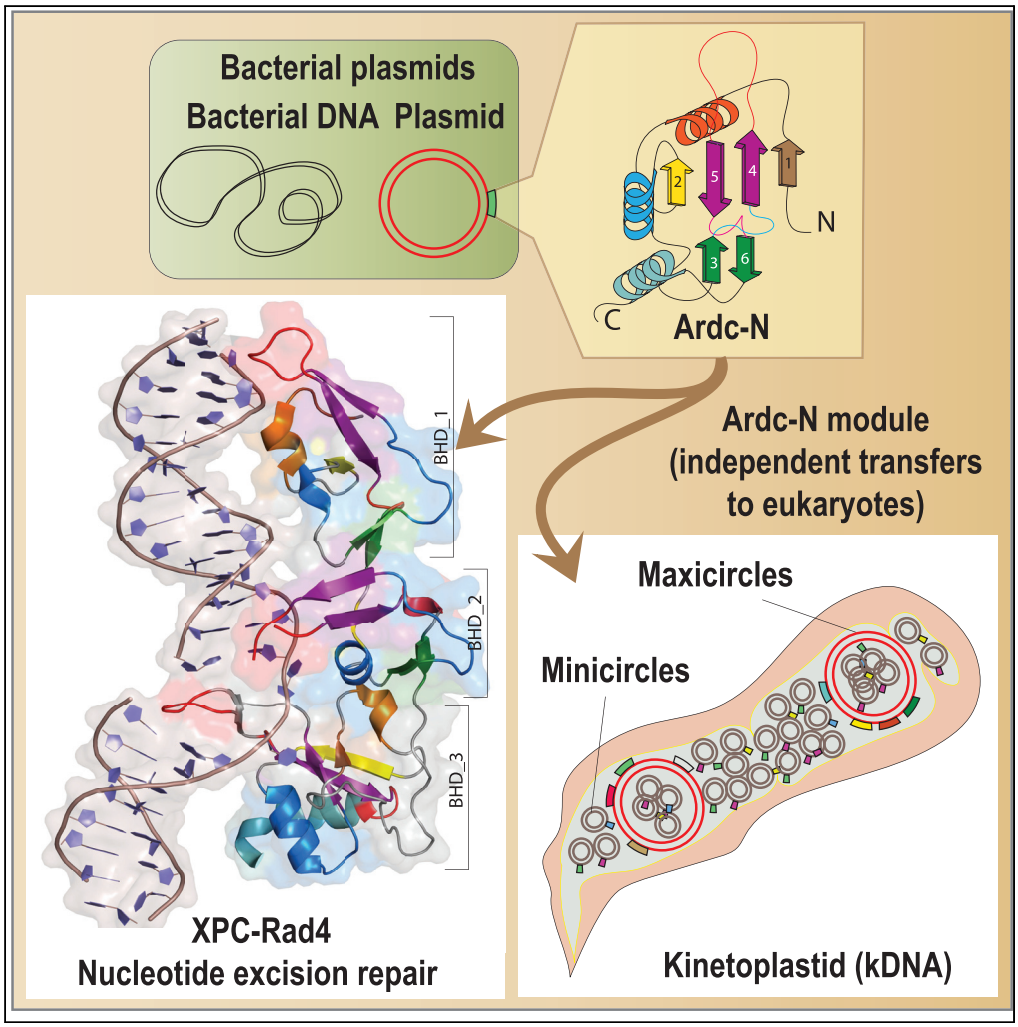
iScience

CellPress

Article

Unexpected Evolution of Lesion-Recognition Modules in Eukaryotic NER and Kinetoplast DNA Dynamics Proteins from Bacterial Mobile Elements



Arunkumar Krishnan, A. Maxwell Burroughs, Lakshminarayan M. Iyer, L. Aravind

aravind@ncbi.nlm.nih.gov

HIGHLIGHTS

Two eukaryotic acquisitions of the ArdC-N domain from bacterial mobile elements

The first spawned the β-hairpin domains of the nucleotide excision repair proteins

The second gave rise to Tc-38-like proteins involved in kinetoplastid kDNA dynamics

Multiple selfish-element-derived components relate to plasmid-like features of kDNA

Krishnan et al., iScience 9, 192–208
November 30, 2018
https://doi.org/10.1016/j.isci.2018.10.017

iScience

CellPress

Article

Unexpected Evolution of Lesion-Recognition Modules in Eukaryotic NER and Kinetoplast DNA Dynamics Proteins from Bacterial Mobile Elements

Arunkumar Krishnan,[1,2] A. Maxwell Burroughs,[1,2] Lakshminarayan M. Iyer,[1] and L. Aravind[1,3,*]

SUMMARY

The provenance of several components of major uniquely eukaryotic molecular machines are increasingly being traced back to prokaryotic biological conflict systems. Here, we demonstrate that the N-terminal single-stranded DNA-binding domain from the anti-restriction protein ArdC, deployed by bacterial mobile elements against their host, was independently acquired twice by eukaryotes, giving rise to the DNA-binding domains of XPC/Rad4 and the Tc-38-like proteins in the stem kinetoplastid. In both instances, the ArdC-N domain tandemly duplicated forming an extensive DNA-binding interface. In XPC/Rad4, the ArdC-N domains (BHDs) also fused to the inactive transglutaminase domain of a peptide-N-glycanase ultimately derived from an archaeal conflict system. Alongside, we delineate several parallel acquisitions from conjugative elements/bacteriophages that gave rise to key components of the kinetoplast DNA (kDNA) replication apparatus. These findings resolve two outstanding questions in eukaryote biology: (1) the origin of the unique DNA lesion-recognition component of NER and (2) origin of the unusual, plasmid-like features of kDNA.

INTRODUCTION

Diverse selfish elements including bacteriophages, plasmids, and conjugative transposons possess the capacity for proliferation within the cell or the genome of their hosts. Thus, they are unceasingly entwined in multilevel conflicts with the host and other co-resident genetic elements, which possesses mechanisms to combat the negative effects of these entities on its own fitness (Aravind et al., 2012; Smith and Price, 1973; Werren, 2011). Such inter- and intra-genomic biological conflicts have spawned numerous molecular adaptations that function as "biochemical armaments" in both cellular genomes and the selfish elements: prime examples include restriction-modification (Ishikawa et al., 2010; Kobayashi, 2001), toxin-antitoxin (Yamaguchi et al., 2011), CRISPR/Cas (Makarova et al., 2011), and polyvalent protein systems (Iyer et al., 2017), among others. Intriguingly, examination of some of these above-listed prokaryotic conflict systems has also led to the realization that they are potential evolutionary "nurseries" for molecular innovation spurred by the pressures for rapid adaptations. These adaptations are then disseminated via lateral transfer and used in functional contexts, which are very distinct from their original role in biological conflicts. Thus, we see numerous molecular adaptations, such as methylases, demethylases, and oxidases originally involved in the synthesis of peptide secondary metabolites in bacteria and DNA-binding domains such as the HIRAN domain involved in the replication apparatus of caudate bacteriophages were later recruited for distinctive functions in eukaryotic chromatin protein complexes (Aravind et al., 2011; Kaur et al., 2018; Zhang et al., 2014). Likewise, components of the RNAi systems (RNaseH fold containing PIWI domains; distinct family of RNA-modifying primpol domains in kinetoplastids) and specialized components of DNA repair and DNA recombination (several helicases and nucleases belonging to the nucleotide excision repair (NER) pathway and recombinational pathway) have their ultimate evolutionary roots in the prokaryotic conflict systems (Aravind et al., 1999, 2012; Burroughs et al., 2014; Burroughs and Aravind, 2016; Zhang et al., 2012).

The anti-restriction factor ArdC is one such module, transmitted during the invasion of bacterial hosts from the plasmids pSA and RP4. The ArdC protein has been previously demonstrated to bind single-stranded DNA (ssDNA) (Belogurov et al., 2000), and they are the among the founding members of a unique class of proteins termed "polyvalent proteins." These polyvalent proteins are characterized by a combination of domains, often enzymatic, with disparate biochemical activities in the same polypeptide that are deployed by bacteriophages, plasmids, and certain conjugative transposons (Iyer et al., 2017). They mediate biological conflicts of these elements with their hosts by deploying a diverse class of biochemical activities

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[2]These authors contributed equally

[3]Lead Contact

*Correspondence: aravind@ncbi.nlm.nih.gov

https://doi.org/10.1016/j.isci.2018.10.017

alongside or immediately after invasion to help establish the element and counter host defenses. One such domain, which might be combined with ArdC in polyvalent proteins, is the TraC-like primase domain, which is part of the machinery facilitating conjugation-coupled replication of plasmids such as RP4 (Belogurov et al., 2000; Miele et al., 1991; Rees and Wilkins, 1990). Our recent study of counter-host strategies deployed by plasmids and phages showed that the classic ArdC protein of pSA contains two globular domains: a distinct N-terminal domain (ArdC-N) with the DNA-binding function and a C-terminal zincin-like metallopeptidase (MPTase) domain (Figure 1A). The ArdC-N is one of the most prevalent domains in the aforementioned polyvalent proteins and is coupled with multiple domains possessing an array of disparate effector activities or other DNA-binding domains (Iyer et al., 2017) (Figure 1A). Together, these observations suggested that ArdC might perform its anti-restriction function by coating ssDNA via the ArdC-N domain during invasion and also possibly by deploying the C-terminal MPTase domain to target the restriction endonucleases (REases) for cleavage or autoproteolytically releasing other effector domains coupled to the ArdC-N domain in polyvalent proteins (Iyer et al., 2011).

In the course of that study (Iyer et al., 2017), we also detected significant sequence similarities between the ssDNA-binding ArdC-N domain and the *Trypanosoma* Tc-38 (p38) protein. Tc-38 is a DNA-binding protein that associates with the structurally complex DNA network of the kinetoplastid mitochondrion known as kinetoplast DNA (kDNA) (Liu et al., 2006). The kDNA consists of two recognizable classes of DNA "circles": the maxi- and the minicircles. The maxicircles are larger DNA rings (20–40 kbp) found in dozens of identical copies encoding rRNAs and cryptic genes. The minicircles, in contrast, are a class of small DNA rings (0.5–2.5 kbp) displaying remarkable sequence heterogeneity found in several thousand copies per kDNA network and encode guide RNAs that act as templates for directing RNA editing of maxicircle-derived cryptic transcripts (Jensen and Englund, 2012; Liu et al., 2005; Lukes et al., 2002; Shapiro, 1993). The Tc-38 protein plays a crucial role in replication and maintenance of kDNA, functioning as an ssDNA-binding protein at the replication origin, and largely influencing the count and supercoiling of minicircles (Duhagon et al., 2003, 2009; Liu et al., 2006). In addition, our searches pointed to a potential evolutionary relationship between ArdC-N and the DNA-binding domains of the NER XPC/Rad4 protein (Min and Pavletich, 2007), which operates on DNA segments containing CPD lesions.

Building on these observations, we detail herein the unification of the ArdC-N domain with the Tc-38-like and C-terminal DNA-binding domains of the XPC/Rad4 proteins. We trace the evolutionary trajectory of this newly recognized DNA-binding fold and find that the ArdC-N module was horizontally transferred to eukaryotes from bacterial conjugative elements, likely twice independently. On one occasion, transfer of the ArdC-N played a role in the emergence of the lesion-recognition domains of XPC/Rad4 protein. On another occasion, it was recruited for a role in kDNA binding. This version of the ArdC-N domain underwent extensive expansion in the kinetoplastid lineage, possibly complementing the diversification/expansion of kDNA circles. We also show that the catalytically inactive transglutaminase (TGL) domain of Rad4 emerged from ancestral archaeal peptide-N-glycanases (PNGases) and then fused with an ArdC-N from a bacterial plasmid giving rise to the extant form of XPC/Rad4. These findings throw light on the provenance of certain eukaryotic systems that have thus far remained largely inscrutable. They also improve our understanding of the mechanism of eukaryotic NER and kDNA replication in kinetoplastids.

RESULTS

Identification and Structural Analysis of Eukaryotic Homologs of the ArdC-N Domain

The Tc-38 Family of Kinetoplastid-Specific DNA-Binding Proteins Contains Multiple Copies of the ArdC-N Domain

Using prokaryotic ArdC-N domain as search seeds, we initiated recursive sequence profile searches using the PSI-BLAST program against the non-redundant protein database of the National Center for Biotechnology Information. Although the initial iterations recovered the prototypical prokaryotic versions of these domains in the polyvalent proteins, we surprisingly also recovered proteins with this domain from eukaryotes albeit with domain architectures completely unlike those from prokaryotes. For example, a search initiated with an ArdC-N domain from *Salmonella enterica* (GenBank: WP_023226849.1: residues 1–140) recovered a significant relationship with Tc-38-like ssDNA-binding protein from the deep-branching kinetoplastid and those from crown-group kinetoplastids such as *Trypanosoma vivax* (see Transparent Methods). A reciprocal search using a Tc-38 homolog from *Perkinsela* (GenBank: KNH07778.1) easily recovered the Tc-38 family proteins from other kinetoplastids and several bacterial ArdC-N homologs with

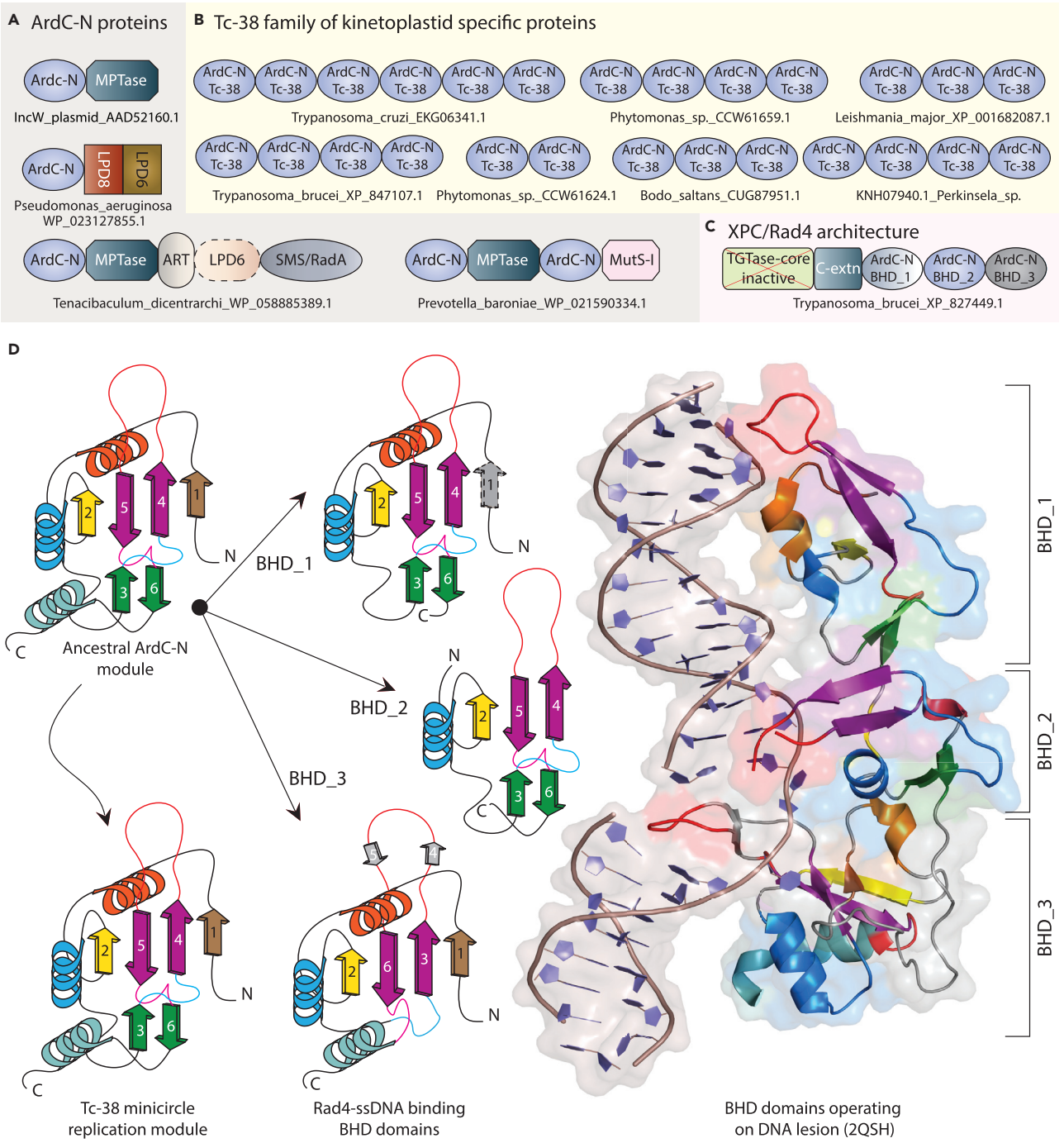


Figure 1. Domain Architecture and Structural Fold of the ArdC-N Homologs

(A–C) Domain architectures of a few exemplars of ArdC-N (A), Tc-38 (B), and XPC/Rad4 (C) proteins. Proteins are labeled by their full species name and accession number.

(D) Schematic topological rendering on the left shows the ssDNA-binding structural scaffold shared between the ancestral ArdC-N module, Tc-38 minicircle replication module, and the BHDs of the XPC/Rad4. The major β-sheet of this domain is formed by four β-strands (strands 1,2,4 and 5), whereas the β-strands 3 and 6 form the minor sheet: the characteristic long hairpin loop connects the central two strands of the major sheet. Ribbon diagram on the right shows DNA-binding interface of the BHD domains (PDB ID: 2QSH): secondary structure elements are colored the same as shown on the topological diagram (left). The long hairpin loop is inserted into the DNA double helix in the recognition of the DNA damage site in nucleotide excision repair.

e-values reaching 1×10^{-5} in PSI-BLAST iteration 2. This affirmed the presence of an ArdC-N domain in Tc-38. Subsequently, multiple searches using diverse Tc-38 sequences as search seeds successively recovered further related sequences from the kinetoplastids, pointing to a large expansion of Tc-38-like ssDNA-binding proteins from diverse kinetoplastids, including early branching representatives such as *Perkinsela* sp and *Bodo saltans* (see exemplars in Figure 1B). Tailored searches for further versions in other euglenozoans such as the diplonemids and euglenids failed to recover reliable homologs of Tc-38, thus suggesting that Tc-38 is specific to kinetoplastids (Data S5).

The So-Called BHD Domains of Nucleotide Excision Repair (NER) Protein Rad4/XPC Are ArdC-N Domains

To further investigate the ArdC-N domain, a hidden Markov Model (HMM) profile constructed from a multiple alignment of ArdC-N sequences was searched against a database of HMM profiles constructed from the Pfam database (Finn et al., 2016) and individual Protein Databank (PDB) (Rose et al., 2017) entries (see Methods) with the HHpred program (Alva et al., 2016). These searches surprisingly detected a significant relationship between the ArdC-N domain and the Pfam profile "BHD_2," one of three domains labeled BHD hitherto exclusively observed in the C-terminal DNA-binding region of the NER proteins XPC/Rad4 (p value: 2.2×10^{-8}, probability: 96.7%; PDB ID: 2QSH [Min and Pavletich, 2007]; p value: 8×10^{-7}, probability: 93.8%). Reverse profile-profile searches of the sequences corresponding to the "BHD_2" model from Pfam recovered not just the eukaryotic XPC/Rad4 proteins but also bacterial exemplars of the ArdC-N domain (see Transparent Methods). The BHD_2 is the central domain in the tripartite organization of the C-terminal DNA-binding region of XPC/Rad4, flanked N terminally by the BHD_1 and C terminally by the BHD_3 domains (Min and Pavletich, 2007). N terminal to the BHD_1–3 module, the XPC/Rad4 proteins are further fused to an inactive TGL fold domain (see following sections) (Anantharaman et al., 2001) (Figure 1C). Visual inspection of the individual BHD domain structures and analysis of concordance in secondary structure elements strongly suggested a relationship across the three domains (Figure 1D). Likewise, pairwise structural homology searches (see Methods) initiated with the individual BHD domains as queries confirmed relationship between the three BHDs (see Transparent Methods).

Shared Structural Core and Conserved Features of the ArdC-N domain

Based on the multiple sequence alignments constructed using representatives of the ArdC-N domain from diverse taxa, we investigated the structural scaffold of the ArdC-N domain, specifically informed by the crystal structures of the versions found in the Rad4 protein (Figures 1D and 2; also see Data S1, S2, S3, and S4). These structure-informed alignments together with structure predictions indicated that the ancestral core of the ArdC-N domain is a rather distinctive structure with no close relationship to any other protein fold. These observations also dispel a previously held view that the BHDs in XPC/Rad4 were OB-fold domains (Clement et al., 2010; Maillard et al., 2007). The ArdC-N domain is characterized by a couple of β-sheets: the major one formed by up to four strands and the minor one by two strands. The polypeptide chain crosses over at the point of entry and exit to the central two strands of the major sheet, and the crossover is bounded by the two strands of the minor sheet (Figure 1D). Furthermore, the crossover region has a distinctive meander at the N terminus and a single 3_{10} helical turn bounded by less-structured segments at the C-terminus. This specific structural feature is termed the "squiggle" (Figure 1D) (Burroughs et al., 2006; Dai et al., 2006, 2009). Together, these sheets form an open barrel-like structure. Despite the striking structural conservation, distinct ArdC-N clades display high sequence heterogeneity (Figure 2; Data S1, S2, and S3). Nonetheless, one subtle yet notable conserved sequence signature across the ArdC-N domains marks the squiggle: a hhsxxQ motif (with "h" representing a hydrophobic residue, "s" representing a small residue, and "x" representing any residue; the first hydrophobic residue is usually aliphatic, whereas the second is aromatic) (Figures 2A and 2B; Data S1 and S2). The glutamine that marks the end of the motif and occurs just before the second strand of the second sheet is mostly conserved, notable exceptions being BHD_1 and BHD_3 (Figures 2C and 2D; Data S1, S2, and S3). This glutamine appears to play a role in maintaining the distinctive structure by stabilizing the 3_{10} helical turn via a contact with the backbone. Furthermore, comparable squiggles observed in other, distinct protein folds have been linked to regions displaying conformational flexibility in a fold (Burroughs et al., 2006; Dai et al., 2006, 2009). The ArdC-N squiggle and the accompanying conserved glutamine residue could similarly facilitate conformational change during recognition of specific features in DNA.

Diverse ArdC-N domains often conserve certain aromatic residues, which are implicated in mediating DNA contacts (Maillard et al., 2007). A prime example is a conserved hydrophobic position (most frequently a

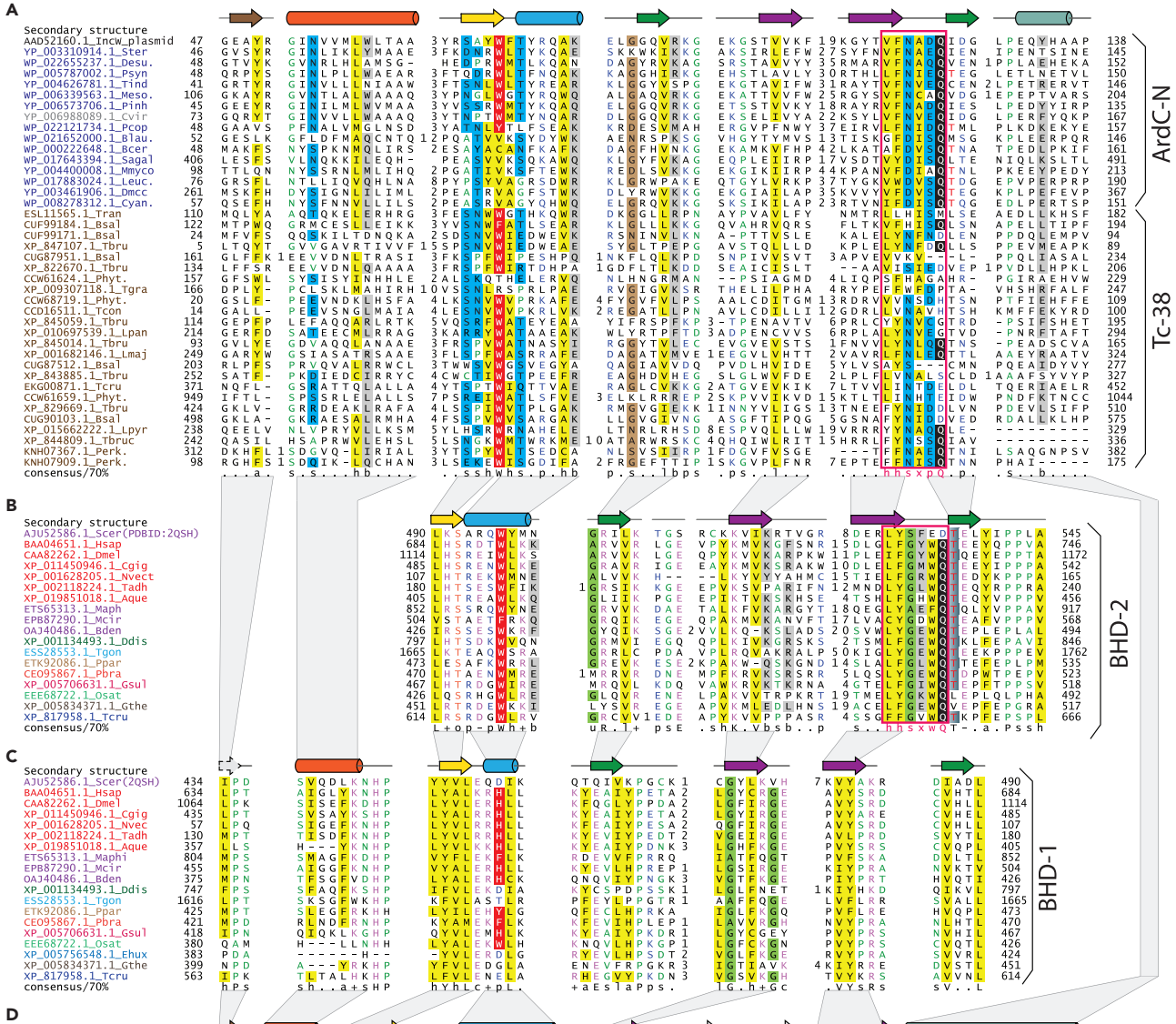


Figure 2. Sequence Features of the ArdC-N Domain

(A–D) Multiple sequence alignments of ArdC-N/Tc-38 (A), BHD_2 (B), BHD_1 (C), and BHD_3 (D) domains. Secondary structure elements are shown on the top and colored the same as shown in topological diagrams in Figure 1D. The characteristic hhsxxQ motif is highlighted as a red rectangular box. Polar and small residues shared between the ArdC-N and Tc-38 to the exclusion of other domains are highlighted in blue and light brown, respectively. Conserved aromatic residues (typically a tryptophan, phenylalanine, and histidine) are highlighted in red.

tryptophan) observed at the beginning of helix-2 (helix-1 for BHD_2) across most versions of the domain (Figure 2). The BHD versions of the ArdC-N domain in XPC/Rad4 further indicate that different representatives of the domain might potentially adopt distinct modes of binding DNA (Min and Pavletich, 2007). Of

the three copies, the BHD_2 version remained closest to the plasmid ArdC-N at the sequence level but lost certain structural elements (Figures 1D and 2B). Conversely, the BHD_1 and BHD_3 versions have diverged considerably at the sequence level while retaining most core structural features (Figures 1D, 2C, and 2D). However, a common denominator is the contact made with ssDNA or double-stranded DNA by the long hairpin loop connecting the central two strands of the major sheet (Figure 1D). Structural studies on the Rad4 protein implicate the deep insertion of this loop into the DNA double helix in the recognition of the DNA damage site in NER (Min and Pavletich, 2007) (Figure 1D). Furthermore, version-specific contacts with the DNA backbone are mediated by the helices downstream of the first and second strands of the core (Figure 1D).

Unraveling the Complex Evolutionary History and Functional Implications of Proteins Containing the ArdC-N Domain in Eukaryotes

Eukaryotes Acquired the ArdC-N Domain through Two Independent Transfers from Bacteria

Notably, searches initiated with Tc-38-like or BHD family members never directly recover each other as immediate hits but recover different bacterial ArdC-N domains as their best hits. This observation is further consistent with (1) distinct phyletic distribution of BHD domains and Tc-38: Tc-38 proteins are restricted to the kinetoplastids (Data S5), whereas the BHD-domain-containing XPC/Rad4 family is widespread across eukaryotes (Data S6) and (2) subtle yet clearly distinct conservation patterns shared by the bacterial ArdC-N with the BHDs and Tc-38 families to the exclusion of the other: the bacterial ArdC-N and BHD_2 specifically share a threonine residue immediately following the conserved Q of the hhsxxQ motif, whereas in Tc-38 the equivalent residue is mostly hydrophobic (Figures 2A and 2B; Data S1, S2, and S3). Conversely, the ArdC-N domains in Tc-38 domains share several features establishing a close relationship with the bacterial ArdC-N domains to the exclusion of the BHDs: (1) clear conservation of both N-terminal α-helices containing shared polar residues (typically an asparagine, a serine, and a threonine), located at the beginning of helices 1 and 2 and strand 2 (Figure 2A; Data S1, S2, and S3); (2) similarly, shared polar residues, typically a serine, an aspartate, or a glutamate, as well as an asparagine, were found upstream of the characteristic Q residue (Figure 2A); and (3) strong conservation of a small residue, typically a glycine, between the second conserved helix and the β-strand. These observations suggest that the prokaryotic ArdC-N domains were acquired independently on two distinct occasions by the eukaryotes: one of these acquisitions led to the more broadly distributed BHD versions found in the XP-C/Rad4 proteins, whereas the second acquisition led to the kinetoplastid-specific Tc-38-like versions. To better understand this dual acquisition of the eukaryotic ArdC-N domains, we next systematically investigated the evolutionary histories of the XP-C/Rad4 and TC-38-like proteins and explored the potential functional implications of their constituent domains.

The Origin of XPC/Rad4 through the Confluence of Domains with Distinct Evolutionary Histories Acquired from Archaea and Conjugative Elements

The XP-C/Rad4 contains a fusion of the ArdC-N domains (BHD_1:3) to an inactive TGL domain that displays the papain-like peptidase fold. This TGL domain is specifically related to the catalytically active version found in the PNGase. They are unified by a conserved C-terminal extension to the exclusion of other members of the TGL superfamily (Anantharaman et al., 2001). However, the origin of this architectural linkage between the TGL and ArdC-N domains in XPC/Rad4 was largely obscure. To elucidate this evolutionary event, we surveyed the genomic data available since our earlier study relating to the TGL domain. Iterative PSI-BLAST searches recovered sequences from Thaumarchaeota (GenBank: ALI37408.1, OLE40692.1, OLC36996.1) as the closest related non-eukaryotic homologs of the XP-C/Rad4-PNGase TGL domain. Using these sequences as search seeds, we recovered additional homologs of archaeal PNGases from Euryarchaeota and Candidatus Micrarchaeota belonging to the DPANN group of archaea; further searches with these recovered homologs from other archaeal lineages including the Crenarchaeota (Data S7). We then used these sequences together with the eukaryotic XP-C/Rad4 and PNGases to construct a phylogenetic tree based on their shared TGL domain. Earlier identified related TGL domains found in ky/cyk3 and YebA (Anantharaman et al., 2001) served as outgroups in this analysis. Saliently, the tree showed that that the TGL domains from Thaumarchaeota are the closest sister group of the eukaryotic XP-C/Rad4 and PNGase TGL domains (Data S8: tree files). In support of this specific relationship, the thaumarchaeal versions share with eukaryotic PNGases a unique Zn-binding domain N terminal to the TGL domain (Figure 3A). In contrast to the thaumarchaeal and eukaryotic versions, all other archaeal PNGase homologs are predicted cell surface proteins as indicated by their N-terminal signal peptide and C-terminal transmembrane helix (Figure 3A). This is consistent with the recent proposal that the archaeal progenitor of

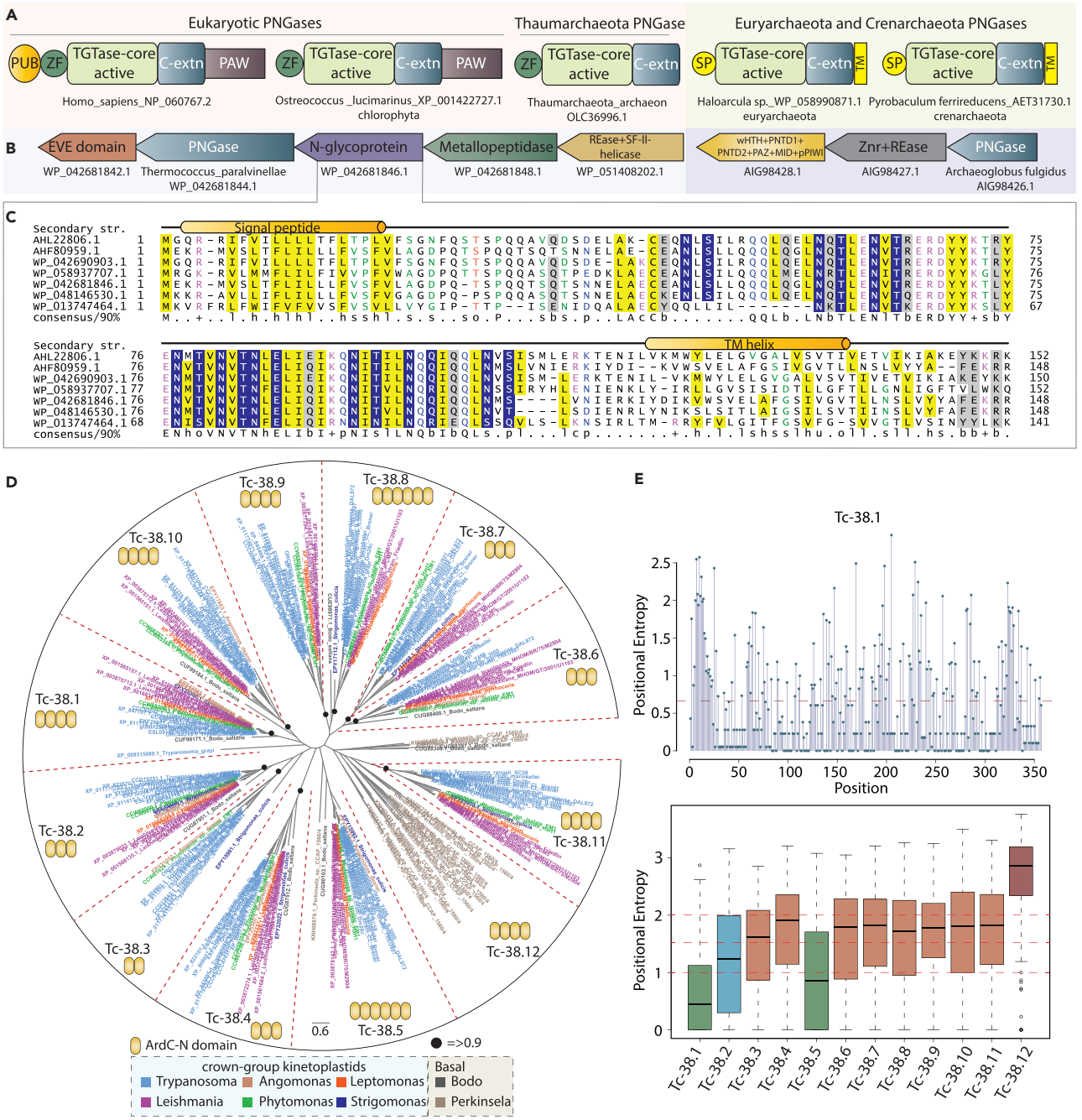


Figure 3. Domain Architectures and Gene Neighborhoods of PNGases and Classification of Tc-38 Family Proteins

(A) Domain architectures of archaeal and eukaryotic PNGases.

(B) Gene neighborhoods of PNGases from *Thermococcus paralvinellae* (GenBank: WP_042681844.1) and *Archaeoglobus fulgidus* (GenBank: AIG98426.1) are shown. The N-glycoprotein, a substrate for the active PNGase for de-N-glycosylation, is operonically linked.

(C) Multiple sequence alignment of the predicted N-glycoprotein showing characteristic NxT/S and NxE/Q repeats (highlighted in deep blue).

(D) Phylogenetic tree showing the Tc-38 family expansions grouped into 12 distinct clusters. *Perkinsela sp.* encoding a lineage-specific expansion is highlighted in brown. The number of ArdC-N domains present within the same polypeptide for each distinct cluster are shown as insets (domains colored yellow). Black dots denote branch support values (90% or more) estimated using ultrafast bootstrap method (1000 replicates) as implemented in the IQ-TREE software.

(E) Positional entropy values for the slow-evolving Tc-38.1 cluster with mean entropy value (dotted horizontal line in red). Furthermore, a box plot is shown comparing global entropy values for the 12 distinct Tc-38 clusters. Mean entropy values of >2, >1.5, <1.5, and <1 are colored deep brick, light brown, light blue, and green, respectively.

the eukaryotes was derived from within an assembly of archaea, which includes the Thaumarchaeota and the Asgardarchaeota (Zaremba-Niedzwiedzka et al., 2017).

Examination of the gene neighborhood contexts of the closest archaeal homologs of the eukaryotic XP-C/Rad4-PNGase TGL domains (Figure 3B; Data S7) showed that in several Euryarchaeota the PNGase homolog is coupled with a gene coding for a predicted membrane-associated protein of 160–180 amino acids with six conserved repeats containing an NxT/S motif and two additional NxQ/E motifs (Figure 3C). Such repeats with conserved asparagines suggest that this polypeptide is an N-glycoprotein and is the likely substrate for de-N-glycosylation by the associated PNGase homolog. Sequence alignments demonstrated that, like the eukaryotic PNGases, but unlike XPC/Rad4, all archaeal homologs, including Thaumarchaeota, retain an intact catalytic triad, suggesting that they are active enzymes (Data S9) that have the capacity to catalyze a PNGase-like reaction. Together these observations indicate that the archaeal PNGase homologs might function similar to the eukaryotic PNGases. Interestingly, these archaeal PNGases also show further linkages to genes encoding proteins with an EVE domain with the PUA-like fold implicated in the recognition of modified bases in DNA (Iyer et al., 2013), an REase-fold DNase domain, and in some cases, a PIWI family protein involved in RNA-dependent restriction (Aravind and Koonin, 2000; Burroughs et al., 2013, 2014; Zhang et al., 2016) (Figure 3B). Hence, these gene neighborhoods are likely to code for systems potentially involved in the discrimination and restriction of invasive elements with DNA containing modified bases, similar to other such biological conflict systems with comparable components that are widespread across prokaryotes (Iyer et al., 2009, 2011, 2013). However, the noteworthy twist in these systems is the inclusion of a cell-surface PNGase and/or an N-glycosylated glycoprotein component. Therefore, these appear to be dual action systems featuring cell-surface components interacting with the invasive elements in the extracellular compartment to potentially preclude their entry via de-glycosylation of surface receptors and associated restriction components that would target their DNA intracellularly.

In terms of phyletic spread, PNGases are present across all sampled major eukaryotic lineages and XPC/Rad4 is present in most except for the diplomonads, breviatea, and centroheliozoa (Data S10). These observations imply that the ancestral eukaryote likely possessed a PNGase-like protein inherited from their archaeal progenitor. Among the basal eukaryotic lineages, parabasalids (e.g., *Trichomonas*) possess a second paralogous PNGase-like protein, which, like XPC/Rad4, is inactive but lacks an association with the ArdC-N domains. Thus, early in eukaryotic evolution, the ancestral PNGase gave rise to two paralogs, one retaining the catalytic activity and the second becoming inactive as seen in the parabasalids. The earliest occurrences of the inactive TGL together with the ArdC-N domains are found in euglenozoans and heteroloboseans (Data S10). Hence, only after the basal-most eukaryotic lineages branched off, this inactive version appears to have fused with the ArdC-N domain acquired from a bacterial conjugative element, resulting in the core architecture of the extant XPC/Rad4. Within the context of the XPC/Rad4, the ArdC-N domain further underwent a triplication.

Distinct Components Recognize DNA Lesions Repaired by NER in Archaea and Eukaryotes

The above-reconstructed scenario allows us to explain the conundrum of the unique components of the NER system found in eukaryotes. NER, which is required for the removal of bulky and helix-distorting alterations of DNA (Shuck et al., 2008) (e.g., thymine dimers), is found across the bacterial and archaeo-eukaryotic branches of life. In both bacterial and archaeo-eukaryotic branches of life, the mechanistic details are comparable with distinct components that, respectively, recognize the DNA lesion and catalyze local unwinding of the DNA duplex and the incision endoDNases, which cut out the ssDNA segment with the lesion (Rouillon and White, 2011). However, these components are unrelated in the bacterial and the archaeo-eukaryotic branches of life. In the former, it is mediated by the UvrABC complexes (Webster et al., 2012), whereas in the latter, XPB/Rad25 and XPD/Rad3 helicases catalyze the unwinding activity and XPF/ERCC1 functions as one of the conserved incision endoDNases (Coin et al., 1998; Evans et al., 1997; Fan and DuPrez, 2015; Prakash and Prakash, 2000). However, the lesion-recognition component is not conserved between archaea and eukaryotes. Studies in archaea have implicated the SSB(RPA) protein in this role, which also has an important function in binding ssDNA during replication initiation (Cubeddu and White, 2005). Hence, our finding that early in eukaryotic evolution the ArdC-N domains were recruited for the lesion recognition role indicates that an alternative protein with comparable ssDNA recognition capacity to SSB(RPA) displaced it. Notably, our analysis also shows that it was likely acquired by the eukaryotes from the replication system of a plasmid/conjugative transposon probably borne by a bacterial endosymbiont. This might suggest that the ssDNA replication intermediates of these conjugative elements were

better structural analogs of the DNA lesions that trigger NER in eukaryotes and the ArdC-N domain was accordingly better equipped to recognize them than the ancestral SSB(RPA).

Co-option of the TGL Domain for Stabilizing XPC/Rad4 against Proteasomal Destruction

In eukaryotic NER, other proteins hr23B/Rad23B, centrin-2, DDB1, DDB2, and cullin-4 augment DNA lesion-recognition by XPC/Rad4 either by increasing its stability, binding efficiency and/or increasing the range of distinct DNA adducts that are recognized (Araki et al., 2001; Fitch et al., 2003; Nishi et al., 2005; Reardon et al., 1996; Sugasawa et al., 1996). Some of these, namely, DDB1, DDB2, and cullin-4, emerged relatively late in eukaryotic evolution via duplication of more widely distributed paralogs (A.M. Burroughs, L. Aravind, unpublished). Centrin-2 and hr23B/Rad23B are present across eukaryotes, including the basal lineages, and are involved in more general functions (Boutros et al., 2011; Chen and Madura, 2008; Cunningham et al., 2014; Dantuma et al., 2009; Martinez-Sanz et al., 2010; Resendes et al., 2008). This points to a recruitment to NER from these more general, ancestral functions. Most notable among these is the hr23B/Rad23, which contains a ubiquitin-like domain, an XPC/Rad4-binding domain (R4BD), and two ubiquitin-associated (UBA) domains. Interestingly, the R4BD interacts with the TGL domain of both XPC/Rad4 and PNGase (Hirayama et al., 2015; Suzuki et al., 2001). In this context, our findings also help explain how the other module of XPC/Rad4, the TGL domain, was recruited for the NER from a protein originally involved in glycoprotein degradation (PNGase). The shared architectural features of thaumarchaeal and eukaryotic PNGases suggests that an enzyme for the intracellular degradation of misfolded N-glycosylated glycoproteins likely emerged in their common ancestor from an earlier archaeal PNGase-like cell-surface protein involved in deglycosylation of cell-surface glycans. By the time of the ancestral eukaryote this was coupled with ubiquitin-mediated protein degradation through the 26S proteasomal system, to which the PNGase is recruited by a complex formed with hr23B/Rad23 (Hirayama et al., 2015; Suzuki et al., 2001, 2016). The similar Rad4•Rad23 complex formed via the inactive TGL domain instead protects XPC/Rad4 from proteasomal degradation and allows it to more efficiently complete NER (Dantuma et al., 2009). Given that the ArdC-N domain of the bacterial conjugative elements is also most frequently fused to an active or inactive peptidase domain (Figure 1A) (Iyer et al., 2017), it is probable that the version first acquired by the eukaryotes not only recognized DNA lesions but also mediated other interactions via this peptidase domain. This was likely displaced by the inactive peptidase-like TGL domain from the PNGase paralog, which now allowed a comparable interaction with hr23B/Rad23 and its sequestration for stabilizing XPC/Rad4 against proteasomal destruction. This opposing regulation via proteasomal degradation and hr23B/Rad23 stabilization was probably selected because it allowed a threshold-based regulation of the DNA-incision activity of NER, which if unregulated might result in unnecessary and deleterious DNA breaks.

The Recruitment of ArdC-N in Kinetoplastids and Its Subsequent Lineage-Specific Expansions

Other than in kinetoplastids, we failed to identify any Tc-38-like proteins from other major classes of euglenozoans such as the Diplonemida, Euglenida, and Symbiontida. However, it should be noted that the sequence data from the other euglenozoans is currently limited. Two distinct classification methods indicated that these Tc-38-like proteins form 12 clades of paralogs (see Methods). We name these clades Tc-38.1 to Tc-38.12 (Data S4). Eleven of these (Tc-38.1–11) contain members from the crown kinetoplastid lineage, the Trypanosomatidae (*Trypanosoma*, *Leishmania*, *Leptomonas*, *Phytomonas*, *Angomonas*, *Strigomonas*) (Figure 3D). Furthermore, 7 of these 11 paralog clades could be traced back to the more basal *Bodo saltans*, with the versions from this organism emerging as the basal-most member of each of those clades in the phylogenetic tree (Figure 3D). Two additional Tc-38-like proteins in *Bodo* did not unambiguously associate with any of these paralogous clades. The 12th clade of TC-38-like proteins is exclusively composed of a large lineage specific expansions (LSE) of 30 members from the basal kinetoplastid *Perkinsela* (Figure 3D). Hence, the most plausible evolutionary scenario would be that the representatives of at least seven of these paralogous clusters were already present in the last common ancestor of Bodonidae and Trypanosomatidae, suggesting an early diversification via LSE resulting in the founding members of the paralogous groups, somewhere close to the divergence of Bodonidae from *Perkinsela*. Given that the other euglenozoans apparently lack homologs of Tc-38, the second currently known acquisition of ArdC-N domains in eukaryotes likely happened at the base of kinetoplastida through the transfer of an ArdC-N domain from a plasmid/conjugative element (see below) probably residing in an endosymbiotic bacterium. In the course of the expansion of this protein family during kinetoplastid evolution we observe the following: (1) Repeated duplications of the ArdC-N domain and in some cases losses of the ArdC-N domain within the same polypeptide. Thus, these paralog clades are characterized by anywhere between

two and six copies of ArdC-N domains, with three or four repeats being most frequent (Figure 3D and Data S4). (2) Extreme sequence diversification of the individual ArdC-N domain repeats both within the same polypeptide and across members from the same paralog clade. (3) Structural divergence within the loop region in the central β-hairpin of the domain and in some cases expansions of long regions of low complexity linking adjacent ArdC-N domains (Data S4).

Evidence for Tc-38-like Proteins Playing a Key Role in kDNA Replication in Kinetoplastids

Our analysis of the diversification of the Tc-38 family also throws light on their functions in the kDNA replication system and the diversification of kDNA structures. The universal minicircle sequence (UMS) sequence with a G-quartet at the start (e.g., UMS in *Trypanosoma brucei* reads GGGGTTGGTGTA) is characteristic of the minicircle replication origins (Liu et al., 2005; Lukes et al., 2002; Ntambi et al., 1986). Tc-38 was shown to have a specific preference for binding TG-rich repeat regions comprising the UMS (Duhagon et al., 2009). Moreover, it also binds the hexamer, a sequence found on the lagging template strand at the start of the first Okazaki fragment. However, a similar UMS-binding function has also been attributed to another protein named the universal minicircle-binding protein (UMSBP) (Tzfati et al., 1992, 1995). This has led to the debate as to whether Tc-38 or UMSBP is the primary ssDNA-binding protein in the kDNA replication system. We present several lines of evidence based on our analysis of the Tc-38-like proteins and the available experimental data that have a bearing on which of these is the principal minicircle origin recognition protein. First, Tc-38-like proteins are specific to kinetoplastids, as would be expected for a protein with a kDNA-specific function. However, UMSBP belongs to a family of proteins with the Zn-knuckle that are distributed throughout eukaryotes and implicated in functions such as ssRNA binding in various RNA-processing systems like the mRNA maturation and splicing apparatuses (Anantharaman et al., 2002). Specifically, UMSBP does not exhibit the diversity and the expansions of the Tc-38-like family, complementing the diversification of kDNA circles in kinetoplastids (see below). Second, whereas the ArdC-N domain binds ssDNA in the related context of bacterial conjugative element replication intermediates, the CCHC-type zinc (Zn)-knuckle repeats found in UMSBP are known to be promiscuous in binding single-stranded nucleic acids and G-quartet structures (Benhalevy et al., 2017). Indeed, UMSBP, in addition to binding the template strand of the UMS, also quite promiscuously binds the complementary strand to the hexamer instead of the template strand (Abu-Elneel et al., 1999). Third, although the RNAi knockdowns of UMSBP led to defects in kDNA circle replication, the primary effects are seen in the segregation of daughter networks with additional effects on nuclear division (Milman et al., 2007). Tc-38 is enriched in the mitochondrial fraction across most stages of the kinetoplastid life cycle, and its RNAi knockdown results in the loss of kDNA and accumulation of "free" minicircles detached from the central network, known as the fraction S minicircles (Duhagon et al., 2009; Liu et al., 2006). Together, these observations favor Tc-38 being the primary UMS-associated circle replication origin-binding protein, with UMSBP likely playing a more general role in ssDNA binding in the kinetoplast and elsewhere.

In functional terms, based on the model of the triple ArdC-N domains in XPC/Rad4 (Min and Pavletich, 2007) we propose that the Tc-38-like proteins in kinetoplastids can bind a considerable DNA stretch encompassing up to 24 base pairs (Min and Pavletich, 2007) or more associated with the minicircle replication origins. Moreover, in parallel to the ArdC-N domains in prokaryotic conjugative elements, Tc-38 likely acts as a ssDNA-protecting agent (Duhagon et al., 2003). Tc-38-like proteins probably protect the integrity of the kDNA circles by providing a bulwark against supercoiling and enabling polymerase procession (Liu et al., 2006) similar to the role of ArdC-N during the transfer of the ssDNA replication intermediate of a conjugative element to a new host. These observations suggest that Tc-38 binds single-stranded minicircle DNA following initial helicase and topoisomerase activity, stabilizing the unwound structure during the initiation of leading and lagging strand syntheses. Loss of Tc-38, therefore, results in a continuation of the unwinding without initiation of synthesis (Liu et al., 2006). Other observations point to a more pervasive role for the diversified repertoire of Tc-38-like proteins in kinetoplastids paralleling the evolutionary diversification of kDNA. We observed varying degrees of sequence divergence across the Tc-38 paralog clades ranging from well-conserved slow-evolving clades to rapidly diversifying clades (Figure 3E). For instance, Tc-38.1, which contains the minicircle replication origin binding Tc-38 protein (also known as p38), is strongly conserved across both bodonids and trypanosomatids (as indicated by a low entropy value) (Figure 3E). Such conservation indicates a selective pressure to likely preserve both the binding features of the Tc-38.1 and the corresponding sequence determinants recognized by it. However, other paralogous clusters display considerable sequence diversity (Figure 3E). Notably, the knockdown of the Tc-38 gene by itself does not appear to affect all DNA circles (Liu et al., 2006), suggesting that some of the paralog clades

have probably been optimized for the recognition of sequences in specific groups of circles. Furthermore, a great diversity of distinctive kDNA network structures have been described across kinetoplastids. For instance, kDNA network structure observed in different basal kinetoplastids include pro- and poly-kDNA, where minicircles exist as monomeric units and often are covalently closed and topologically relaxed; pan-kDNA where minicircles are mainly monomeric, but unlike pro- and poly-kDNA they form a supercoiled structure; and mega-kDNA, an unusual form of kDNA where they do not constitute actual minicircles, but minicircle-like sequences are tandemly linked to larger maxicircle-like molecules (Lukes et al., 2002). Thus, the observed ArdC-N domain diversity in the Tc-38-like family might have gone hand in hand with the emergence of diverse kDNA network structures.

Tc-38 Is Joined by Several Other Components of the kDNA Replication Apparatus in Being Derived from Selfish Replicons

To contextualize our observation concerning the provenance of the Tc-38-like proteins we expanded our analysis of other kDNA replication components. Importantly, we observed that in a phylogenetic tree the kinetoplastid topoisomerase IA is lodged within a clade, which is otherwise mostly composed of topoisomerases from bacterial plasmids/conjugative elements that also code for an ArdC-N domain protein (e.g., *Helicobacter cetorum* plasmid topo IA, GenBank: AFI04135.1; Figure 4A). Strikingly, these mobile element topoisomerases are linked in a conserved gene neighborhood, with the gene for the ArdC-N domain protein as their immediate downstream neighbor (e.g., *Helicobacter cetorum* plasmid ArdC-N, GenBank AFI04134.1) (Figure 4A). This suggests that both the topo IA and the ArdC-N precursor of the Tc-38 protein were likely to have been acquired from a common plasmid source and recruited together for kDNA replication. The primary kDNA polymerases are members of the bacterial Pol I family. In the phylogenetic tree, the three paralogous kinetoplastid enzymes Pol IB, IC, and ID formed respective clusters with internal branching, largely in agreement with kinetoplastid phylogeny; the homologs from bacteriophage-infecting gammaproteobacteria are placed basal to these (Figure 4B). This indicates that the acquisition of the DNA pol I from a bacteriophage likely occurred in the common ancestor of the kinetoplastids and subsequently underwent multiple rounds of duplication giving rise to the four paralogs found across all kinetoplastids. Trypanosomatids possess two ATP-dependent DNA ligases k-α and k-β that are involved in kDNA replication. Interestingly, our sequence profile searches recovered a third divergent and previously unreported kinetoplastid-specific ATP-dependent ligase in both trypanosomatids and *Bodo saltans* (e.g., GenBank: CUI15152.1), which in searches recover ligases k-α and k-β, suggesting a specific relationship to the known kDNA ligases (Figure 4C). These could be traced back to *Bodo*, whereas the homologs from *Perkinsela* are placed basal to the overall clade separating k-α and k-β, suggesting that the emergence of the three distinct ligase clades happened in the common ancestor of trypanosomatids and *Bodo* (Figure 4C). In turn, all kDNA ligases form a clade with those from gammaproteobacterial bacteriophages to the exclusion of other DNA ligases (Figure 4C). These results emphatically established that at least two groups of key kDNA replicative enzymes, the DNA pol I and the ATP-dependent DNA ligases, were acquired from gammaproteobacterial bacteriophages. Aside from these, a member of the archaeo-eukaryotic primase superfamily PPL1, which functions as both a primase and a polymerase (primpol), has been proposed to be involved in both mitochondrial and nuclear DNA replication (Garcia-Gomez et al., 2013). We have earlier demonstrated that all eukaryotic primpols were ultimately inherited from a nucleocytoplasmic large DNA virus (NCLDV)-like source (Burroughs and Aravind, 2016; Iyer et al., 2005). Thus, several key kDNA replication components appear to have been acquired from not only plasmids/conjugative elements but also bacteriophages and eukaryotic viruses.

These are joined by a second category of kDNA replication proteins, which either have a deep evolutionary history as mitochondrial replication proteins or function as in eukaryotic nuclear replication. These include two other topoisomerases (DNA topoisomerase IB and IIA) involved in topological manipulations of kDNA (Bakshi and Shapiro, 2004; Scocca and Shapiro, 2008; Strauss and Wang, 1990), DNA polymerase κ homologs of the polY family functioning in translesion repair in most eukaryotes (Ohmori et al., 2009; Rajao et al., 2009), two DNA polymerase β (DNA pol-β) enzymes (including DNA pol-β-PAK [Saxowsky et al., 2003]), six distinct helicases (TbPIF1, TbPIF2, TbPIF4, TbPIF5, TbPIF7, and TbPIF8 [Liu et al., 2009]) related to the eukaryotic dual mitochondrial and nuclear PIF1-type helicase (Boule and Zakian, 2006), and the TbHslVU protease (Li et al., 2008). The primary eukaryotic mitochondrial DNA polymerase pol Q/pol θ (Pol IA in *T. brucei*) was retained in kinetoplastids but was apparently relegated to a role in repair (Hoeijmakers and Weijers, 1980; Klingbeil et al., 2002). The third category of proteins recruited to kDNA replication includes genuinely kinetoplastid-specific proteins lacking homologs elsewhere. One of these, p93, appears

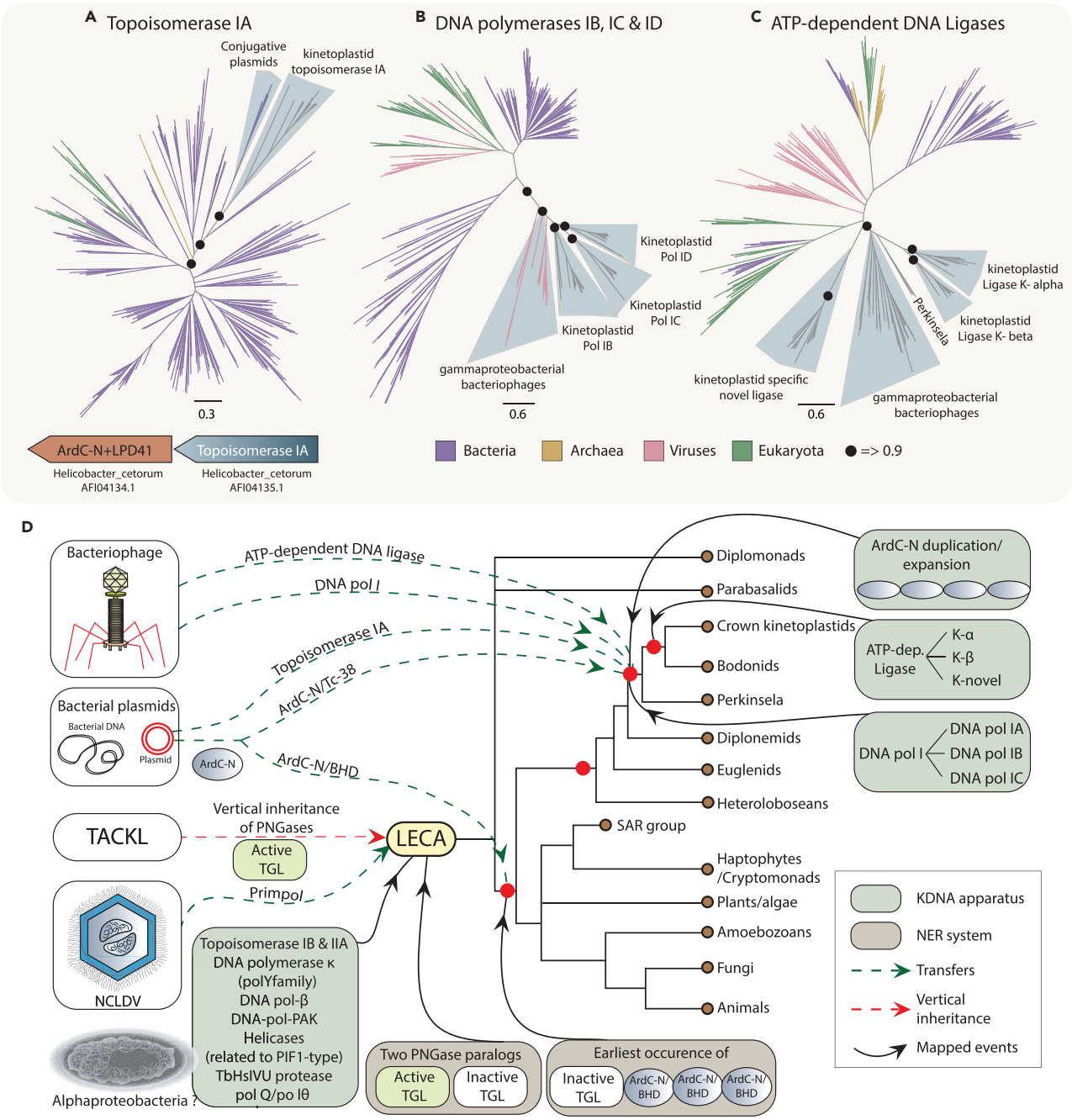


Figure 4. Evolutionary Scenarios for kDNA Replication and NER Components

(A–C) Phylogenetic trees showing key members of the kDNA replication apparatus, namely, topoisomerase IA (A), DNA polymerases IB, IC, and ID (B), and ATP-dependent DNA ligases (C), being acquired from plasmids/conjugative elements or bacteriophages. Operonically linked ArdC-N and topoisomerase IA in plasmids/conjugative elements are shown for *Helicobacter cetorum*.

(D) Schematic diagram showing the summary of the evolutionary events that shaped the kinetoplastid replication and NER systems in eukaryotes. Major evolutionary events are mapped on to a simplified eukaryotic tree. Transfers from plasmids/conjugative elements, bacteriophages, and NCLDV sources are highlighted in green dotted arrows. Events related to the kDNA replication apparatus and the NER repair system are highlighted in green and gray boxes, respectively, and are mapped to the respective nodes on the eukaryotic tree (using curved lines in black with arrowheads).

to have a poorly understood role in minicircle replication (Li et al., 2007). We could trace the provenance of p93 to *Bodo*; however, no homolog was found in *Perkinsela*.

DISCUSSION

In this study, we show that the N-terminal domain of the anti-restriction protein ArdC of plasmids/conjugative elements was acquired by eukaryotes on two independent occasions and incorporated into two different functional systems. The first acquisition, which happened early in eukaryotic evolution, gave rise to the so-called BHDs of the XPC/Rad4. The second acquisition in the common ancestor of the kinetoplastids gave rise to the Tc-38 like proteins that extensively diversified in the context of the kDNA replication and dynamics. In both instances, the ArdC-N domain underwent multiple duplications within the same polypeptide in eukaryotes to form an extensive DNA-binding interface that could span a substantial length of DNA. In the case of the BHDs, this was accompanied by substantial structural modifications in each of the copies. These findings resolve two outstanding questions regarding very different systems in eukaryotes. One, with more general implications, explains the origin of the eukaryote-specific primary DNA lesion-recognition component in the NER system, several of whose core components were inherited from the archaeal progenitor (Rouillon and White, 2011; Scharer, 2013; Shuck et al., 2008). We show that a mobile-element-derived DNA-binding protein, potentially acquired through a plasmid-bearing endosymbiont, was re-purposed for this function supplanting the ancestral protein SSB(RPA) in this system process. The other acquisition, again likely occurring via a plasmid-bearing bacterial endosymbiont, with more specific implications helps explain how the plasmid-like replication and unusual structure of the kDNA of kinetoplastids might have emerged. In the case of the Tc-38-like proteins their role in kDNA replication is likely to be similar to their ancestral role in mobile element replication, whereas in the case of XPC/Rad4 there was a shift to a DNA repair role. In this regard, it might be noted that other DNA-binding domains involved in DNA repair, such as the MutS-I domain, are also shared with the plasmid replication proteins like ArdC-N (Iyer et al., 2017). In both cases, the incorporation of the ArdC-N domains into eukaryotic systems involved a complex history of physical and/or functional linkage with other components with shared or distinct evolutionary trajectories. In the case of the XPC/Rad4 proteins, they were fused to an inactive TGL related to the active version found in PNGases that catalyze sugar removal in glycan degradation. Here we establish that these PNGase domains have a deep history in archaea, where they likely performed an ancestral role in the removal of sugar moieties from proteins. The fusion of the TGL domain with the ArdC-N domain was likely a key factor in the emergence of the eukaryote-specific regulation of NER via the ubiquitin-proteasome system.

From our survey, it is apparent that the kDNA replication apparatus represents a major reconfiguration of the ancestral mitochondrial replication system inherited by eukaryotes from an alphaproteobacterial ancestor (Aravind et al., 2006, 2012; Roger et al., 2017). This seems to coincide with the previously documented euglenozoan mitochondrial genome scrambling event following which distinctly modified mitochondrial genome structures emerged in different lineages. Based on the mechanistic details of kDNA replication it had been proposed that the precursor of the kDNA circles was perhaps a plasmid harbored within the mitochondrion of the common ancestor of kinetoplastids (Lukes et al., 2002). Our study objectively evaluates this proposal and presents an evolutionary scenario for the origin of kDNA replication system (Figure 4D). First, our analysis indicates that the key components of the kDNA replication system emerged from independent replicons found in bacteria, namely, plasmids/conjugative elements and bacteriophages. There is no evidence for these elements residing in the eukaryotic mitochondrion at a time long after the stem eukaryote, where the endosymbiosis first occurred. However, eukaryotes including kinetoplastids harbor other bacterial endosymbionts (e.g., certain bodonids, strigomonads, and novymonads [Catta-Preta et al., 2015; Du et al., 1994; Harmer et al., 2018]), which might contain such selfish elements. As the mitochondrial genome degenerated, it reached a size comparable to that of these selfish replicons, which code for their own replication proteins. Given that these are likely optimized for the dedicated replication of such small replicons, they in part displaced the ancestral components of the mitochondrial replication system. In parallel, as we had earlier shown (Burroughs and Aravind, 2016), central components of the RNA editing system of the kinetoplast such as the RNA ligases and end-processing enzymes were also derived from bacteriophage RNA repair systems. Together, with the acquisition of key DNA replication components, such as the Tc-38 and Topo IA from plasmids, DNA pol I and ATP-dependent ligases from bacteriophages, and the primpol from NCLDVs, the above probably facilitated the unique structural developments that characterize kDNA. In addition, they were supplemented by components already present in the ancestral eukaryote as well as lineage-specific innovations.

These findings add to a growing body of case studies that are revealing key aspects of the provenance of eukaryote-specific systems. For instance, the Vrr-Nuc, a nuclease from bacterial selfish elements, has been recruited into a parallel eukaryotic DNA repair complex, the Fanconi anemia complex, required for the repair of interstrand cross-links (Iyer et al., 2006; MacKay et al., 2010). Strikingly, an associated ssDNA-binding domain, the HIRAN domain, which is also derived from bacterial selfish elements (Hishiki et al., 2015; Iyer et al., 2006; Kile et al., 2015), has been recruited in eukaryotes for the recognition of the 3′ends of stalled replication forks to facilitate their regression to control DNA damage. In both cases, these domains acquired from the bacterial selfish elements have been physically or functionally linked to components of the ubiquitin-proteasome system to regulate the actual process of DNA repair. Similarly, we had earlier shown that another eukaryotic ssDNA-binding protein involved in recombination, Rad52, had its origin in ssDNA-binding proteins involved in bacteriophage genome recombination (de Souza et al., 2010; Iyer et al., 2002). Thus, the sudden emergence of several components and systems for the safeguarding of the larger and linearly segmented genomes of eukaryotes appear to have widely recruited pre-adapted catalytic and DNA-binding functions found in prokaryotic selfish elements. This emphasizes the role of selfish elements resident in bacterial endosymbionts, both from the mitochondrial progenitor and from others, as major contributors to eukaryote-specific systems. In more general terms, over the years we have presented several lines of evidence for the extensive re-use of domains that ultimately originated in prokaryotic biological conflict systems in eukaryote-specific systems (Aravind et al., 2012, 2014). The opportunities for rapid evolution seen in such conflict systems, which are under intense selective pressures due to their direct effects of fitness, has enabled them to explore a wide structure and substrate space thereby furnishing pre-adaptations that could be utilized elsewhere (Zhang et al., 2014).

In conclusion, our findings can help guide specific experiments on DNA recognition components both in the context of NER and kDNA dynamics. Furthermore, our discovery of bacterial prototypes for some key DNA recognition components in eukaryotes provides a clear picture of the structural diversity of domains, such as the ArdC-N domain. This could help guide the structural engineering of DNA-binding domains for recognition of specific features such as lesions in DNA.

Limitations of the Study

Both in the context of the kDNA replication dynamics and the XPC/Rad4 NER system tracing the earliest eukaryotic occurrence of the ArdC-N domain or the time point of its recruitments in eukaryotes is based on the currently publicly available genomic datasets. Although we have comprehensively surveyed all available eukaryotic genomes, the genomic data for euglenozoans (excluding the kinetoplastids), such as the Diplonemida, Euglenida, and Symbiontida, as well as for some of the basal excavates are limited. Future availability of genomic sequences from these lineages would further help establish if there are any alterations to the precise temporal reconstruction of the two independent recruitment events of the ArdC-N domain reported in this study.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods and 10 data files and can be found with this article online at https://doi.org/10.1016/j.isci.2018.10.017.

ACKNOWLEDGMENTS

This work was supported by an NIH postdoctoral visiting fellowship (A.K.) and the intramural funds (L.M.I., A.M.B., and L.A.) of the National Library of Medicine at the National Institutes of Health, USA. The funding institute had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

AUTHOR CONTRIBUTIONS

L.A., A.M.B., and L.M.I. conceived the project and directed its management. A.K., A.M.B., and L.A. performed all computational analyses and analyzed the data. A.K. and A.M.B. wrote the first draft of the manuscript. A.K. prepared all the figures and Supplemental Information. L.A. edited the manuscript to prepare the final version. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

All identified sequences and complete data associated with the manuscript is also available at the FTP site: ftp://ftp.ncbi.nih.gov/pub/aravind/ardcn/ardcn.html.

Received: July 17, 2018
Revised: October 15, 2018
Accepted: October 15, 2018
Published: November 30, 2018

REFERENCES

Abu-Elneel, K., Kapeller, I., and Shlomai, J. (1999). Universal minicircle sequence-binding protein, a sequence-specific DNA-binding protein that recognizes the two replication origins of the kinetoplast DNA minicircle. J. Biol. Chem. *274*, 13419–13426.

Alva, V., Nam, S.Z., Soding, J., and Lupas, A.N. (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res. *44*, W410–W415.

Anantharaman, V., Koonin, E.V., and Aravind, L. (2001). Peptide-N-glycanases and DNA repair proteins, Xp-C/Rad4, are, respectively, active and inactivated enzymes sharing a common transglutaminase fold. Hum. Mol. Genet. *10*, 1627–1630.

Anantharaman, V., Koonin, E.V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. *30*, 1427–1464.

Araki, M., Masutani, C., Takemura, M., Uchida, A., Sugasawa, K., Kondoh, J., Ohkuma, Y., and Hanaoka, F. (2001). Centrosome protein centrin 2/caltractin 1 is part of the xeroderma pigmentosum group C complex that initiates global genome nucleotide excision repair. J. Biol. Chem. *276*, 18665–18672.

Aravind, L., Abhiman, S., and Iyer, L.M. (2011). Natural history of the eukaryotic chromatin protein methylation system. Prog. Mol. Biol. Transl. Sci. *101*, 105–176.

Aravind, L., Anantharaman, V., Zhang, D., de Souza, R.F., and Iyer, L.M. (2012). Gene flow and biological conflict systems in the origin and evolution of eukaryotes. Front. Cell Infect. Microbiol. *2*, 89.

Aravind, L., Burroughs, A.M., Zhang, D., and Iyer, L.M. (2014). Protein and DNA modifications: evolutionary imprints of bacterial biochemical diversification and geochemistry on the provenance of eukaryotic epigenetics. Cold Spring Harb. Perspect. Biol. *6*, a016063.

Aravind, L., Iyer, L.M., and Koonin, E.V. (2006). Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr. Opin. Struct. Biol. *16*, 409–419.

Aravind, L., and Koonin, E.V. (2000). Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system. Genome Res. *10*, 1172–1184.

Aravind, L., Walker, D.R., and Koonin, E.V. (1999). Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. *27*, 1223–1242.

Bakshi, R.P., and Shapiro, T.A. (2004). RNA interference of *Trypanosoma brucei* topoisomerase IB: both subunits are essential. Mol. Biochem. Parasitol. *136*, 249–255.

Belogurov, A.A., Delver, E.P., Agafonova, O.V., Belogurova, N.G., Lee, L.Y., and Kado, C.I. (2000). Antirestriction protein Ard (Type C) encoded by IncW plasmid pSa has a high similarity to the "protein transport" domain of TraC1 primase of promiscuous plasmid RP4. J. Mol. Biol. *296*, 969–977.

Benhalevy, D., Gupta, S.K., Danan, C.H., Ghosal, S., Sun, H.W., Kazemier, H.G., Paeschke, K., Hafner, M., and Juranek, S.A. (2017). The human CCHC-type zinc finger nucleic acid-binding protein binds G-rich elements in target mRNA coding sequences and promotes translation. Cell Rep. *18*, 2979–2990.

Boule, J.B., and Zakian, V.A. (2006). Roles of Pif1-like helicases in the maintenance of genomic stability. Nucleic Acids Res. *34*, 4147–4153.

Boutros, R., Lorenzo, C., Mondesert, O., Jauneau, A., Oakes, V., Dozier, C., Gabrielli, B., and Ducommun, B. (2011). CDC25B associates with a centrin 2-containing complex and is involved in maintaining centrosome integrity. Biol. Cell *103*, 55–68.

Burroughs, A.M., Allen, K.N., Dunaway-Mariano, D., and Aravind, L. (2006). Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J. Mol. Biol. *361*, 1003–1034.

Burroughs, A.M., Ando, Y., and Aravind, L. (2014). New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. Wiley Interdiscip. Rev. RNA *5*, 141–181.

Burroughs, A.M., and Aravind, L. (2016). RNA damage in biological conflicts and the diversity of responding RNA repair systems. Nucleic Acids Res. *44*, 8525–8555.

Burroughs, A.M., Iyer, L.M., and Aravind, L. (2013). Two novel PIWI families: roles in inter-genomic conflicts in bacteria and Mediator-dependent modulation of transcription in eukaryotes. Biol. Direct *8*, 13.

Catta-Preta, C.M., Brum, F.L., da Silva, C.C., Zuma, A.A., Elias, M.C., de Souza, W., Schenkman, S., and Motta, M.C. (2015). Endosymbiosis in trypanosomatid protozoa: the bacterium division is controlled during the host cell cycle. Front. Microbiol. *6*, 520.

Chen, L., and Madura, K. (2008). Centrin/Cdc31 is a novel regulator of protein degradation. Mol. Cell. Biol. *28*, 1829–1840.

Clement, F.C., Camenisch, U., Fei, J., Kaczmarek, N., Mathieu, N., and Naegeli, H. (2010). Dynamic two-stage mechanism of versatile DNA damage recognition by xeroderma pigmentosum group C protein. Mutat. Res. *685*, 21–28.

Coin, F., Marinoni, J.C., Rodolfo, C., Fribourg, S., Pedrini, A.M., and Egly, J.M. (1998). Mutations in the XPD helicase gene result in XP and TTD phenotypes, preventing interaction between XPD and the p44 subunit of TFIIH. Nat. Genet. *20*, 184–188.

Cubeddu, L., and White, M.F. (2005). DNA damage detection by an archaeal single-stranded DNA-binding protein. J. Mol. Biol. *353*, 507–516.

Cunningham, C.N., Schmidt, C.A., Schramm, N.J., Gaylord, M.R., and Resendes, K.K. (2014). Human TREX2 components PCID2 and centrin 2, but not ENY2, have distinct functions in protein export and co-localize to the centrosome. Exp. Cell Res. *320*, 209–218.

Dai, J., Finci, L., Zhang, C., Lahiri, S., Zhang, G., Peisach, E., Allen, K.N., and Dunaway-Mariano, D. (2009). Analysis of the structural determinants underlying discrimination between substrate and solvent in beta-phosphoglucomutase catalysis. Biochemistry *48*, 1984–1995.

Dai, J., Wang, L., Allen, K.N., Radstrom, P., and Dunaway-Mariano, D. (2006). Conformational cycling in beta-phosphoglucomutase catalysis: reorientation of the beta-D-glucose 1,6-(Bis) phosphate intermediate. Biochemistry *45*, 7818–7824.

Dantuma, N.P., Heinen, C., and Hoogstraten, D. (2009). The ubiquitin receptor Rad23: at the crossroads of nucleotide excision repair and

proteasomal degradation. DNA Repair (Amst) *8*, 449–460.

de Souza, R.F., Iyer, L.M., and Aravind, L. (2010). Diversity and evolution of chromatin proteins encoded by DNA viruses. Biochim. Biophys. Acta *1799*, 302–318.

Du, Y., Maslov, D.A., and Chang, K.P. (1994). Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. Proc. Natl. Acad. Sci. U S A *91*, 8437–8441.

Duhagon, M.A., Dallagiovanna, B., Ciganda, M., Ruyechan, W., Williams, N., and Garat, B. (2003). A novel type of single-stranded nucleic acid binding protein recognizing a highly frequent motif in the intergenic regions of *Trypanosoma cruzi*. Biochem. Biophys. Res. Commun. *309*, 183–188.

Duhagon, M.A., Pastro, L., Sotelo-Silveira, J.R., Perez-Diaz, L., Maugeri, D., Nardelli, S.C., Schenkman, S., Williams, N., Dallagiovanna, B., and Garat, B. (2009). The Trypanosoma cruzi nucleic acid binding protein Tc38 presents changes in the intramitochondrial distribution during the cell cycle. BMC Microbiol. *9*, 34.

Evans, E., Moggs, J.G., Hwang, J.R., Egly, J.M., and Wood, R.D. (1997). Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. EMBO J. *16*, 6559–6573.

Fan, L., and DuPrez, K.T. (2015). XPB: an unconventional SF2 DNA helicase. Prog. Biophys. Mol. Biol. *117*, 174–181.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Fitch, M.E., Nakajima, S., Yasui, A., and Ford, J.M. (2003). In vivo recruitment of XPC to UV-induced cyclobutane pyrimidine dimers by the DDB2 gene product. J. Biol. Chem. *278*, 46906–46910.

Garcia-Gomez, S., Reyes, A., Martinez-Jimenez, M.I., Chocron, E.S., Mouron, S., Terrados, G., Powell, C., Salido, E., Mendez, J., Holt, I.J., et al. (2013). PrimPol, an archaic primase/polymerase operating in human cells. Mol. Cell *52*, 541–553.

Harmer, J., Yurchenko, V., Nenarokova, A., Lukes, J., and Ginger, M.L. (2018). Farming, slaving and enslavement: histories of endosymbioses during kinetoplastid evolution. Parasitology *145*, 1–13.

Hirayama, H., Hosomi, A., and Suzuki, T. (2015). Physiological and molecular functions of the cytosolic peptide:N-glycanase. Semin. Cell. Dev. Biol. *41*, 110–120.

Hishiki, A., Hara, K., Ikegaya, Y., Yokoyama, H., Shimizu, T., Sato, M., and Hashimoto, H. (2015). Structure of a Novel DNA-binding domain of helicase-like transcription factor (HLTF) and its functional implication in DNA damage tolerance. J. Biol. Chem. *290*, 13215–13223.

Hoeijmakers, J.H., and Weijers, P.J. (1980). The segregation of kinetoplast DNA networks in *Trypanosoma brucei*. Plasmid *4*, 97–116.

Ishikawa, K., Fukuda, E., and Kobayashi, I. (2010). Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. DNA Res. *17*, 325–342.

Iyer, L.M., Babu, M.M., and Aravind, L. (2006). The HIRAN domain and recruitment of chromatin remodeling and repair activities to damaged DNA. Cell Cycle *5*, 775–782.

Iyer, L.M., Burroughs, A.M., Anand, S., de Souza, R.F., and Aravind, L. (2017). Polyvalent proteins, a pervasive theme in the intergenomic biological conflicts of bacteriophages and conjugative elements. J. Bacteriol. *199*, e00245–17.

Iyer, L.M., Koonin, E.V., and Aravind, L. (2002). Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC Genomics *3*, 8.

Iyer, L.M., Koonin, E.V., Leipe, D.D., and Aravind, L. (2005). Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. Nucleic Acids Res. *33*, 3875–3896.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle *8*, 1698–1710.

Iyer, L.M., Zhang, D., Burroughs, A.M., and Aravind, L. (2013). Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. Nucleic Acids Res. *41*, 7635–7655.

Iyer, L.M., Zhang, D., Rogozin, I.B., and Aravind, L. (2011). Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic Acids Res. *39*, 9473–9497.

Jensen, R.E., and Englund, P.T. (2012). Network news: the replication of kinetoplast DNA. Annu. Rev. Microbiol. *66*, 473–491.

Kaur, G., Iyer, L.M., Subramanian, S., and Aravind, L. (2018). Evolutionary convergence and divergence in archaeal chromosomal proteins and Chromo-like domains from bacteria and eukaryotes. Sci. Rep. *8*, 6196.

Kile, A.C., Chavez, D.A., Bacal, J., Eldirany, S., Korzhnev, D.M., Bezsonova, I., Eichman, B.F., and Cimprich, K.A. (2015). HLTF's ancient HIRAN domain binds 3' DNA ends to drive replication fork reversal. Mol. Cell *58*, 1090–1100.

Klingbeil, M.M., Motyka, S.A., and Englund, P.T. (2002). Multiple mitochondrial DNA polymerases in *Trypanosoma brucei*. Mol. Cell *10*, 175–186.

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res. *29*, 3742–3756.

Li, Y., Sun, Y., Hines, J.C., and Ray, D.S. (2007). Identification of new kinetoplast DNA replication proteins in trypanosomatids based on predicted S-phase expression and mitochondrial targeting. Eukaryot. Cell *6*, 2303–2310.

Li, Z., Lindsay, M.E., Motyka, S.A., Englund, P.T., and Wang, C.C. (2008). Identification of a bacterial-like HslVU protease in the mitochondria of *Trypanosoma brucei* and its role in mitochondrial DNA replication. PLoS Pathog. *4*, e1000048.

Liu, B., Liu, Y., Motyka, S.A., Agbo, E.E., and Englund, P.T. (2005). Fellowship of the rings: the replication of kinetoplast DNA. Trends Parasitol. *21*, 363–369.

Liu, B., Molina, H., Kalume, D., Pandey, A., Griffith, J.D., and Englund, P.T. (2006). Role of p38 in replication of *Trypanosoma brucei* kinetoplast DNA. Mol. Cell. Biol. *26*, 5382–5393.

Liu, B., Wang, J., Yaffe, N., Lindsay, M.E., Zhao, Z., Zick, A., Shlomai, J., and Englund, P.T. (2009). Trypanosomes have six mitochondrial DNA helicases with one controlling kinetoplast maxicircle replication. Mol. Cell *35*, 490–501.

Lukes, J., Guilbride, D.L., Votypka, J., Zikova, A., Benne, R., and Englund, P.T. (2002). Kinetoplast DNA network: evolution of an improbable structure. Eukaryot. Cell *1*, 495–502.

MacKay, C., Declais, A.C., Lundin, C., Agostinho, A., Deans, A.J., MacArtney, T.J., Hofmann, K., Gartner, A., West, S.C., Helleday, T., et al. (2010). Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. Cell *142*, 65–76.

Maillard, O., Solyom, S., and Naegeli, H. (2007). An aromatic sensor with aversion to damaged strands confers versatility to DNA repair. PLoS Biol. *5*, e79.

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct *6*, 38.

Martinez-Sanz, J., Kateb, F., Assairi, L., Blouquit, Y., Bodenhausen, G., Abergel, D., Mouawad, L., and Craescu, C.T. (2010). Structure, dynamics and thermodynamics of the human centrin 2/hSfi1 complex. J. Mol. Biol. *395*, 191–204.

Miele, L., Strack, B., Kruft, V., and Lanka, E. (1991). Gene organization and nucleotide sequence of the primase region of IncP plasmids RP4 and R751. DNA Seq. *2*, 145–162.

Milman, N., Motyka, S.A., Englund, P.T., Robinson, D., and Shlomai, J. (2007). Mitochondrial origin-binding protein UMSBP mediates DNA replication and segregation in trypanosomes. Proc. Natl. Acad. Sci. U S A *104*, 19250–19255.

Min, J.H., and Pavletich, N.P. (2007). Recognition of DNA damage by the Rad4 nucleotide excision repair protein. Nature *449*, 570–575.

Nishi, R., Okuda, Y., Watanabe, E., Mori, T., Iwai, S., Masutani, C., Sugasawa, K., and Hanaoka, F. (2005). Centrin 2 stimulates nucleotide excision repair by interacting with xeroderma pigmentosum group C protein. Mol. Cell. Biol. *25*, 5664–5674.

Ntambi, J.M., Shapiro, T.A., Ryan, K.A., and Englund, P.T. (1986). Ribonucleotides associated with a gap in newly replicated kinetoplast DNA minicircles from *Trypanosoma equiperdum*. J. Biol. Chem. *261*, 11890–11895.

Ohmori, H., Hanafusa, T., Ohashi, E., and Vaziri, C. (2009). Separate roles of structured and unstructured regions of Y-family DNA polymerases. Adv. Protein Chem. Struct. Biol. *78*, 99–146.

Prakash, S., and Prakash, L. (2000). Nucleotide excision repair in yeast. Mutat. Res. *451*, 13–24.

Rajao, M.A., Passos-Silva, D.G., DaRocha, W.D., Franco, G.R., Macedo, A.M., Pena, S.D., Teixeira, S.M., and Machado, C.R. (2009). DNA polymerase kappa from *Trypanosoma cruzi* localizes to the mitochondria, bypasses 8-oxoguanine lesions and performs DNA synthesis in a recombination intermediate. Mol. Microbiol. *71*, 185–197.

Reardon, J.T., Mu, D., and Sancar, A. (1996). Overproduction, purification, and characterization of the XPC subunit of the human DNA repair excision nuclease. J. Biol. Chem. *271*, 19451–19456.

Rees, C.E., and Wilkins, B.M. (1990). Protein transfer into the recipient cell during bacterial conjugation: studies with F and RP4. Mol. Microbiol. *4*, 1199–1205.

Resendes, K.K., Rasala, B.A., and Forbes, D.J. (2008). Centrin 2 localizes to the vertebrate nuclear pore and plays a role in mRNA and protein export. Mol. Cell. Biol. *28*, 1755–1769.

Roger, A.J., Munoz-Gomez, S.A., and Kamikawa, R. (2017). The origin and diversification of mitochondria. Curr. Biol. *27*, R1177–R1192.

Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. *45*, D271–D281.

Rouillon, C., and White, M.F. (2011). The evolution and mechanisms of nucleotide excision repair proteins. Res. Microbiol. *162*, 19–26.

Saxowsky, T.T., Choudhary, G., Klingbeil, M.M., and Englund, P.T. (2003). *Trypanosoma brucei* has two distinct mitochondrial DNA polymerase beta enzymes. J. Biol. Chem. *278*, 49095–49101.

Scharer, O.D. (2013). Nucleotide excision repair in eukaryotes. Cold Spring Harb. Perspect. Biol. *5*, a012609.

Scocca, J.R., and Shapiro, T.A. (2008). A mitochondrial topoisomerase IA essential for late theta structure resolution in African trypanosomes. Mol. Microbiol. *67*, 820–829.

Shapiro, T.A. (1993). Kinetoplast DNA maxicircles: networks within networks. Proc. Natl. Acad. Sci. U S A *90*, 7809–7813.

Shuck, S.C., Short, E.A., and Turchi, J.J. (2008). Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. Cell Res. *18*, 64–72.

Smith, J.M., and Price, G.R. (1973). The logic of animal conflict. Nature *246*, 15.

Strauss, P.R., and Wang, J.C. (1990). The TOP2 gene of *Trypanosoma brucei*: a single-copy gene that shares extensive homology with other TOP2 genes encoding eukaryotic DNA topoisomerase II. Mol. Biochem. Parasitol. *38*, 141–150.

Sugasawa, K., Masutani, C., Uchida, A., Maekawa, T., van der Spek, P.J., Bootsma, D., Hoeijmakers, J.H., and Hanaoka, F. (1996). HHR23B, a human Rad23 homolog, stimulates XPC protein in nucleotide excision repair in vitro. Mol. Cell. Biol. *16*, 4852–4861.

Suzuki, T., Huang, C., and Fujihira, H. (2016). The cytoplasmic peptide:N-glycanase (NGLY1) - structure, expression and cellular functions. Gene *577*, 1–7.

Suzuki, T., Park, H., Kwofie, M.A., and Lennarz, W.J. (2001). Rad23 provides a link between the Png1 deglycosylating enzyme and the 26 S proteasome in yeast. J. Biol. Chem. *276*, 21601–21607.

Tzfati, Y., Abeliovich, H., Avrahami, D., and Shlomai, J. (1995). Universal minicircle sequence binding protein, a CCHC-type zinc finger protein that binds the universal minicircle sequence of trypanosomatids. Purification and characterization. J. Biol. Chem. *270*, 21339–21345.

Tzfati, Y., Abeliovich, H., Kapeller, I., and Shlomai, J. (1992). A single-stranded DNA-binding protein from *Crithidia fasciculata* recognizes the nucleotide sequence at the origin of replication of kinetoplast DNA minicircles. Proc. Natl. Acad. Sci. U S A *89*, 6891–6895.

Webster, M.P., Jukes, R., Zamfir, V.S., Kay, C.W., Bagneris, C., and Barrett, T. (2012). Crystal structure of the UvrB dimer: insights into the nature and functioning of the UvrAB damage engagement and UvrB-DNA complexes. Nucleic Acids Res. *40*, 8743–8758.

Werren, J.H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc. Natl. Acad. Sci. U S A *108* (*Suppl 2*), 10863–10870.

Yamaguchi, Y., Park, J.H., and Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. Annu. Rev. Genet. *45*, 61–79.

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature *541*, 353.

Zhang, D., Burroughs, A.M., Vidal, N.D., Iyer, L.M., and Aravind, L. (2016). Transposons to toxins: the provenance, architecture and diversification of a widespread class of eukaryotic effectors. Nucleic Acids Res. *44*, 3513–3533.

Zhang, D., de Souza, R.F., Anantharaman, V., Iyer, L.M., and Aravind, L. (2012). Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol. Direct *7*, 18.

Zhang, D., Iyer, L.M., Burroughs, A.M., and Aravind, L. (2014). Resilience of biochemical activity in protein domains in the face of structural divergence. Curr. Opin. Struct. Biol. *26*, 92–103.

ISCI, Volume *9*

Supplemental Information

Unexpected Evolution of Lesion-Recognition Modules in Eukaryotic NER and Kinetoplast DNA Dynamics Proteins from Bacterial Mobile Elements

Arunkumar Krishnan, A. Maxwell Burroughs, Lakshminarayan M. Iyer, and L. Aravind

Transparent Methods

Iterative sequence profile searches were performed using the PSI-BLAST (Altschul et al., 1997) and JACKHMMER (Eddy, 2009) programs against the nr protein database of the National Center for Biotechnology Information (NCBI). Similarity-based clustering implemented in the BLASTCLUST (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) was used for classification/paralogous grouping and culling of nearly identical sequences. The length (L) and score (S) threshold parameters were variably set between 1.2 to .3 to obtain clusters at different thresholds. Profile-Profile searches were run using the HHpred program either against the PDB or Pfam databases (Finn et al., 2016; Rose et al., 2017). Multiple sequence alignments were built using the KALIGN (Lassmann et al., 2009) and GISMO (Neuwald and Altschul, 2016) programs, followed by manual adjustments based on profile-profile and structural alignments. Secondary structures were predicted with the JPred (Cole et al., 2008) program. Phylogenetic relationships were derived using an approximate maximum likelihood (ML) method as implemented in the FastTree program (Price et al., 2010): corresponding local support values were also estimated as implemented. To increase the accuracy of topology in FastTree, we increased the number of rounds of minimum-evolution subtree-prune-regraft (SPR) moves to 4 (-spr 4) as well as utilized the options -mlacc and -slownni to make the ML nearest neighbor interchanges (NNIs) more exhaustive. Phylogenetic tree topologies were also derived using ML methods based on the edge-linked partition model as implemented in the IQ-TREE software (Nguyen et al., 2015): branch supports were obtained using the ultrafast bootstrap method (1000 replicates). Gene neighborhoods was retrieved by a Perl script that extracts the upstream and downstream genes of the query gene and uses BLASTCLUST to cluster the proteins to identify conserved gene neighborhoods. Position-wise Shannon entropy (H) was computed using a custom script written in the R language using the equation

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

where M is the number of amino acid types and P is the fraction of residues of amino acid type *i*. The Shannon entropy for any given position in the MSA ranges from 0 (absolutely conserved one amino acid at that position) to 4.32 (all 20 amino acid residues equally represented at that position). Structural visualization and manipulations were performed with the PyMol (http://www.pymol.org) program. The in-house TASS package, which comprises a collection of Perl scripts, was used to automate aspects of large-scale analysis of sequences, structures, and genome context.

Iterative sequence profile searches and profile-profile comparisons in delineating the eukaryotic homologs of the ArdC-N domain.

While investigating the ArdC-N domain in prokaryotes using recursive sequence profile searches (PSI-BLAST program) against the NCBI-NR database, we surprisingly recovered proteins with this domain from eukaryotes albeit with domain architectures completely unlike those of the prokaryotes. Some of the search examples are furnished here. For example, a search initiated with an ArdC-N domain from *Salmonella enterica* against the NCBI NR database (WP_023226849.1: residues 1 to 140) recovered a significant relationship with Tc-38-like ssDNA-binding protein from the deep-branching kinetoplastid *Perkinsela* sp. (KNH05906.1 (e-value: 3e-06); KNH07778.1 (3e-04); KNH08381.1 (3e-04) in iteration 2 with max target sequences set to 20000) and those from crown-group kinetoplastids such as *Trypanosoma vivax* (CCC49616.1). A reciprocal search using Tc-38-like from *Perkinsela* (KNH07778.1) easily recovered

the TC-38 family proteins from other kinetoplastids and several bacterial ArdC-N homologs with e-values reaching 1e-05 in PSI-BLAST iteration 2. This affirmed the presence of an ArdC-N domain in Tc-38.

Similarly, while analyzing the domain using the profile-profile comparisons as implemented in the HHpred program, we surprisingly detected a significant relationship between the ArdC-N domain and the Pfam profile "BHD_2", one of three domains labeled BHD hitherto exclusively observed in the C-terminal DNA-binding region of the nucleotide excision repair proteins XPC/Rad4 (p-value: 2.2E-08, probability: 96.7%; PDB-ID: 2QSH (Min and Pavletich, 2007), p-value: 8E-07, probability: 93.8%). Reverse profile-profile searches of the sequences corresponding to the "BHD_2" model from Pfam recovered not just the eukaryotic XPC/Rad4 proteins but also bacterial exemplars of the ArdC-N domain. For instance, an iterative sequence profile-based search initiated with a BHD_2 sequence from the fungus *Candida albicans* against the NCBI NR database (XP_712990.1; residues 446 to 507) retrieved the ArdC-N domain in *Fictibacillus phosphorivorans* (WP_066238349.1, e-value: 1e-06, iteration: 5), *Bacillus oceanisediminis* (WP_019379886.1, e-value: 2e-04, iteration: 5), *Bacillus* sp. (WP_009336457.1, e-value: 3e-04, iteration: 5), among others.

We also ran further structure-based homology searches with the DALI program to study the relationships across the three tandemly duplicated ArdC-N domains (BHD domains) of the XPC-RAD4. For example, a pairwise search initiated with the BHD_1 domain recovers the region corresponding to the BHD_2 domain with scores strongly indicative of a relationship (Z-score: 4.6). The reverse search initiated with BHD_2 likewise recovers BHD_2 (Z-score: 7.2) and BHD_1 (z--score: 3.6). Thus, the above observations expand the range of established ArdC-N domains beyond the classical plasmid-interacting prokaryotic ArdC-N family to include the Tc-38-like and BHD families from eukaryotes.

Supplemental references.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research *25*, 3389-3402.

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic acids research *36*, W197-201.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome informatics International Conference on Genome Informatics *23*, 205-211.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A.*, et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic acids research *44*, D279-285.

Lassmann, T., Frings, O., and Sonnhammer, E.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic acids research *37*, 858-865.

Min, J.H., and Pavletich, N.P. (2007). Recognition of DNA damage by the Rad4 nucleotide excision repair protein. Nature *449*, 570-575.

Neuwald, A.F., and Altschul, S.F. (2016). Bayesian Top-Down Protein Sequence Alignment with Inferred Position-Specific Gap Penalties. PLoS computational biology *12*, e1004936.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution *32*, 268-274.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one *5*, e9490.
Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z.*, et al.* (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic acids research *45*, D271-D281.