined. Therefore, in my opinion, the conclusion that "Our co- evolution analysis suggests that DNA methylation-related functionalities of CDCA7 and HELLS are inherited from LECA" should be softened, as the evidence for this scenario is not very compelling and seems premature in the absence of molecular data from more species.

We appreciate this reviewer's thorough reading of our manuscript.

Regarding the citation issues, we will cite Dunican 2013 and Dunican 2015. In addition, we went through the manuscript to update the citations.

As pointed out by the reviewer, the role of CDCA7 in genome DNA methylation was extensively studied in Velasco et al 2018. The result, together with Thijssen et al (2015), and Unoki et al. (2018), supports the idea that ZBTB24, CDCA7 and HELLS act within the same pathway to promote DNA methylation, the pattern of which is overlapping but distinct from DNMT3B-mediated methylation. This observation suggests that a ZBTB24- CDCA7-HELLS

mechanism for DNA methylation may involve an alternative DNMT. Interestingly, our analysis of the gene presence-absence pattern revealed that the presence of CDCA7 coincides with DNMT1 more than DNMT3 genes. Indeed, while CDCA7 is lost from diverse branches of eukaryote species, genomes encoding CDCA7 always encode HELLS, and almost always encode DNMT1. Based on this observation, we speculate the role of CDCA7 is tightly linked to HELLS and DNA methylation throughout evolution.

As pointed out by Reviewer 1, the link between CDCA7, HELLS and DNA methylation has not been determined experimentally across these species. However, based on our previously published and unpublished data, we are confident about the functional interaction between CDCA7 and HELLS in Xenopus laevis and Homo sapiens.

Furthermore, the importance of HELLS homologs in DNA methylation has been extensively studied in human, mice and plants. We hope our current study will motivate the field to experimentally test the evolutionary conservation of HELLS-CDCA7 interaction, as well as their importance in DNA methylation, in other species.

> *The authors used BLAST searches to characterize the evolutionary conservation of CDCA7 family proteins in vertebrates. From Figure 2A, it seems that they identify a LEDGF binding motif in CDCA7/JPO1. Is this correct and if yes, could you please elaborate and show this result? This is interesting and important to clarify because previous literature (Tesina et al 2015) reports a LEDGF binding motif only in CDCA7L/JPO2.*

We searched for a LEDGF binding motif ({E/D}-X-E-X-F-X-G-F, also known as IBM described in Tesina et al 2015) in vertebrate CDCA7 proteins, and reported their positions in Figure 2A. Examples of identified LEDGF-binding motifs are now presented in Fig. 2C.

> *To provide evidence for a potential evolutionary co-selection of CDCA7, HELLS and the DNA methyltransferases (DNMTs) the authors performed CoPAP analysis. Throughout the manuscript, it is unclear to me what the authors mean when referring to "DNMT3". In the Material and Methods section, the authors mention that human DNMT3A was used in BLAST searches to identify proteins with DNA methyltransferase domains. Does this mean that "DNMT3" should be DNMT3A? And if yes, should "DNMT3" be corrected to "DNMT3A"? Is there a reason that "DNMT3A" was chosen for the BLAST searches?*

As described in the Methods section, both Human DNMT1 and DNMT3A were used to initially identify any proteins containing a domain homologous to the DNA methyltransferase catalytic domain. Within Metazoa, if their orthologs exist, the top hit from BLAST search using human DNMT1 and DNMT3A show E-value 0.0, and thus their orthology is robust. This is even true for DNMT1 and DNMT3 homologs in the sponge Amphimedon queenslandica, which is one of the earliest-branching metazoan species. For other DNMTs, such as DNMT2, DNMT4, DNMT5, DNMT6, we conducted separate BLAST searches using those proteins as baits as described in Methods. The methyltransferase domain was then isolated using the NCBI conserved domains search. The selected DNMT domain sequences were aligned with CLUSTALW to generate a phylogenetic tree to further classify DNMTs. In response to reviewer #2's comments, we also generated another multi-sequence alignment of DNMTs using MUSCLE v5 and conducted maximum-likelihood-based phylogenetic tree assembly using IQ-TREE (new Fig. S6). The overall topology of these trees is consistent except for orphan DNMTs. It has been suggested that vertebrate DNMT3A and DNMT3B are derived from duplication of a DNMT3 gene of chordates ancestor (e.g., Liu et al 2020, PMID 31969623). As such many invertebrates encode only one DNMT3. As previously shown (Yaari et al., 2019, PMID 30962443), plants have two distinct DNMT3-like protein family, the 'true DNMT3' and DRM, the plant specific de novo DNMT that is often considered to be a DNMT3 homolog (see Reviewer 2's comment). Our phylogenetic analysis successfully deviated the clade of DNMT3 and DRM from the rest of DNMTs (Figure S6). Yaari et al noted that PpDNMT3a and

PpDNMT3b, the two DNMT3 orthologs encoded by the basal plant Physcomitrella patens, are not orthologs of mammalian DNMT3A and DNMT3B, respectively. Therefore, to minimize such nomenclature confusions, any DNMTs that belong to either the DNMT3 or DRM clades indicated in Figure S6 are collectively referred to as 'DNMT3' throughout the paper (see Figure S2 for overview).

> *CoPAP analysis revealed that CDCA7 and HELLS are dynamically lost in the Hymenoptera clade and either co-occurs with DNMT3 or DNMT1/UHRF1 loss, which seems important. Unfortunately, the authors do not provide sufficient information in their figures or supplementary data about what is already known regarding DNA methylation levels in the different Hymenoptera species to further consider a potential impact of this observation. What is "the DNA methylation status" of all these organisms? This information cannot be easily retrieved from Table S2. A clearer presentation of what is actually known already would improve this paragraph.*

As the DNA methylation status of the species in the Hymenoptera clade has not been comprehensively tested, we initially did not include this information to Figure 7. However, during the course of the revision, we realized that Bewick et al.2017 (PMID 28025279) reported that DNA methylation is absent from the braconid wasp Aphidius ervi. We originally conducted synteny analysis on Aphidius gifuensis, which has a chromosome-level genome assembly with annotated proteins available in NCBI, whereas annotated proteins for Aphidius ervi protein are not available in NCBI. By conducting tBLASTn search against the Aphidius ervi genome, we now found that the presence/absence pattern of CDCA7, HELLS, DNMT1, DNMT3 and UHRF1 in Aphidius ervi is identical to that of Aphidius gifuensis, with a caveat that genome assembly of Aphidius ervi is at scaffold-level. In other words, DNA methylation, DNMT1 and CDCA7 are absent in Aphidius ervi, where 5mC is undetectable. Additionally, we also realized that the DNA methylation status reported for some species in Bewick et al. 2017 was inferred from the CpG frequency instead of the direct experimental detection of methylated cytosines. Therefore, we have amended Table S3 to indicate the presence of 5mC only for those species where this was experimentally tested. As such, we now consider the DNA methylation status of Fopius arisanus, which lacks DNMT1 and CDCA7, to be unknown.

Altogether, among the 17 Hymenoptera species that we analyzed (listed in the amended Table S3), the 8 species that have detectable DNA methylation all encode CDCA7, whereas the 2 species that do not have detectable DNA methylation lack CDCA7. We will note this finding in the revised text, and include the known 5mC status in the new Figure 7.

> *Furthermore, A. thaliana DDM1, and mouse and human Lsh/Hells are known to preferably promote DNA methylation at satellite repeats, transposable elements and repetitive regions of the genome. On the other hand, DNA methylation in insects and other invertebrates occurs in genic rather than intergenic regions and transposable elements (e.g. Bewick et al 2017; Werren JH PlosGenetics 2013). It would be helpful to elaborate on these differences.*

We were aware of this interesting point, which was discussed in the third paragraph of the Discussion. To better illustrate this point, we now expanded the Discussion (page 14) to speculate about the role of DNA methylation in insects, where emerging evidence indicates the importance of DNMT1 in meiosis. It should be noted that, in the Arabidopsis ddm1 mutant, reduction of CG methylation of gene bodies is common (50% of all methylated euchromatic genes) (Zemach et al, 2013). In addition, hypomethylation is not limited to satellite repeats and transposable elements in ICF patients defective in HELLS or CDCA7 (Velasco et al., 2018).

Reviewer #2 (Public Review):

In this manuscript, Funabiki and colleagues investigated the co-evolution of DNA methylation and nucleosome remolding in eukaryotes. This study is motivated by several observations: (1) despite being ancestrally derived, many eukaryotes lost DNA methylation and/or DNA methyltransferases; (2) over many genomic loci, the establishment and maintenance of DNA methylation relies on a conserved nucleosome remodeling complex composed of CDCA7 and HELLS; (3) it remains unknown if/how this functional link influenced the evolution of DNA methylation. The authors hypothesize that if CDCA7-HELLS function was required for DNA methylation in the last eukaryote common ancestor, this should be accompanied by signatures of co-evolution during eukaryote radiation.

To test this hypothesis, they first set out to investigate the presence/absence of putative functional orthologs of CDCA7, HELLS and DNMTs across major eukaryotic clades. They succeed in identifying homologs of these genes in all clades spanning 180 species. To annotate putative functional orthologs, they use similarity over key functional domains and residues such as ICF related mutations for CDCA7 and SNF2 domains for HELLS. Using established eukaryote phylogenies, the authors conclude that the CDCA7-HELLS-DNMT axis arose in the last common ancestor to all eukaryotes. Importantly, they found recurrent loss events of CDCA7-HELLS-DNMT in at least 40 eukaryotic species, most of them lacking DNA methylation.

Having identified these factors, they successfully identify signatures of co-evolution between DNMTs, CDCA7 and HELLS using CoPAP analysis - a probabilistic model inferring the likelihood of interactions between genes given a set of presence/absence patterns. As a control, such interactions are not detected with other remodelers or chromatin modifying pathways also found across eukaryotes. Expanding on this analysis, the authors found that CDCA7 was more likely to be lost in species without DNA methylation.

In conclusion, the authors suggest that the CDCA7-HELLS-DNMT axis is ancestral in eukaryotes and raise the hypothesis that CDCA7 becomes quickly dispensable upon the loss of DNA methylation and/or that CDCA7 might be the first step toward the switch from DNA methylation-based genome regulation to other modes.

The data and analyses reported are significant and solid. However, using more refined phylogenetic approaches could have strengthened the orthologous relationships presented. Overall, this work is a conceptual advance in our understanding of the evolutionary coupling between nucleosome remolding and DNA methylation. It also provides a useful resource to study the early origins of DNA methylation related molecular process. Finally, it brings forward the interesting hypothesis that since eukaryotes are faced with the challenge of performing DNA methylation in the context of nucleosome packed DNA, loosing factors such as CDCA7-HELLS likely led to recurrent innovations in chromatin-based genome regulation.

Strengths:

- *The hypothesis linking nucleosome remodeling and the evolution of DNA methylation.*
- *Deep mapping of DNA methylation related process in eukaryotes.*
- *Identification and evolutionary trajectories of novel homologs/orthologs of CDCA7.*

- *Identification of CDCA7-HELLS-DNMT co-evolution across eukaryotes.*

Weaknesses:

- *Orthology assignment based on protein similarity.*
- *No statistical support for the topologies of gene/proteins trees (figure S1, S3, S4, S6) which could have strengthened the hypothesis of shared ancestry.*

We appreciate the reviewers' accurate summary, nicely emphasizing the importance of the our study. We agree that better phylogenetic analysis for orthology assignment will strengthen our conclusion. Having anticipated this weakness, however, we specifically conducted a CoPAP analysis exclusively for Ecdysozoa specieswhich supported our major conclusion, as orthology assignment is straightforward in these species. For example, if we conduct BLAST search against the clonal raider ant Oocerea biroi protein dataset using human HELLS as a query, top 1 hit is a protein sequence annotated as one of three isoforms of 'lymphoid-specific helicase" (i.e., HELLS), with E value 0.0. Similarly, the top BLAST hit from the Oocerea biroi dataset using human DNMT1 as a query also returns with isoforms of DNMT1 with E value 0.0. As such, there are little disputes in orthology assignment in Ecdysozoa. Outside of Chordata, classification of DNMTs, particularly in Excavata and SAR, require more extensive identification in these supergroups. Our current orthology assignment for the major targets in this study (HELLS, DNMT1, DNMT3, DNMT5) is largely consistent with published results (Ponger et al., 2005 PMID 15689527; Huff et al, 2014 PMID 24630728; Yaari et al., 2019 PMID 30962443; Bewick et al., 2019 PMID 30778188). However, while we are preparing this response and re-crosschecking our assignments with these references, we realized that we had erroneously missed DNMT5 orthologs in Leucosporidium creatinivorum, Postia placenta, Armillaria gallica and Saitoella complicata, and a DNMT6 ortholog in Fragilariopsis cylindrus. We also recognized that DNMT4 orthologs were identified in Fragilariopsis cylindrus and Thalassiosira pseudonana in Huff et al 2014 (PMID 24630728), but in our phylogenetic analysis, these proteins form a distinct clade between DNMT1/Dim-2 and DNMT4 (original Figure S6), although the confidence level of this classification by Huff et al was not strong. To resolve this potential confusion in DNMT annotations, we generated new multiple sequence alignments with MUSCLE v5 and IQ-TREE 2 (maximum likelihood-based method, coupled with selection of optimal substitution model and bootstrapping). The tree topology was not significantly altered between the two methods, except for the unambiguous location of orphan DNMTs and DNMT4-related proteins. To avoid unnecessary confusion in the DNMT annotations, we decided to present MUSCLE-IQ- TREE for the DNMT phylogenetic tree and classification (new Fig. S6). The raw results of IQ-TREE analysis for CDCA7/zf-4CXXC_R1, HELLS SNF2 domain, and DNMTs are included as Dataset S1-S3. We then conducted CoPAP analysis using the corrected classification. As it is not clear a priori if fungal specific CDCA7-like proteins (now referred to as CDCA7F with class II zf-4CXXC_R1) should be considered CDCA7 orthologs, we conducted CoPAP against two lists; the first list includes CDCA7F in the CDCA7 group, whereas the second list includes a separate category of class II zn-4CXXC_R1, which includes CDCA7F. Both results show slightly different topology in the coevolutionary linkages but support our major conclusion that CDCA7 coevolved with DNMT1-UHRF1 and HELLS. These new CoPAP results are shown in Fig. S7.

Reviewer #1 (Recommendations For The Authors):

Summary

Last sentence: "...a unique specialized role of CDCA7 in HELLS-dependent DNA methylation maintenance...". What do the authors mean?

Our analysis strongly indicates that CDCA7 is dispensable in systems lacking HELLS and DNMT (particularly DNMT1). In other words, species preserve CDCA7 only if it has both HELLS and DNMT1 (or in some cases DNMT5). The importance of HELLS homologs in DNA methylation has been extensively studied in human, mouse and plants. However, in these studies, substantial DNA methylation remains despite the defective HELLS/DDM1 (especially in euchromatic regions). Additionally, there are species (e.g., Bombyx mori) that have DNMT1 and detectable DNA methylation but lacks HELLS and CDCA7. These observations suggest that the role of CDCA7 must be unique and specialized in a way that it is strongly coupled to HELLS-dependent DNA methylation (but not HELLS-independent DNA methylation), and that this function of CDCA7 seems to be inherited from the last eukaryotic common ancestor.

Introduction

- *page 3: "DNMTs are largely subdivided into maintenance and de novo DNMTs" - Which species are the authors referring to?*

As described in the cited reference (Lyko 2018), maintenance DNA methylation and de novo DNA methylation are well accepted functional classification of DNA methylation. It is also currently accepted that distinct DNMTs execute maintenance DNA methylation or de novo DNA methylation, although crosstalk between these processes has been reported. Therefore, we stated, “DNMTs are largely subdivided into maintenance DNMTs and de novo DNMTs”, and this subdivision is species independent.

- *page 3" "Maintenance DNMTs recognize hemimethylated CpGs. " - Can the authors please define the species and/or literature they are referring to? This seems important to clarify. For instance, mammalian DNMT1 requires a co-factor, UHRF1, which recognizes hemimethylated DNA and H3K9me3 (Bostick et al 2007).*

We meant to describe, “Maintenance DNMTs directly or indirectly recognize hemimethylated CpGs…”. The specific requirement of UHRF1 for DNMT1-mediated maintenance DNA methylation is explained in the subsequent sentence “In animals…”. In the case of Cryptococcus neoformans, DNMT5 recognizes hemimethylated DNA independently of UHRF1 in vitro to execute maintenance methylation.

- *page 3: The authors may want to mention that A. thaliana also has a de novo DNA methyltransferase, DRM2, a homolog of the mammalian DNMT3 methyltransferases. This seems important, since they show in Figure 1 that a de novo methyltransferase is found in A. thaliana. Also, later in their manuscript they mention plant de novo DNA methylation.*

Thanks for pointing this out. As shown in Figure 5, we classified plant DRMs as DNMT3-like proteins, but we now note this in the Introduction.

- *page 3: Sentence starting "In about 50% of ICF patients,..." - Why is DNMT3B referred to as "de novo", is it not a de novo DNA methyltransferase?*

You are correct. Quotation marks are now removed to avoid unnecessary confusion.

- *page 4: Sentence starting "Indeed, the importance of HELLS/CDCA7 in DNA methylation maintenance...", - Which references (Han et al., 2020; Ming et al., 2021; Unoki, 2021; Unoki et al., 2020) provide experimental evidence for a role of CDCA7 in DNA methylation maintenance by DNMT1?*

Thanks for pointing out the typo. “/CDCA7” is now removed.

- *page 5: Sentence starting "Indeed, it has been shown that DNMT3A..." - Should DNMTB be DNMT3B?*

Yes. This is now corrected.

Results

- *Page 5: Sentence starting "However, we identified a protein..." - No A. thaliana reference?*

We added Zemach et al 2010, and Chan et al 2005.

- *Figure 2B: "ICF4 mutations" should this be "ICF3 mutations"?*
- *Figure 3: "ICF4 mutations" should this be "ICF3 mutations"?*
- *Figure 4: "ICF4 mutations" should this be "ICF3 mutations"?*
- *Figure S1: Orange colored "CDC7L (fish), CDC7e, CDC7, CDC7L" is there an "A" missing?*
- *Figure S5: "ICF4 mutations" should this be "ICF3 mutations"?*

These typos are now corrected. Thank you.

- *Figure S7: What is "CDCA7(II)" referring to, "zf-4CXXC_R1 class II (plants)"?*

The original CDCA7 (II) included proteins with class II zf-4CXXC_R1, which are found in plants, fungi, Acanthamoeba castellanii and Amphimedon. Among those species, the prototypical CDCA7 orthologs are absent only in fungi. It has been a priori unclear if fungal proteins with class II zf-4CXXC_R1 (now we term CDCA7F) should be included in CDCA7 for CoPAP analysis. Although we originally included CDCA7F in CDCA7, we now show the results of two analyses. In the first one (Fig. S7A) CDCA7F was included in CDCA7, whereas in in the second one (Fig. S7B) CDCA7F was included in the separate category of class II zf-4CXXC_R1. Topologies of two results are slightly different, but they both show coevolutionary linkage between the CDCA7 and DNMT1- UHRF1 cluster.

- *Figure 4 and 5: In the case of preliminary genome assemblies what is the difference between empty squares with dotted lines and filled squares without dotted lines?*

As it is difficult to be certain of a gene’s absence (did the species lose the gene or is it simply not annotated due to incomplete genome coverage?), we illustrated the absence of a gene in

preliminary genome assemblies with an empty square with dotted outline. Since the presence of a gene is evident regardless of the level of genome assembly, the presence of a gene is represented with filled squares with solid lines, even for preliminary genome assemblies.

> - *Figure 1: Why was Mus musculus - one of the main model organisms used for many DNA methylation studies not included? Also what are empty and filled squares?*

Filled and empty squares indicate the presence and absence of the indicated genes, respectively. Clarifying statement is now added in the figure legends. Mus musculus is now included in the figure.

> - *Figure S2: Adding the existence of DNA methylation and DNMT3 in the bottom right part of the figure (overall no of species) would make this panel more informative*

We included this overview to summarize the co-retention of CDCA7, HELLS and maintenance DNMTs across the analyzed species. We decided not to include DNA methylation, since DNA methylation status is known for only a fraction of the listed species. Inclusion of DNMT3 will introduce too many possible gene presence-absence combinations to convey a clear message. However, we now mention in the revised text (page 11, second paragraph) that unlike the prevalent co-retention of DNMT1 in species with CDCA7, we identified several species that possess CDCA7, HELLS and DNMT1 but lack DNMT3. These examples include insects such as the bed bug Cimex lectularius and the red paper wasp Polistes canadensis.

> - *Page 6: Sentence starting "This leucine zipper sequence is highly conserved..." - Figure/Reference missing?*

The sequence alignment of the leucine zipper is now shown in Fig. 2C.

> - *page 6: Sentence starting "In contrast to zf-4CXXC_R1 motif-containing proteins..." - The authors may want to mention the role of the CXXC zf domain in KDM2A/B, DNMT1, MLL1/2 and TET1/3 and what the CDCA7 CXXC zf domain is/could be required for.*

The notion that zf-CXXC binds to nonmethylated CpG is now included. Due to the substantial difference between zf-CXXC and zf-4CXXC_R1, we hesitated to relate the function of zf-4CXXC_R1 with zf-CXXC, but we now discuss a potential role of zf- 4CXXC_R1 in sensing DNA methylation status in Discussion (Page 13).

> - *page 7: Sentence starting "Second, the fifth cysteine is replaced..."- Zoopagomycota" - Figure 4A does not have this labeling, one has to deduce this from Figure 4B.*

We fixed this by including the list of Zoopagomycota species in the main text.

> - *page 7: Sentence containing "Neurospora crassa DMM-1 does not directly regulate DNA methylation or demethylation but rather..." - How does the information about DMM- 1 relate to what is shown in Figure 4B, to CDCA7, HELLS and DNMTs? Please clarify.*

Both Neurospora DMM-1 and Arabidopsis IBM1 contain the JmjC domain and are implicated in an indirect control mechanism of DNA methylation. Since it has never been pointed out that they have a divergent zf-4CXXC_R1 domain, which clearly shares the origin with CDCA7 proteins, we thought that this is important to note. We realized that we did not clearly mark Neurospora XP-956257 as DMM-1 in Fig. 4B. This is now fixed.

> - *Heading "Systematic identification of CDCA7, HELLS and DNMT homologs in eukaryotes". When mentioning CDCA7 the authors may want to decide on the use of one consistent definition of "prototypical (Class I) CDCA7-like proteins (i.e. CDCA7 orthologs)" "Class I CDCA7 proteins". Constantly changing the way how they refer to these proteins is very confusing.*

We now make it clear that we call proteins with class I zf-CXXC_R1 motif CDCA7 orthologs. We also define class II zf-4CXXC_R1 (as those with a substitution at ICF- associated glycine residue). Since no clear CDCA7 orthologs can be found in fungi, we now call fungi proteins with class II zf-4CXXC_R1 “CDCA7F”, implying its ambiguous orthology assignment.

> *Under this heading there is also no mention of DNMTs. Instead, the authors introduce DNMTs under the heading "Classification of DNMTs in eukaryotes" - Please clarify.*

This is now corrected.

> - *page 9: Sentence containing "... presence of DNMT1, UHRF1 and CDCA7 outside of Viridiplantae and Opisthokonta is rare". What does "rare" mean? How is UHRF1 relevant here?*

Among the 32 species outside of Viridiplantae and Opisthokonta, only the Acanthamoeba castellanii genome encodes clear orthologs of DNMT1, UHRF1 and CDCA7. Although it is often difficult to deduce if the selected panel of species is a reasonable representation, we think that it is not unreasonable to state that Acanthamoeba is a rare case to encode this set of proteins outside of Viridiplantae and Opisthokonta. We include UHRF1 since it is a well-established activator of DNMT1, and indeed our CoPAP analysis showed a tight coevolution of UHRF1 with DNMT1. Outside of Viridiplantae and Opisthokonta, only Acanthamoeba castellanii and Naegleria gruberi encode UHRF1. Interestingly, these two species also encode CDCA7 and HELLS.

Having said that, we rephrased this sentence, which reads; “Species that encode a set of DNMT1, UHRF1, CDCA7 and HELLS are particularly enriched in Viridiplantae and Metazoa.”

> - *page 11: Sentence containing "..., that the function of CDCA7-like proteins is strongly linked to HELLS and DNMT1,..." What do the authors mean with "the function of CDCA7-like proteins"? And what happened to DNMT3?*

Our observation that almost all species that contain CDCA7 (including fungal CDCA7F) also have DNMT1 and HELLS, despite the frequent loss of these genes in species that do not contain CDCA7, indicates “that the function of CDCA7-like proteins is strongly linked to HELLS and DNMT1”. We found only 2 species that possesses CDCA7 (class I or class II) but not DNMT1 among the panel of 180 species. These 2 exceptional species, Naegleria gruberi and Taphrina deformans, do encode UHRF1-like proteins and a DNMT (an orphan DNMT in N. gruberi and DNMT4 in T. deformans). In contrast, we found 26 species that possess CDCA7 (or CDCA7F) but not DNMT3 (Table S1), so the linkage between CDCA7 and DNMT3 is weaker.

- *page 11: Sentence containing "..., CDCA7 is lost from this gene cluster in parasitoid wasps, including Ichneumonoidea wasps and chalcid wasps". This sentence is confusing because already in an earlier paragraph the authors say that "Microplitis demolitor lost CDCA7" and in the following sentence they say "...among Ichneumonoidea wasps, CDCA7 appears to be lost in the Braconidae clade, ...". It would greatly help this reader if the authors could streamline these sentences and also decide on whether CDCA7 is lost in M. demolitor or CDCA7 appears to be lost in M.demolitor.*

The confusion was in part due to the difficulty in differentiating between the true loss of a gene versus its apparent absence in a species due to an incomplete genome assembly, including for of M. demolitor. To verify that the loss of CDCA7 was not due to gaps in the genome assembly, we executed the synteny analysis. However, we edited this section to improve the readability (Page 12-13).

> *What could be the role for HELLS/CDCA7 in insect DNA methylation? In several cases, the authors analyses reveal co-evolutionary links between DNMT3 (DNMT3A?) and CDCA7/HELLS. I do not understand why this finding is not really discussed by the authors. Instead there is a strong focus on replication-uncoupled DNA methylation maintenance. Could the authors elaborate why?*

The role of DNA methylation in insects is largely unclear, so discussion must be highly speculative. A recent finding in the clonal raider ant, showing that DNMT1 is not essential for development but is critical for oogenesis, pointed toward a possible more universal role of DNA methylation in meiosis. Stimulated from a finding in Neurospora, where DNA methylation is required for homolog pairing during meiosis, we discuss a speculative model that DNA methylation status acts as a hallmark to distinguish between healthy/young DNA and old/mutated (or competitive/pathogenic) DNA at homolog pairing during meiosis (page 14).

Regarding the cases where CDCA7 and DNMT3 are co-lost, we had discussed about this phenomenon at the last section of Result, stating, “This co-loss of CDCA7 and DNA methylation (together with either DNMT1-UHRF1or DNMT3) in braconid wasps suggests that evolutionary preservation of CDCA7 is more sensitive to DNA methylation status per se than to the presence or absence of a particular DNMT subtype.” Please note that we found several lineages that lacks CDCA7 but has DNMT1 (and DNMT3), whereas almost all species that has CDCA7 also has DNMT1 (but not necessarily DNMT3). Supported with our CoPAP analyses, our results indicate the tight functional link between CDCA7 and DNMT1, but it does not necessarily mean that CDCA7 does not play any role related to DNMT3 or de novo methylation. Clarification of this point and our speculation of how CDCA7 loss is linked to reduced requirement of DNA methylation are discussed in page 13 and 14 with additional texts.

> *Discussion*
>
> - *page 12: Where is the data supporting. "... the red flour beetle Tribolium castaneum possesses DNMT1 and HELLS, but lost DNMT3 and CDCA7"?*

Figure 5, Figure S2 and Table S1. This is now noted in the text.

> - *page 14: Based on which parts of their analyses or evidence from the literature can the authors speculate that "...the evolutionary arrival of HELLS-CDCA7 in eukaryotes might have been required to transmit the original immunity-related*

> *role of DNA methylation from prokaryotes to nucleosome-containing (eukaryotic) genomes"? Please clarify.*

This is inferred from the well-known role of DNA methylation in bacteria for defending against phage viruses. However, it was not correct to state that such a function was inherited from prokaryotes. It should be stated that it was inherited from the last universal common ancestor (LUCA). We also admit that it is not clear if such an immunity-related role was inherited from LUCA, or if it emerged through convergent evolution. Therefore, we amended this description to emphasize our hypothesis that the advent of CDCA7 was “a key step to transmit the DNA methylation system from the LUCA to the eukaryotic ancestor with nucleosome-containing genomes”.

> *Supplementary Figures/Tables*
>
> - *page 26: Table S2 and Table S3, it seems that these tables show data that supports what is shown in Figure 7 and not Figure 5.*

You are correct. Thank you for pointing out the typos.

> *Has the methylation status been assessed in C. glomerata, C. typhae, Chelonus insularis, Diachasma alloeum or Aphidius gifuensis? Please clarify in Table S2.*

Not to our knowledge. However, as we realized that absence of DNA methylation in Aphidius ervi was previously reported (Bewick et al 2017), we now included this data together with presence/absence analysis of DNMT1, UHRF1, DNMT3, CDCA7 and HELLS. Known presence/absence of DNA methylation is now shown in Fig.7.

> ## *Reviewer #2 (Recommendations For The Authors):*
>
> *Recommendation to strengthen the paper:*
>
> *1. Phylogenetics:*
>
> - *Test and report the appropriateness of the substitution model used in protein alignments/trees.*
> - *Use Maximum likelihood methods and/or MCM Bayesian inference to build and report trees with well supported topologies. This is required to properly assign orthology (shared ancestry). This will avoid false interpretation due to technical limitation of similarity-based phylogenies (without statistical support). Figure S1, S3, S4 and S6.*

To address these points, we made new multisequence alignments using MUSCLE v6 and generated phylogenetic trees using the maximum likelihood-based IQ-TREE 2, where multiple models were screened. A consensus tree was generated after 1000 bootstrap replicates from the best alignment and model. The topology and assignment of these new trees were largely consistent with the original trees, except for some corrections in DNMT assignment as discussed below.

1. We realized that we erroneously missed DNMT5 orthologs of Leucosporidium creatinivorum, Postia placenta, Armillaria gallica and Saitoella complicata., and DNMT6 orthologs from Fragilariopsis cylindrus reported in Huff et al 2014 (PMID 24630728). They are now included in the new list and CoPAP analysis.

2. DNMT4 orthologs were identified in Fragilariopsis cylindrus and Thalassiosira pseudonana by Huff et al 2014 (PMID 24630728), but in our original phylogenetic analysis, these proteins form a distinct clade between DNMT1/Dim-2 and DNMT4. The new tree and classification are more consistent with Huff et al, so we present the new tree in Fig. S6 and conducted the classification based on this tree.

Beside Fig. S6, we decided to maintain original Fig. S1, S3 and S4 (with a few adjustments) for better visibility, but we included the results of IQ-TREE analysis as Dataset S1-S3.

The CoPAP analysis based on the revised assignment slightly changed the topology of coevolutionary linkages. In addition, we obtained a slightly different result depending on whether fungal specific CDCA7 with class II zn-4CXXC_R1 (now referred to as CDCA7F) is included as a CDCA7 ortholog or not. Despite this difference, we reproducibly observed the coevolutionary linkage between CDCA7 and DNMT1- UHRF1.

> - *Be more careful with wording: RBH is not sufficient to call gene/proteins orthologs (e.g. Page 8). The above mentioned method will help you support this claim (+ synteny when you can).*

We were aware of this issue. This is why we conducted phylogenetic tree building based on sequence alignment of full-length HELLS (Fig. S3) and SNF2 domain only (Fig. S4), as explained in the text. We found that the RBH criterion is robust in Metazoa; orthologs are easily recognizable with very low E-value (0.0) and extensive homology over the full length of the protein, while synteny is not practical to employ in the diverse set of species.

> - *Also, use "co-retention" or "co-evolution" but not "co-selection" when describing CoPAP results - as CoPAP does not test for signature of natural selection.*

This is a good point and is now corrected.

> - *The statistics (p-val...) underlying the CoPAP analyses should be explained.*

The explanation is now added in Methods section.

“A method to calculate p-value for CoPAP was described previously (Cohen et al., 2012, PMID 22962457). Briefly, for each pair of tested genes, Pearson's correlation coefficient was computed. Parametric bootstrapping was used to compute a p-value by comparing it with a simulated correlation coefficient calculated based on a null distribution of independently evolving pairs with a comparable exchangeability (a value reporting the likelihood of gene gain and loss events across the tree).”

> *1. Figure S2 and S3 could be improved for readability*

After consideration of this criticism, we decided to keep their original formats for following reasons.

Figure S2. The purpose of this list is to better visualize the comprehensive list shown in Table S2. A consolidated list is already shown in Figure 5. An alternative choice is to make a diagram where individual species names are unreadable. This kind of presentation is seen in many published papers, but we found that they are not helpful to check the details. As this is a supplementary figure, we prefer to show the detailed data that can be visible without a specialized software.

Figure S3. This figure is included to show which SNF2 family proteins are more likely to be misassigned as HELLS/DDM1 orthologs. We believe that the figure serves this purpose.

> *1. What is the meaning of the coloring patterns of ICF residues in znf?*

ICF residues are highlighted as light blue in the schematics to indicate its conservation. In the alignment, the coloring reflects the level of conservation within the shown set of proteins, and the choice of coloring was set by Jalview.

> *1. To improve clarity: the introduction could be more focused on evolutionary considerations and functional link between CDCA7-HELLS and DNMTs.*

We revised the first paragraph of the introduction to illustrate this point.

> *1. Could indicate the CDC7A loss / DNA methylation hypothesis in the abstract.*

We now included this hypothesis in the Abstract.