Evolutionary Insights into the Length Variation of DNA Damage Response Proteins Across Eukaryotes

Dominic Wiredu-Boakye [1,*], Laurence Higgins [1], Ondřej Gahura [2], Anzhelika Butenko [2,3,4], Guy Leonard [5], Mark A. Freeman [6], Árni Kristmundsson [7], Karen Moore [1], Jamie W. Harrison [1], Shani Mac Donald [1], Vyacheslav Yurchenko [3], Bryony A. P. Williams [1], Richard Chahwan [8,*]

[1]Faculty of Health and Life Sciences, Clinical and Biomedical Sciences, University of Exeter, Exeter, Devon, UK
[2]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budejovice, Czechia
[3]Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czechia
[4]Faculty of Sciences, University of South Bohemia, České Budějovice 370 05, Czechia
[5]Department of Biology, University of Oxford, Oxford, UK
[6]Conservation Medicine and Ecosystem Health, Biomedical Sciences, Ross University School of Veterinary Medicine, Basseterre, St. Kitts
[7]Fish Disease Laboratory, Institute for Experimental Pathology, University of Iceland, Reykjavík, Iceland
[8]Cancer Immunobiology Laboratory, Institute of Experimental Immunology, University of Zurich, Zurich, Switzerland

*Corresponding author: E-mails: chahwan@immunology.uzh.ch; d.wiredu-boakye@exeter.ac.uk.

Accepted: April 16, 2025

Abstract

Across the tree of life, DNA damage response (DDR) proteins play a pivotal, yet dichotomous role in organismal development and evolution. Here, we present a comprehensive analysis of 432 DDR proteins encoded by 68 genomes, including that of *Nucleospora cyclopteri*, an intranuclear microsporidia sequenced in this study. We compared the DDR proteins encoded by these genomes to those of humans to uncover the DNA repair-ome across phylogenetically distant eukaryotes. We also performed further analyses to understand if organismal complexity and lifestyle play a role in the evolution of DDR protein length and conserved domain architecture. We observed that the genomes of extreme parasites such as *Paramicrocytos*, *Giardia*, *Spironucleus,* and certain microsporidian lineages encode the smallest eukaryotic repertoire of DDR proteins and that pathways involved in modulation of nucleotide pools and nucleotide excision repair are the most preserved DDR pathways in the eukaryotic genomes analysed here. We found that DDR and DNA repair proteins are consistently longer than housekeeping and metabolic proteins. This is likely due to the higher number of physical protein–protein interactions which DDR proteins are involved. We find that although DNA repair proteins are generally longer than housekeeping proteins, their functional domains occupy a relatively smaller footprint. Notably, this pattern holds true across diverse organisms and shows no dependence on either lifestyle or mitochondrial status. Finally, we observed that unicellular organisms harbour proteins that are tenfold longer than their human homologues, with the extra amino acids forming interdomain regions with a clearly novel albeit undetermined function.

Key words: DNA lesions, genome compaction, DNA damage signalling, protein length, intracellular parasites.

Introduction

Eukaryotic cells endure thousands of DNA lesions each day (Jackson and Bartek 2009). Recognizing, signalling, and repairing DNA damage is therefore essential for cellular maintenance and for ensuring accurate transfer of DNA to daughter cells. The DNA damage response (DDR) is vital for the proper development and disease prevention of multicellular organisms. For example, in adaptive immunity, antibody diversification requires control of several DDR proteins (Sheppard et al. 2018; Cervantes-Gracia et al. 2021). In some unicellular parasites, such as *Trypanosoma brucei*, DDR proteins mediate host immune system evasion by

Significance

The recent explosion of genomic data for non-model organisms provides an avenue to answer several questions about the evolutionary trajectory and dispensability of DNA damage response (DDR) proteins. Protein sequence analyses, including those of *Nucleospora cyclopteri* sequenced herein, show that DDR and DNA replication proteins are, on average, two times longer than their housekeeping counterparts, regardless of organismal lifestyle or mitochondrial status. Furthermore, our data support the hypothesis that at least in some parasitic lineages, protein length compaction happened prior to the emergence of parasitic lifestyles. Finally, we show that in the analysed proteomes, the "modulation of nucleotide pools' and Nucleotide Excision Repair pathway are the most preserved DDR pathways, with DPOLA and ERCC3 being the most preserved DDR proteins.

changing their protective variant surface glycoprotein (VSG) coat. This process occurs close to telomeres and is facilitated by the presence of double-strand breaks (DSBs) in DNA (Horn and McCulloch 2010). Crucially, loss of specific DDR proteins such as those involved in DNA mismatch repair (MMR) can lead to hypermutations. The genomic variation that arises from these hypermutations in turn allows some unicellular organisms to evolve and exploit new habitats and rapidly adapt to new environmental stressors (Steenwyk et al. 2019; Phillips et al. 2021). Hypermutations could lead to genomic expansion or contraction accompanied by long or short protein-coding genes, respectively. The evolutionary trajectory towards genomic expansion or reduction is, however, dependent on fitness. Previous studies have shown that the evolutionary trajectory in the genomes of organisms with hypermutations is often biased towards reduction (Eisen and Hanawalt 1999; Steenwyk et al. 2019). DDR pathways are, therefore, critical for both multicellular and unicellular eukaryotes. However, these pathways are only well-studied in humans and a handful of model organisms, while various understudied eukaryotic lineages, for which genomic data exists, remain underexplored.

Protein-coding genes, encompassing those involved in DDR pathways, evolve through many mechanisms, such as shrinkages, expansions, base mutations, duplications, and fusions (Levinson and Gutman 1987; Björklund et al. 2006; Giacomelli et al. 2007; McDonald et al. 2011). It has been demonstrated that in most genomes, the genetic sequences of functional domains of these proteins are constrained with most of the nucleotide variation occurring in the interdomain regions (Kurland et al. 2007; Wang et al. 2011; Light et al. 2013), although exceptions have been documented in the genomes of obligate intracellular parasites (Nakjang et al. 2013).

With the rise of next-generation sequencing and the availability of raw data repositories, genomes of numerous eukaryotic organisms, especially those that are difficult to culture or those residing in various environmental niches, are now publicly accessible. It is captivating to discern how DDR pathway evolution aligns with the emergence of diverse lifestyles among different eukaryotic lineages. One such lifestyle is intranuclear parasitism. Even with their fascinating life strategy and the potential to exploit their host's DDR protein repertoire, the sequenced genomes of only a handful of intranuclear eukaryotes, namely, *Paramicrosporidium saccamoebae*, *Giardia* spp., and *Enterospora canceri* (Adam 2000 ; Corsaro et al. 2014; Wiredu Boakye et al. 2017), are currently publicly available. To this end, we have now sequenced the genome of *Nucleospora cyclopteri*, an intranuclear microsporidia that infects lumpfish (Freeman et al. 2013). We have also used phylogenomic and bioinformatic analyses to investigate the impact intranuclear parasitism, cytoplasmic parasitism, extracellular parasitism, and free-living lifestyles have on DDR protein length evolution. We analysed the homologues of 432 DDR proteins, 35 DNA replication proteins, 42 metabolic proteins, and 21 housekeeping proteins in the genomes of 67 eukaryotes. We used protein functional domain analyses to further investigate the exact sites of shrinkage or expansion in the homologue structures. We detailed which DDR proteins and pathways are preserved and to what degree they differ in protein sequence length and structure. With the inclusion of several genomes of intracellular parasites, known to have undergone genome compaction due to unique evolutionary pressures, these analyses shed light on the minimal DNA repair-ome. In addition, homologues of human DDR proteins that have undergone protein length compaction or expansion were revealed.

This study serves as a valuable resource for those keen on delving deeper into DDR pathways across a spectrum of eukaryotes, especially in the context of non-model organisms, and to understand minimal protein requirements and configurations essential for DDR pathways. In organisms such as the trypanosomatids and humans, DDR factors are often targets for drug development (Genois et al. 2014 ; Vieira-da-Rocha et al. 2019). A better understanding of their evolutionary trajectory could provide better drug targeting strategies and a better understanding of the fundamental biological mechanisms of DDR signalling.

Results

DPOLA and ERCC3 Are the Most Preserved DDR Proteins in Eukaryotes

Although the proteomes used in this study were parsed and pathways classified for orthologues of known human DDR proteins, we found that 5 out of the 432 human DDR proteins (~1%) did not have identifiable orthologues in model organisms, such as *Xenopus* and *Danio* spp. Our DDR pathway classification was performed according to previous work (Pearl et al. 2015). The DDR protein found to be encoded in all the proteomes used in our analysis, including that of *E. coli* (prokaryotic control proteome), was DPOLA (Arrow in Fig. 1a). The preservation of DPOLA is perhaps not surprising, as it is critically important for DNA synthesis (Starokadomskyy et al. 2016). ERCC3 was the only other protein, for which orthologues were identified in all eukaryotic organisms analysed in this study (Arrow in Fig. 1b).

Modulation of Nucleotide Pools and NER are the Most Preserved DDR Pathways in Eukaryotes

None of the analysed eukaryotic proteomes contained orthologues of all 432 human DDR proteins. However, some DDR pathways had more constituent proteins in the analysed proteomes than others. For example, on average, 4 out of 6 proteins in the nucleotide pool modulation pathway had orthologues in the eukaryotic proteomes used in this study. In other words, on average, 65% of the proteins in the human nucleotide pool modulation pathway had orthologues encoded in the eukaryotic genomes analysed. We refer to this value as the preservation value for the pathway; supplementary table S2, Supplementary Material online lists the preservation values for each of the DDR pathways investigated in this study and shows that modulation of nucleotide pools and NER pathways are the most preserved in the eukaryotic genomes investigated. In general, there is a significant difference in the mean pathway preservation across the four lifestyles investigated (free-living, extracellular, cytoplasmic, and intranuclear [one-way ANOVA, F (3, 64) = 13.08, $P < 0.0001$]). More specifically, pathways were generally more preserved in free-living organisms than in symbionts (Fig. 2, left). Among the organisms analysed, *Paramicrocytos*, *Giardia*, *Spironucleus*, *Enterospora*, and *Nucleospora* harboured the smallest eukaryotic DDR protein repertoires. Specifically, their proteomes contained only 66, 71, 71, 85, and 102 of the 432 human DDR proteins investigated, respectively. Our results indicate that DNA replication and housekeeping pathways are more preserved than metabolic and DDR pathways across all investigated lifestyles (Fig. 2, right). Moreover, lifestyle significantly influences pathway preservation, as evidenced by a notable interaction effect between lifestyle and the four pathway groups (two-way ANOVA, F(9, 256) = 2.293, $P = 0.0172$).

Assessing the Impact of ROS-producing Organelles on Protein Length

We compared the protein lengths between organisms with plastid and ATP-producing mitochondria and no ATP-producing mitochondria to understand if the presence of reactive-oxygen-species-producing organelles, such as ATP-producing mitochondria and plastids, has an impact on protein size evolution (Fig. 3). The classification of organisms with ATP-producing mitochondria was based on Müller et al's classification (Classes 1–4 = organisms with ATP-producing mitochondria, Class 5 = organisms without ATP-producing mitochondria) (Müller et al. 2012). Our results show there was no significant effect of plastid/mitochondrial status on protein length. Furthermore, our data show that DDR and DNA replication proteins are longer compared to their housekeeping and metabolic counterparts in both organisms with plastid/ATP-producing mitochondria and no ATP-producing mitochondria. Moreover, the impact of plastid/mitochondrial status on protein length does not differ between DDR, DNA replication, metabolic, and housekeeping pathways (two-way ANOVA, F (3, 264) = 1.937, $P = 0.1239$) (Fig. 3).

Protein Length Expansion Lies in Regions of Unknown Function

An unexpected observation from this study was that some proteomes contained proteins that were considerably longer than their human homologues. This included proteomes of unicellular organisms, such as *Endotrypanum*, *Neospora,* and *Plasmodium*. Typical examples of these unusually long proteins in unicellular organisms are MLH1, RUVB1, and ATR (MOLV88_240014700, XP_012897480.1 and KAH7831821.1). Due to the widespread misannotation of protein-coding genes in published genomes, often caused by the reliance of annotation programmes on gene models from related organisms that also contain errors, it is likely that the lengths of some of the extremely long proteins identified in our analyses have been affected by this issue. To address this, we systematically cross-referenced the predicted protein lengths with publicly available transcriptomic databases (e.g. EupathDB), focusing on the largest protein in the most conserved orthologue set in each pathway. Of the 20 proteins investigated, only one showed definite evidence of misannotation (UNG in *Blastocystis:* XP_012896248.1; see notes in supplementary table S1, Supplementary Material online). It is, however, unlikely that most of the proteins listed above represent errors in gene model predictions as there is transcriptomic evidence that covers the entire length of their predicted coding sequence in the EupathDB database or the gene model predictions were performed with mRNA

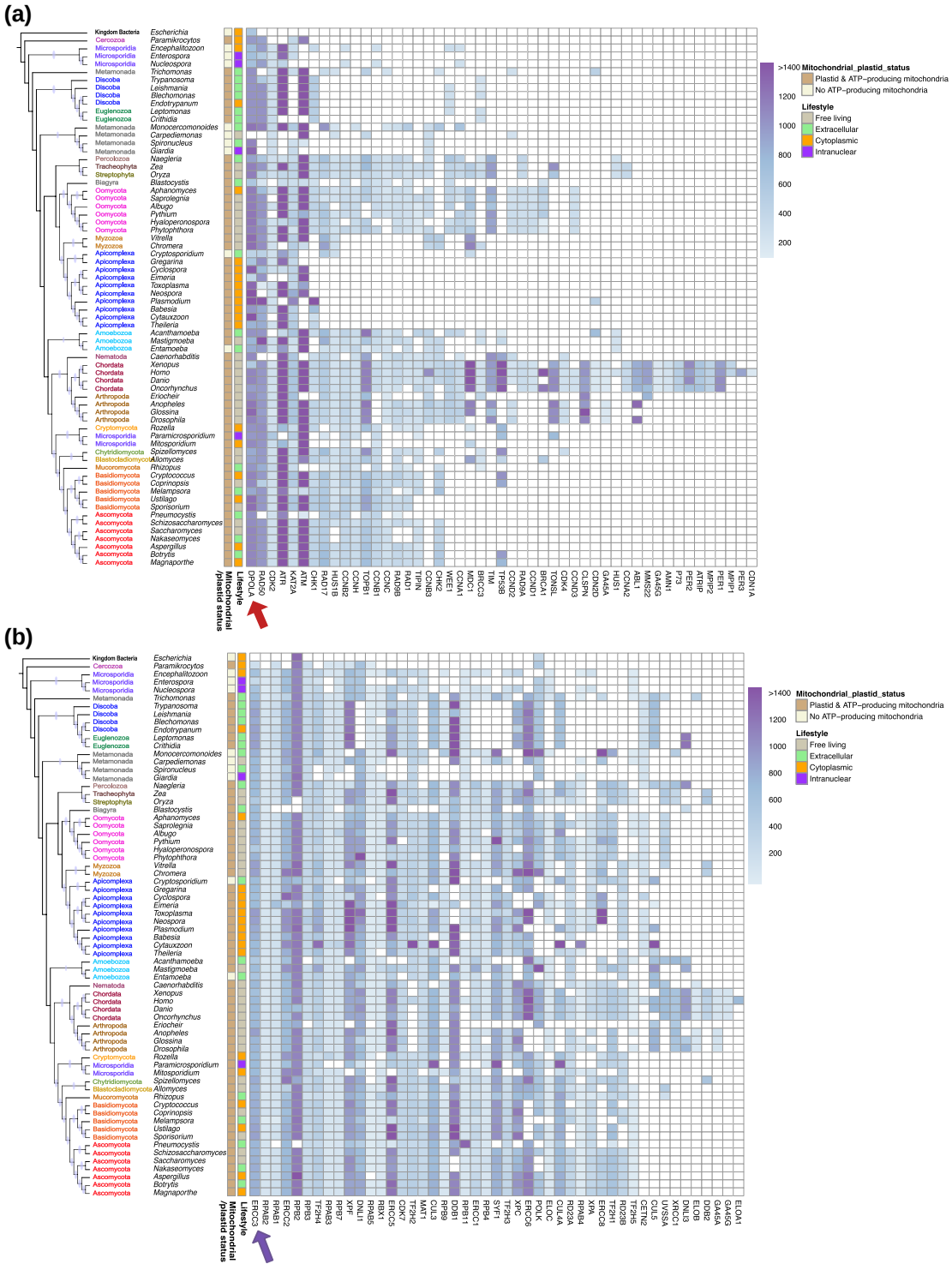

Fig. 1. Profiling of DNA damage repair pathways. The heatmap shows the presence and length of constituent orthologues. Proteins that were not found to be encoded by the analysed genomes are represented by blank spaces. The phylogenetic positions in the cladogram are derived from maximum likelihood analyses performed on a concatenated alignment of 52 proteins. Oval symbols on the phylogenetic tree represent bootstrap support values for the corresponding nodes greater than 70. Proteins were grouped on the *x*-axis according to the DDR pathway they are part of a) *Checkpoint Factors*: Arrow pointing to DPOLA which is the most preserved DDR protein across the 68 organisms analysed in this study, including bacteria. b) *Nucleotide Excision Repair Pathway (NER)*: Arrow pointing to ERCC3 which is the most preserved DDR protein amongst the eukaryotes analysed here. Other pathways analysed in this study can be found in supplementary fig. S1–S21, Supplementary Material Online.

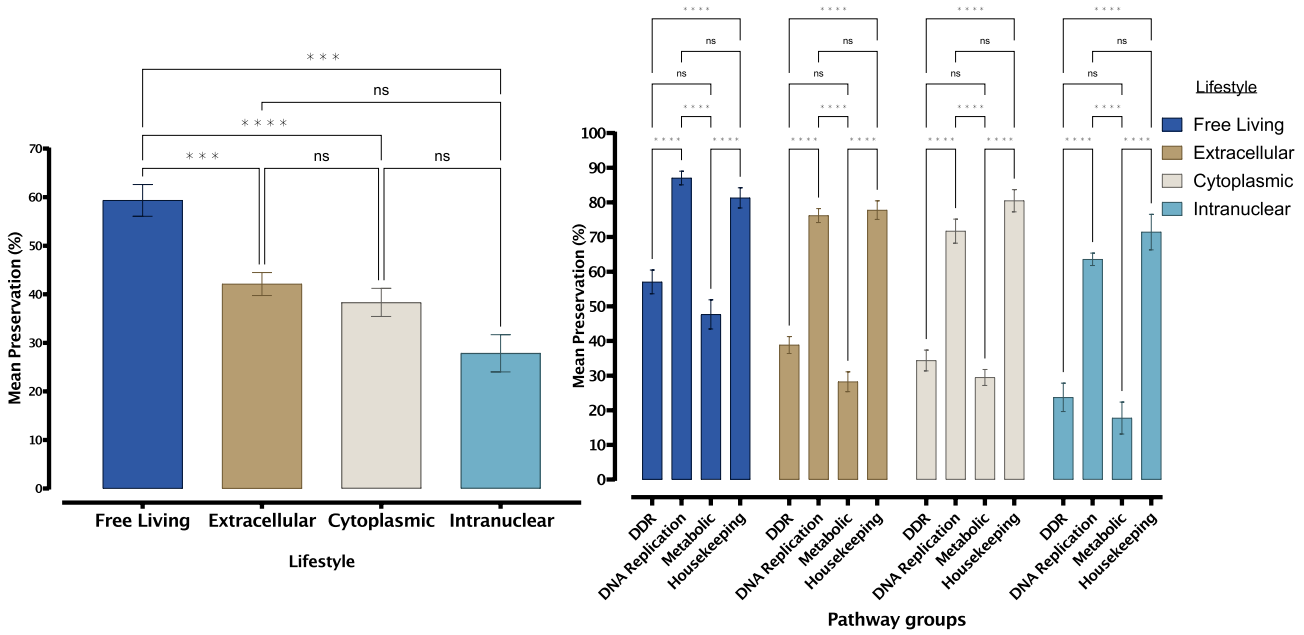

Fig. 2. Mean preservation of DDR pathways: left) Analyses of the four lifestyles investigated show that there is a significant difference in overall mean preservation between lifestyles (one-way ANOVA, $F(3, 64) = 13.08$, $P < 0.0001$) with a gradual decline in preservation as organisms move from a free-living lifestyle to an intranuclear lifestyle. right) Mapping the preservation of DDR, DNA replication, metabolic, and housekeeping pathways against different lifestyles show that DNA replication and housekeeping pathways are more preserved across all lifestyles. Error bars represent the standard error of the mean.

sequencing data (see notes in supplementary table S1, Supplementary Material online).

We performed a Spearman's correlation test between protein length, interdomain length, and functional domain length to determine the hotspots for the observed protein length expansion. There was a very strong positive correlation between protein length and interdomain length ($r(66) = 0.940$, $P < 0.0001$). Similarly, there was a moderate correlation between protein length and functional domain length ($r(66) = 0.464$, $P < 0.0001$) and between Interdomain length and functional domain length ($r(66) = 0.449$, $P < 0.0001$). These results suggest that the increase in protein length in the pathways analysed is primarily due to the expansion of interdomain regions (supplementary fig. S24, Supplementary Material online). A one-way ANOVA test was performed to compare the effect of the four pathway groups on functional domain and interdomain length. This showed that there was a significant difference in mean functional domain length between all four pathway groups, with DNA replication proteins harbouring the longest functional domain regions ($P < 0.0001$) (supplementary fig. S25, Supplementary Material online). Analyses on interdomain length revealed that similarly, there was a significant difference in the length of interdomain regions between all four pathway groups analysed with DDR proteins and housekeeping proteins harbouring the longest and shortest interdomain regions, respectively ($P < 0.0001$) (supplementary fig. S25, Supplementary Material online). On average, DDR and DNA replication proteins are approximately 84% longer than housekeeping proteins and 24% longer than metabolic proteins.

Analysis of DDR, DNA Replication, Metabolic and Housekeeping Protein Lengths Across Eukaryotes With Different Lifestyles

Our analyses revealed that DDR and DNA replication proteins are longer than metabolic and housekeeping proteins, regardless of the organism's lifestyle (Fig. 4). The effect that these four pathway groups have on the protein's length does not differ between lifestyles (two-way ANOVA, $F(9,256) = 0.9415$, $P = 0.4896$).

Genome Statistics of Nucleospora Cyclopteri

The nuclear genome of *Nucleospora cyclopteri* assembled in this study is 4.76 Mb in size spanning across 1,209 scaffolds. It contains 2,939 predicted genes. The GC content is 26.3%, which is within the range of previously sequenced Microsporidia belonging to the family Enterocytozoonidae (22.44—40.15%; Wiredu Boakye et al. 2017 [Table 1]).

Comparison of DDR Proteins With Extreme Lengths Using Structure Prediction

To get a visual representation of how interdomain protein extensions and/or deletions affect tertiary protein structures, we compared experimentally determined- or

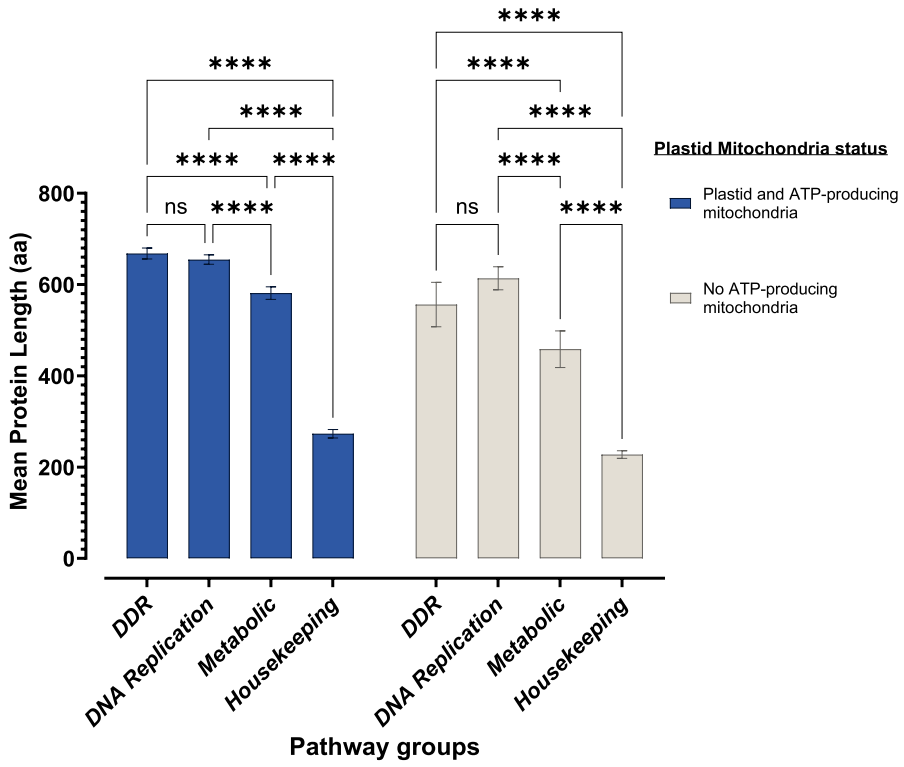

Fig. 3. Comparing the length of DDR proteins in organisms with different plastid and mitochondrial status. The way that functional pathway groups (DDR, DNA replication, metabolic and housekeeping) influenced protein length did not depend on plastid/mitochondrial status ($P = 0.1239$). Functional pathway groups, however, significantly influenced protein length ($P < 0.0001$), with DDR and DNA replication proteins being consistently longer than their housekeeping counterparts across all plastid/mitochondria status groups.

predicted structures of the human or Saccharomyces cerevisiae homologue, with the homology-modelled structures of the smallest and the longest homologues. As expected, in most cases, the homology models of the longest proteins correspond only to the part of their human or yeast homologues, whose structure is available or can be modelled from structures of orthologues in other species. However, in some cases, additional 3D folds were indicated in the extended regions. Therefore, we predicted structures of the longest orthologues of 11 selected proteins using AlphaFold 3 (Abramson et al. 2024) and used the models to search multiple structural and domain databases for structural similarity search by FoldSeek (van Kempen et al. 2024). The approach showed that 6 out of 11 proteins tested are predicted to contain additional domains, such as pleckstrin-homology (TYDP1), helicase (PIF1), ATPase (MRE11), or helical repeat (PARP1, TOP1) domains, in their extensions, indicating possibly extended interaction networks or functions. The remaining proteins exhibit mostly multiple insertions without any predicted folds. The predicted structures of the shortest proteins consistently corresponded to the representatives from model species, but typically lacked long stretches of amino acids that encode functional domains in their human or yeast homologues (supplementary table S5, Supplementary Material online).

Discussion

Proteins Involved in DNA Replication Pathways are Preserved Even in Extreme Parasites

DNA replication stood out as the most preserved pathway amongst the four pathway groups in our analyses. Understandably, transcription and DNA replication for both eukaryotes and prokaryotes is indispensable, and so the retention of proteins involved in this pathway across our analysed proteomes should not be surprising (Samson and Bell 2016). However, considering some of the proteomes analysed here belong to extreme parasites such as microsporidians that have undergone extreme gene loss, it was interesting to see that even in these parasites, DNA replication pathway conservation remained high. Amongst DDR pathways, modulation of nucleotide pools and nucleotide excision repair pathways (NER) were the most preserved. This was expected given that proteins involved in modulation of nucleotide pools, such as Ribonucleoside-diphosphate reductase (RIR), RIR1 and

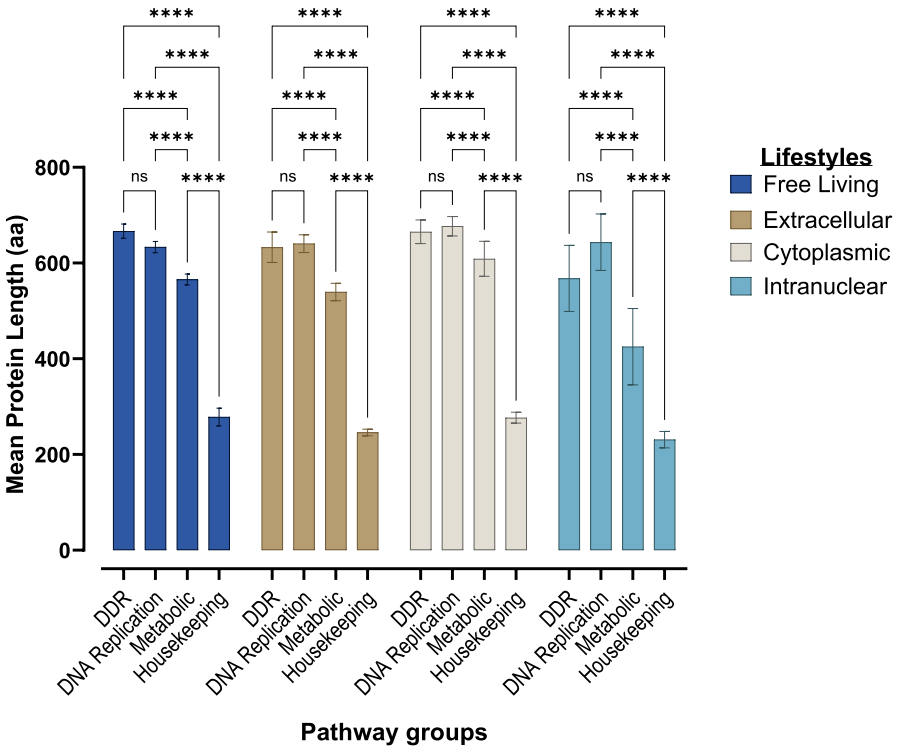

Fig. 4. Assessing the effect of lifestyle on the length of proteins in DDR, DNA replication, metabolic, and housekeeping pathways. Proteins belonging to DDR and DNA replication pathways are longer than those in metabolic and housekeeping pathways, regardless of the organism's lifestyle ($P < 0.0001$). Error bars represent the standard error of the mean.

RIR2, are responsible for de novo dNTP synthesis and for maintaining balanced nucleotide pools (Kunz et al. 1994). Despite the critical nature of DNA synthesis for life, the proteomes of *Spironucleus, Entamoeba*, and *Giardia* did not appear to contain any ortholog of the human RIR proteins (Also known as type I RNR proteins). This observation is consistent with the literature suggesting that their absence is an adaptation to extreme parasitism. *Giardia*, for instance, might have developed specialised nucleotide transporters to syphon nucleotides from its host (Baum et al. 1989; Adam 2001; Loftus et al. 2005; Xu et al. 2014). We also found that the proteome of *Monocercomonoides* did not contain noticeable RIR proteins. Instead, the proteome of this amitochondriate flagellate contained four nucleoside triphosphate proteins, which are believed to substitute the function of RIR proteins in de novo nucleotide synthesis and regulation (Karnkowska et al. 2019). This is peculiar as ribonucleoside triphosphate reductase proteins, also referred to as type III RNRs, are typically found in bacteria and archaea (Fontecave et al. 2002). BLAST search results suggest these proteins may have been horizontally acquired, but further phylogenetic analyses are required to confirm this hypothesis. Similarly, the proteome of *Blastocystis* did not contain any homologues of the RIR proteins. Instead, it contained orthologues of the type III RNRs. Considering type III RNRs are inactivated by oxygen, it is possible that these organisms, together with the other eight organisms investigated in this analyses that contained at least one RIR protein as well as type III RNRs (*Aphanomyces*, *Acanthamoeba*, *Phytophthora*, *Naegleria*, *Allomyces*, *Pythium,* and *Saprolegnia*), may use this protein during their anaerobic life stages. The list of type III RNRs found in the above-mentioned organisms can be found in supplementary table S3, Supplementary Material online. Although *Trichomonas* appeared not to possess type I RNR proteins, there is evidence to support the presence of type II RNR proteins in the *Trichomonas* proteome that has either archaeal or eukaryal origins (Lundin et al. 2010). In line with published data that suggest a reduced repertoire of DDR proteins in *Carpediemonas*, we observed that it lacked any type of RNRs in its proteome (Salas-Leiva et al. 2021). This is particularly intriguing as *Carpediemonas* is a free-living organism and cannot benefit from syphoning host nucleotides as some closely related parasites in the Metamonada lineage do (Karnkowska et al. 2019). It is therefore still unclear how or where *Carpediemonas* gets its nucleotides from for DNA replication and repair.

Nucleotide excision repair (NER) is one of the principal DNA repair pathways, responsible for removing a wide range of DNA lesions typically induced by exogenous

Table 1 Comparing the genome assembly statistics for *Nucleospora cyclopteri* with other members of the Enterocytozoonidae family

	N. cyclopteri	*H. eriocheir*	*H. eriocheir canceri*	*E. canceri*	*E. hepatopenaei*
Assembly Size (Mb)	4.76	4.57	4.84	3.10	3.26
GC %	26.3	22.44	23.16	40.15	25.45
Number of contigs	1209	1300	2344	537	64
N_{50}	6279	17,583	3349	11,128	125,008
# genes	2681	2716	3058	2179	2540
Coverage (X)	920	4477.89	63.18	288	363

chemical agents or UV irradiation. NER is subdivided into two subpathways: global genome repair (GG-NER), which operates throughout the genome, and transcription-coupled repair (TC-NER), which specifically removes lesions from the transcribed strand of actively transcribed genes. Xeroderma pigmentosum group C-complementing protein (XPC) is a core protein required for GG-NER, whereas ERCC6 and ERCC8 are essential for TC-NER (Balajee and Bohr 2000). Given NER's critical role in repairing UV-induced damage, the patchy preservation of its core proteins across the proteomes analysed was initially surprising. However, this observation aligns with previous studies. For instance, Sekelsky and colleagues have previously reported the absence of core TC-NER proteins in *Drosophila* and other arthropods (Sekelsky et al. 2000). Recent studies using excision repair sequencing and genome-wide repair mapping have highlighted the versatility of XPC, demonstrating that XPC could function in both GG- and TC-NER in Drosophila. This suggests that, in closely related arthropods lacking canonical TC-NER proteins, XPC may compensate for the absence of canonical TC-NER proteins (Deger et al. 2022). The absence of core NER proteins in members of the Basidiomycota fungal clade has also been previously documented. In these fungi, alternative excision repair pathways and photoreactivation are believed to substitute canonical NER pathway function (Wong et al. 2019). Notably, *Giardia*, *Spironucleus*, *Enterospora*, and *Paramikrocytos* were the only organisms in our dataset lacking proteins from both canonical NER subpathways. This finding is corroborated by previous studies for *Paramikrocytos*, *Giardia*, and *Spironucleus* (Adam 2000; Feltrin et al. 2020; Onuț-Brännström et al. 2023). Although one previous study predicted the presence of NER core proteins in *Giardia* and *Spironucleus* (Karnkowska et al. 2019), our BLAST searches using NER proteins such as ERCC6 from the closely related metamonad *Monocercomonoides* resulted in low query coverage and low sequence identity (<30%), casting doubt over the existence of functional NER proteins in these organisms. It remains unclear if, and how, parasites such as *Giardia* and *Spironucleus*, which have an environmental life cycle stage, repair DNA damage induced by UV irradiation. Finally, the complete absence of all core NER proteins in the microsporidian *Enterospora* is particularly intriguing, given that other closely related microsporidians, such as *Nucleospora*, encode at least one of these proteins. While the possibility of incomplete genome coverage or annotation errors cannot be ruled out, the apparent lack of most NER proteins in *Enterospora* strongly suggests a genuine loss of this pathway, which warrants further investigation.

Homologous recombination (HR) and nonhomologous end joining (NHEJ) are fundamental biological processes essential for the repair of double-strand DNA breaks. RAD51 and its out-paralogues (RAD51B, RAD51C, RAD51D, XRCC2, and XRCC3) are key regulators of HR in eukaryotes; their absence is particularly lethal in higher eukaryotes but not in other eukaryotic lineages (Bleuyard et al. 2005; Markmann-Mulisch et al. 2007). Nonetheless, our analyses show that nearly all of the eukaryotic proteomes examined contain RAD51. In those few lineages lacking RAD51, such as *Phytophthora*, *Melamspora*, *Aspergillus*, Metamonda, and Euglenozoa (see supplementary fig. S15, Supplementary Material online), we identified either a homologue of one of its out-paralogues or DMC1, a protein with considerable sequence similarity and chemical properties to RAD51 (Lan et al. 2020; Steinfeld et al. 2019). This finding suggests that the HR pathway is likely conserved across these diverse organisms. We find that core NHEJ proteins such as DNLI4, XRCC5, and XRCC6 (KU70/80) are absent in lineages in our analyses with parasitic lifestyles (supplementary fig. S12, Supplementary Material online). Considering the level of functional overlap between HR and NHEJ, it is plausible that genome-streamlining pressures associated with parasitic lifestyles led to the loss of the NHEJ machinery. A more comprehensive exploration of the NHEJ evolution in parasites can be found in Nenarokova et al. (2019).

The evolutionary conservation of certain proteins categorized within DDR pathways, such as POLA1, ERCC3, TOP2B, among others, likely reflects their fundamental roles in cellular processes such as DNA replication and transcriptional regulation, rather than their auxiliary roles in DDR. We recognize that these proteins might be highly conserved for reasons beyond their involvement in DDR pathways. However, we chose to adhere to the genome-wide classification established by Pearl et al. (2015) for consistency and comparability across past studies. We do, however, acknowledge that these categorisations may not perfectly capture the primary biological functions of such proteins. Future efforts to refine DDR protein classifications

could improve the specificity of pathway assignments and facilitate a more nuanced understanding of their evolutionary conservation.

Absence of an Intrinsic Energy Source May Have Led to the Evolution of Smaller Protein Homologues in Eukaryotes Without ATP-Producing Mitochondria

Poorly preserved pathways and shorter proteins are well-known adaptations of parasitism (Williams et al. 2002; Beznoussenko et al. 2007; Corradi et al. 2010; Keeling et al. 2010). Our analyses showed that not all parasites had poorly preserved pathways and short proteins, suggesting that parasitism is not always accompanied by a lack of pathway preservation and gene compaction culminating in shorter proteins; corroborating conclusions from Butenko et al. (2020) and Salas-Leiva et al (2021). Exploring the influence of ROS-producing organelles on protein length across distantly related parasites is challenging, as this effect can be masked by the distinct evolutionary pressures acting on each species. However, the microsporidians' parasitic lifestyle, coupled with the presence of both species that possess ATP-producing mitochondria and those that lack them in this lineage, presents a unique opportunity to investigate the relationship between parasitism and the influence of ROS-producing organelles on protein length. *Paramicrosporidia*, a microsporidian with an ATP-producing mitochondria, encode much longer proteins as compared to other members of the phylum Microsporidia analysed here that do not possess ATP-producing mitochondria (*Encephalitozoon, Enterospora*, and *Nucleospora*) (supplementary fig. S22, Supplementary Material online). The absence of intrinsic ATP generation through oxidative phosphorylation is likely to have posed a bioenergetic constraint that ultimately fostered the evolution of smaller genes and, hence, shorter proteins in microsporidians, as smaller genes are less energetically expensive to maintain, express, and replicate. Conversely, it is generally accepted that the ATP-rich environment provided by a functional mitochondrion fosters the evolution of larger genes that, in turn, encode longer proteins (Lane and Martin 2010; Lane 2011). However, intracellular abundance of ATP alone is unlikely to be the driving force behind the evolution of longer proteins (Lynch and Marinov 2015). Longer proteins must confer an adaptive advantage, especially when occurring in extremely compacted genomes such as those of extreme parasites. Here, we speculate that the presence of ROS-producing organelles, such as ATP-producing mitochondria, increases the selective pressure for the evolution of longer proteins. Considering the deleterious effects of ROS on protein stability (Zuo et al. 2015), it is possible that the amino acids in regions of low complexity of the elongated proteins are used to shield functional domains from the harmful effects of ROS. It has been demonstrated that protein sequence elongation at the C- or N-terminal can confer extra stability to the protein (Matsuura et al. 1999; Liu et al. 2001). Interestingly, several of the extensions observed in the very long homologues analysed here were in the C- or N-terminal of the core protein (supplementary table S5, Supplementary Material online). Furthermore, low complexity regions, such as those featured in the very long proteins in this study, have long been known to have mechanistic functions. Some of these include interaction with phospholipids (Robison et al. 2016), coordinating metal ions (Zhu and Karlin 1996), and having adhesive properties (Haritos et al. 2010; So et al. 2016). A future experiment to test this hypothesis would be to expose homologous proteins with extreme lengths (i.e. a very long homologue and a very short homologue) to oxidative agents and measure the rate of protein denaturation. As genomes of obligate parasites with functional mitochondria that are closely related to parasites without functional mitochondria become publicly available, this hypothesis can be further explored.

Length Variation Between DDR, DNA Replication, Metabolic and Housekeeping Proteins Suggest They Are Influenced by Different Evolutionary Pressures

Proteins involved in the DNA replication process are intimately involved in several DDR processes. As such, it was not surprising to find that there was no difference in length between proteins in these two pathway groups. Our results, however, showed that DDR and DNA repair proteins were consistently longer than metabolic and housekeeping proteins (Fig. 4). It is not clear from our analyses why DDR and DNA repair proteins would evolve to be larger than their metabolic and housekeeping counterparts. However, previous studies have suggested a positive correlation between protein size and expression profiles (Warringer and Blomberg 2006; Moutinho et al. 2019). Thus, highly expressed proteins, such as metabolic and housekeeping proteins, may experience evolutionary pressure to undergo shrinkage to make translation and post-translational folding more energy-efficient (Kim et al. 2014; Uhlen et al. 2015; von Stechow and Olsen 2017). Evidence from transcriptomic analyses from several organisms corroborates this hypothesis as genes involved in DNA repair pathways are not as heavily transcribed as their metabolic counterparts (Ohtsu et al. 2007; Shin et al. 2018; Callejas-Hernández et al. 2019; Mardanov et al. 2020; Fan et al. 2022).

Furthermore, longer proteins are hypothesized to evolve in response to a heightened need for protein–protein interactions within the cell (Cavalier-Smith 2005). In order to assess this hypothesis in relation to our data, we performed searches for 54 representative DDR and housekeeping proteins on the STRING protein–protein interaction database (Szklarczyk et al. 2021). We found that on average, a

DDR protein was projected to have 15 physical interactions, while a housekeeping protein only had 6 (supplementary fig. S23, Supplementary Material online). This underscores the possible role of expanded interdomain regions in DDR proteins in aiding their increased protein–protein interactions. We further found that while DNA repair proteins are typically longer than metabolic and housekeeping proteins (Fig. 4) ($P < 0.0001$), domains of DNA repair proteins occupy a relatively smaller footprint compared to their metabolic and housekeeping counterparts (where footprint is the length covered by the functional domain relative to the length of the protein). Within that smaller domain footprint, DDR proteins often pack more individual functional domains (Fig. 5). That is, the increased length of DDR proteins for the most part is not due to the inclusion of multiple functional domains or expansion of single functional domains but actually due to the expansion of the interdomain space. The packaging of multiple functional domains within a small footprint further speaks to the involvement of DDR proteins in multiple interactions and the importance of the interdomain regions in aiding this function. This pattern appears to have no dependence on organismal lifestyle or mitochondrial status. In-paralogues can serve as an evolutionary sandbox for the birth of novel functional proteins. During speciation, these duplicated genes may accumulate mutations that expand or reduce the length of the proteins they encode (McClune and Laub 2020). While we observed notable differences in protein length among certain in-paralogues in the proteomes analysed, the longest proteins in the pathways we examined were generally single-copy with the exception of ARI1A in *Anopheles*. However, the difference in length between the paralogues of ARI1A in *Anopheles* was relatively small. As such, it is unlikely that the extreme lengths in protein sizes observed in some of the unicellular organisms investigated here are a result of the genetic plasticity that paralogues have.

Intranuclear Lifestyle Is Not Accompanied by Significant Protein Length Reduction

Our analyses contained the only three published genomes of intranuclear eukaryotes, as well as a genome of a fourth intranuclear microsporidia we have sequenced for the first time as part of this study. We did not observe any correlation between intranuclear lifestyle and protein size. Furthermore, our ortholog clustering did not identify any protein families that were unique to all four intranuclear parasites, and our protein length analyses did not identify any unique patterns for this group of organisms, leaving the question of how these parasites are adapted to their unique lifestyle unanswered.

Materials and Methods

Sampling Wild Lumpfish and Purification of *Nucleospora cyclopteri* Spores

Atlantic lumpfish (*Cyclopterus lumpus*) were sampled from a commercial fishing boat operating from the coast of

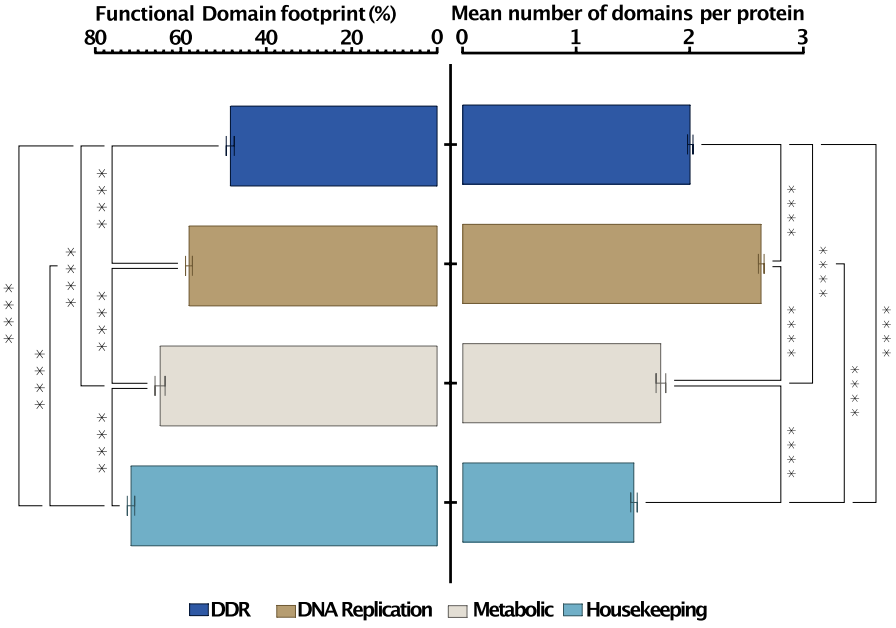

Fig. 5. Assessing the footprint of functional domains between proteins in DDR, DNA replication, metabolic, and housekeeping pathways. DDR and DNA replication proteins have relatively small functional domain footprints but pack a higher density of functional domains within those compact spaces compared to metabolic and housekeeping proteins.

Iceland in December 2016. During necropsy, kidneys with advanced clinical signs of *Nucleospora* infections were crushed with a sterile pestle and mortar in 1 × PBS. The homogenate was then filtered through a 100-μm mesh followed by cell sieving through 40-μm filter to remove tissue debris, and the filtrate was further purified using a Percoll (Sigma) density gradient centrifugation (Taupin et al. 2006).

Genomic DNA Extraction and Sequencing Protocols

Purified spores were subjected to bead beating followed by phenol/chloroform extraction and ethanol precipitation as previously described (Campbell et al. 2013). *Nucleospora cyclopteri* genomic DNA was used to generate a SPriworks fragment library (Beckman Coulter), which was sequenced using the MiSeq v. 2 platform at the University of Exeter Sequencing Service.

Genome Assembly and Annotation

A total of 39,581,310 raw Illumina paired-end reads with a length of 250 bp were used in the analysis. These reads are deposited in the NCBI SRA database under the accession number SUB12014904. PRINSEQ (Schmieder and Edwards 2011) was used to filter and trim reads identified to have poor quality scores by FASTQC (Andrews 2010). This resulted in 32,709,299 paired-end reads with an average length of 139 bp. These reads were used for the assembly of the *N. cyclopteri* genome with Spades v. 3 (Prjibelski et al. 2020). The sparsity of introns in the genomes of most microsporidians make fast prokaryotic genome annotation tools ideal for their annotation, as has been done in previous microsporidian sequencing studies (Wiredu Boakye et al. 2017). As such, open-reading frames (ORFs), tRNAs and rRNAs of the *N. cyclopteri* genome were predicted by Prokka v. 1.11 (Seemann 2014). Further BLASTP and BLASTX searches were used to annotate predicted ORFs. Contaminating bacterial sequences, found in the initially assembled scaffolds, were removed using a combination of both BLASTP and BLASTX searches for putative ORFs and by BLASTN searches for rRNAs against the NCBI non-redundant protein (nr) and nucleotide (nt) databases, respectively. All BLAST searches were performed using default parameters.

Identification of DDR, DNA Replication, Metabolic, and Housekeeping Orthologues

In this study, the predicted proteins of *N. cyclopteri* and those of 67 other organisms, whose proteomes are publicly available (supplementary table S4, Supplementary Material online), were parsed to identify homologues of DDR proteins. To accomplish this, the predicted proteins encoded by the genomes of these organisms were downloaded from NCBI (Agarwala et al. 2018), Ensembl (Zerbino et al. 2018), EupathDB (Aurrecoechea et al. 2017), or GigaDB (Sneddon et al. 2012) (FASTA files with protein sequences used in this study can be found here: https://drive.google.com/file/d/1q25rde-9vHaZBANvGSbq4bbV5PGl3LSg/view?usp=sharing). These protein sequences were grouped into orthologue families by running Orthofinder (V2.5.5) (Emms and Kelly 2019) with default parameters. Orthofinder assigned 995882 genes (90.1%) to 84761 orthogroups, which suggested our species sampling was good. The Orthofinder orthogroup output folder for human proteins was then parsed for 432 DDR, 35 DNA replication, 38 metabolic, and 21 housekeeping orthologue families. Although Orthofinder reliably detects orthologues in closely related species, its phylogeny-based approach can miss orthologues in distantly related lineages. To address this limitation, we supplemented our orthofinder analysis with orthology clustering data from InParanoiDB. Unlike Orthofinder, InParanoiDB identifies orthologues by comparing pairwise similarity scores for Pfam-predicted domains between pairs of reference proteomes (e.g. between human and yeast), thereby circumventing some of the challenges encountered by phylogeny-centric tools. Following the orthofinder analyses, we constructed a phylogenetic tree using the concatenated sequences of a set of shared proteins between the 68 organisms analysed, enabling us to define broad taxonomic clades. For each clade, we designated a reference organism present in InParanoiDB (e.g. Saccharomyces for Fungi). When Orthofinder failed to detect an orthologue for a given species, we first queried the corresponding human orthologue against the clade's reference organism proteome in InParanoiDB. If the human protein had a confirmed orthologue in the reference proteome, we then used that reference orthologue as "bait" to locate the corresponding orthologue of the query species in the Orthofinder pairing file between the reference organism and the query species. A flow chart explaining our orthologue clustering methodology can be found in supplementary fig. S26, Supplementary Material online. The DDR proteins used in this analysis were those used in Perl et al. (2015). The DNA replication proteins were those listed in the KEGG human DNA replication repository (Kanehisa et al. 2023), whereas the housekeeping proteins included those in Joshi et al. (2022) and Barta et al. (2023). Subsequently, the protein lengths for the predicted orthologues were extracted. The script used for the orthology clustering and extracting protein lengths can be found here: https://github.com/DrDomUoE/DDR_paper.git. To evaluate the performance of our orthologue-calling approach, we focused on orthologue groups that showed >95% conservation across the species analysed. Because these groups are so widely preserved, we reasoned that any reported absences are likely to be due either to methodological errors or rare instances of true loss. For any orthologues found to be missing within these groups, we manually investigated them through BLAST and literature

searches to determine whether they could be detected in closely related species not included in our analyses. This enabled us to distinguish genuine biological absences from those potentially attributable to failure of our orthologue-calling, incomplete genome sequencing, or poor gene calling models. We identified 34 orthologue groups that were at least 95% conserved across all species examined. We next quantified the performance of our orthologue-calling method by counting the number of absences that we predicted to be due to methodological failure. Our estimates indicate that, for each orthologue group, our ortholog calling method successfully identified orthologues in 65 to 67 of the 67 species analysed (i.e. 95.5% to 100% coverage). Based on these figures, the probability of observing the recorded number of missing orthologues purely as a result of orthologue-calling failure (rather than genuine biological absences) was calculated to be between 0% and 17%. During these analyses, it became apparent that the *Eriocheir* proteome used here was severely incomplete due to partial coverage of the published genome assembly. Consequently, protein absences in this genome were excluded from our analysis.

Phylogenetic Tree Construction

For the above-mentioned phylogenetic tree, Fifty-two single-copy othologues from the Orthofinder output were used to build a concatenated alignment for the construction of a species tree. More specifically, the proteins were initially aligned using MUSCLE v. 5.1 (Edgar 2004) with default parameters, trimmed with TrimAl v1.5.rev0 (Capella-Gutierrez et al. 2009) with default parameters. To identify the most appropriate protein model for each alignment, we used iqtree (v2.3.6) with the -m MF option (Minh et al. 2020). We used the concatenate option in Mega to concatenate the alignments from the ortholog families into a single file. We then used the predicted top protein models assigned by iqtree and the alignment length for each orthologue family to create a tabular formatted partition file as prescribed by the RaxML manual. The final concatenated alignment was passed to RaxML(8.2.12) (Stamatakis 2014) to construct a species tree with the following parameters: PTHREADS -T 20 -q, -m PROTGAMMAAUTO, -f a, -# 100, -x 12345, -*P* 54321.

Functional Domain Prediction

Functional domain calling was accomplished by using pfam_scan.pl (Mistry et al. 2007) to query all proteins against a locally installed version of the Pfam-A.hmm database, which was downloaded on the 23 August 2024 onto a local server. This was run with default parameters (Eddy 2011). The number and length of the domains predicted for each protein were extracted from the pfam_scan.pl output file. The bash script we created for parsing the pfam_scan.pl output can be found online: https://github.com/DrDomUoE/DDR_paper.git.

Prediction of Protein Structures

Structures of selected DDR proteins were homology-modelled using SWISS-MODEL (Waterhouse et al. 2018) and Phyre2 (Kelley et al. 2015) online web toolkits. De novo structure prediction was performed using AlphaFold 3 (Abramson et al. 2024). FoldSeek (van Kempen et al. 2024) was used to search for structurally similar proteins and domains in all available databases, including AlphaFold, PDB, and CATH. The predicted structures were superimposed with experimentally determined structures of respective human or *S. cerevisiae* homologues using the MatchMaker tool of UCSF Chimera (Pettersen et al. 2004).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgements

This work was supported by: UK Research and Innovation-Biotechnology and Biological Sciences Research Council [BB/N017773/2] to [RC]; Swiss National Science Foundation [310030_212553; 320030E_215576, CRSK-3_190550; IZSEZ0_204655; IZSEZ0_218166] to [RC]; Novartis Foundation [22B140] to [RC]; Vontobel [41309] to [RC]; Foundation for Research in Science and the Humanities at the University of Zurich [F-41309-01-01] to [RC]; and University of Zurich -University Research Priority Programs [Translational Cancer Research] to [RC]; European Union's Operational Program "Just Transition" (CZ.10.03.01/00/22_003/0000003 LERCO) to [VY]; Czech Science Foundation (23-04769S) to [VY]; University of Exeter Sequencing Service; Wellcome Trust Institutional Strategic Support Fund [WT097835MF]; Wellcome Trust Multi User Equipment Award [WT101650MA]; Medical Research Council Clinical Infrastructure Funding [MR/M008924/1]; BBSRC LOLA award [BB/K003240/1]

Conflict of Interest

The authors have no conflicts of interest to declare.

Data Availability

The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JASPEO000000000. The version described in this article is version JASPEO010000000. In house scripts and datafiles used in this study

can be found online: https://github.com/DrDomUoE/DDR_paper.git

Literature Cited

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024:630(8016):493–993.

Adam RD. The *Giardia lamblia* genome. Int J for Parasitol. 2000:30(4):475–484. https://doi.org/10.1016/S0020-7519(99)00191-5.

Adam RD. Biology of *Giardia lamblia*. Clin Microbiol Rev. 2001:14(3):447–475. https://doi.org/10.1128/CMR.14.3.447-475.2001.

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2018:46(D1):D8–D13. https://doi.org/10.1093/nar/gkx1095.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 2018 September 15

Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, Crouch K, Doherty R, Falke D, Fischer S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic acids Res. 2017:45(D1):D581–D591. https://doi.org/10.1093/nar/gkw1105.

Balajee AS, Bohr VA. Genomic heterogeneity of nucleotide excision repair. Gene. 2000(1–2):15–30. https://doi.org/10.1016/s0378-1119(00)00172-4.

Barta N, Ördög N, Pantazi V, Berzsenyi I, Borsos BN, Majoros H, Páhi ZG, Ujfaludi Z, Pankotai T. Identifying suitable reference gene candidates for quantification of DNA damage-induced cellular responses in human U2OS cell culture system. Biomol. 2023:13:1523. https://doi.org/10.3390/biom13101523.

Baum KF, Berens RL, Marr JJ, Harrington JA, Spector T. Purine deoxynucleoside salvage in *Giardia lamblia*. J Biol Chem. 1989:264(35):21087–21090. https://doi.org/10.1016/S0021-9258(19)30049-3.

Beznoussenko GV, Dolgikh VV, Seliverstova EV, Semenov PB, Tokarev YS, Trucco A, Micaroni M, Di Giandomenico D, Auinger P, Senderskiy IV, et al. Analogs of the Golgi complex in microsporidia: structure and avesicular mechanisms of function. J Cell Sci. 2007:120(7):1288–1298. https://doi.org/10.1242/jcs.03402.

Björklund ÅK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Comput Biol. 2006:2(8):e114. https://doi.org/10.1371/journal.pcbi.0020114.

Bleuyard J-Y, Gallego ME, Savigny F, White CI. Differing requirements for the *Arabidopsis* Rad51 paralogs in meiosis and DNA repair. Plant J. 2005:41(4):533–545. https://doi.org/10.1111/j.1365-313X.2004.02318.x.

Butenko A, Opperdoes FR, Flegontova O, Horák A, Hampl V, Keeling P, Gawryluk RMR, Tikhonenkov D, Flegontov P, Lukeš J. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biol. 2020:18(1). https://doi.org/10.1186/s12915-020-0754-1.

Callejas-Hernández F, Gutierrez-Nogues Á, Rastrojo A, Gironès N, Fresno M. Analysis of mRNA processing at whole transcriptome level, transcriptomic profile and genome sequence refinement of *Trypanosoma cruzi*. Sci Rep. 2019:9(1):1–11. https://doi.org/10.1038/s41598-019-53924-6.

Campbell SE, Williams TA, Yousuf A, Soanes DM, Paszkiewicz KH, Williams BAP. The genome of *Spraguea lophii* and the basis of host-microsporidian interactions. PLoS Genet. 2013:9(8):e1003676. https://doi.org/10.1371/journal.pgen.1003676.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009:25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 2005:95(1):147–175. https://doi.org/10.1093/aob/mci010.

Cervantes-Gracia K, Gramalla-Schmitz A, Weischedel J, Chahwan R. APOBECs orchestrate genomic and epigenomic editing across health and disease. Trends Genet. 2021:37(11):1028–1043. https://doi.org/10.1016/j.tig.2021.07.003.

Corradi N, Pombert J-F, Farinelli L, Didier ES, Keeling PJ. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. Nat Commun. 2010:1(1):77. https://doi.org/10.1038/ncomms1082.

Corsaro D, Walochnik J, Venditti D, Steinmann J, Müller K-D, Michel R. Microsporidia-like parasites of amoebae belong to the early fungal lineage Rozellomycota. Parasitol Res. 2014:113(5):1909–1918. https://doi.org/10.1007/s00436-014-3838-4.

Deger N, Cao X, Selby CP, Gulec S, Kawara H, Dewey EB, Wang L, Yang Y, Archibald S, Selcuk B, et al. CSB-independent, XPC-dependent transcription-coupled repair in *Drosophila*. Proc Natl Acad Sci USA. 2022:119(9–NaN).

Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011:7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004:32(5):1792–1797. https://doi.org/10.1093/nar/gkh340.

Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res. 1999:435(3):171–213. https://doi.org/10.1016/S0921-8777(99)00050-6.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Fan Y, Wang J, Yu K, Zhang W, Cai Z, Sun M, Hu Y, Zhao X, Xiong C, Niu Q, et al. Comparative transcriptome investigation of *Nosema ceranae* infecting eastern honey bee workers. Insects. 2022:13(3):241. https://doi.org/10.3390/insects13030241.

Feltrin RDS, Segatto ALA, de Souza TA, Schuch AP. Open gaps in the evolution of the eukaryotic nucleotide excision repair. DNA Rep. 2020:95:102955. https://doi.org/10.1016/j.dnarep.2020.102955.

Fontecave M, Mulliez E, Logan DT. Deoxyribonucleotide synthesis in anaerobic microorganisms: the class III ribonucleotide reductase. Prog Nucleic Acid Res Mol Biol. 2002:72:95–222.

Freeman MA, Kasper JM, Kristmundsson Á. Nucleospora cyclopteri n. sp., an intranuclear microsporidian infecting wild lumpfish, Cyclopterus lumpus L., in Icelandic waters. Parasi Vectors. 2013:6(1). https://doi.org/10.1186/1756-3305-6-49.

Genois M-M, Paquet ER, Laffitte M-CN, Maity R, Rodrigue A, Ouellette M, Masson J-Y. DNA repair pathways in trypanosomatids: from DNA repair to drug resistance. Microbiol Mol Biol Rev. 2014:78(1):40–73. https://doi.org/10.1128/MMBR.00045-13.

Giacomelli MG, Hancock AS, Masel J. The conversion of 3′ UTRs into coding regions. Mol Biol Evol. 2007:24(2):457–464. https://doi.org/10.1093/molbev/msl172.

Haritos VS, Niranjane A, Weisman S, Trueman HE, Sriskantha A, Sutherland TD. Harnessing disorder: onychophorans use highly unstructured proteins, not silks, for prey capture. Proc R Soc Lond B Biol Sci. 2010:27:3255–3263. https://doi.org/10.1098/rspb.2010.0604.

Horn D, McCulloch R. Molecular mechanisms underlying the control of antigenic variation in African trypanosomes. Curr Opin Microbiol. 2010:13(6):700–705. https://doi.org/10.1016/j.mib.2010.08.009.

Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009:461(7267):1071–1078. https://doi.org/10.1038/nature08467.

Joshi CJ, Ke W, Drangowska-Way A, O'Rourke EJ, Lewis NE. What are housekeeping genes? PLoS Comput Biol. 2022:18(7):e1010295. https://doi.org/10.1371/journal.pcbi.1010295.
Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023:51(D1):D587–D592. https://doi.org/10.1093/nar/gkac963.
Karnkowska A, Treitli SC, Brzoň O, Novák L, Vacek V, Soukal P, Barlow LD, Herman EK, Pipaliya SV, Pánek T, et al. The oxymonad genome displays canonical eukaryotic complexity in the absence of a mitochondrion. Mol Biol Evol. 2019:36(10):2292–4604.
Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S. The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. Genome Biol Evol. 2010:2:304–309. https://doi.org/10.1093/gbe/evq022.
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015:10(6):845–858. https://doi.org/10.1038/nprot.2015.053.
Kim HJ, Na JI, Min BW, Na JY, Lee KH, Lee JH, Lee YJ, Kim HS, Park JT. Evaluation of protein expression in housekeeping genes across multiple tissues in rats. Korean J Pathol. 2014:48(3):193–200. https://doi.org/10.4132/KoreanJPathol.2014.48.3.193.
Kunz BA, Kohalmi SE, Kunkel TA, Mathews CK, McIntosh EM, Reidy JA. Deoxyribonucleoside triphosphate levels: a critical factor in the maintenance of genetic stability. Mutat Res-Genet Toxicol. 1994:318(1):1–64. https://doi.org/10.1016/0165-1110(94)90006-X.
Kurland CG, Canbäck B, Berg OG. The origins of modern proteomes. Biochimie. 2007:89(12):1454–1463. https://doi.org/10.1016/j.biochi.2007.09.004.
Lan WH, Lin S-Y, Kao C-Y, Chang W-H, Yeh H-Y, Chang H-Y, Chi P, Li H-W. Rad51 facilitates filament assembly of meiosis-specific Dmc1 recombinase. Proc Natl Acad Sci USA. 2020:117(21):11257–22521.
Lane N. Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct. 2011:6(1):35. https://doi.org/10.1186/1745-6150-6-35.
Lane N, Martin W. The energetics of genome complexity. Nature. 2010:467(7318):929–934. https://doi.org/10.1038/nature09486.
Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 1987:4(3):203–221. https://doi.org/10.1093/oxfordjournals.molbev.a040442.
Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol. 2013:30(12):2645–2653. https://doi.org/10.1093/molbev/mst157.
Liu JH, Tsai CF, Liu JW, Cheng KJ, Cheng CL. The catalytic domain of a *Piromyces rhizinflata* cellulase expressed in *Escherichia coli* was stabilized by the linker peptide of the enzyme. Enzyme Microb Technol. 2001:28(7-8):582–589. https://doi.org/10.1016/S0141-0229(00)00349-5.
Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, et al. The genome of the protist parasite *Entamoeba histolytica*. Nature. 2005:433(7028):865–868. https://doi.org/10.1038/nature03291.
Lundin D, Gribaldo S, Torrents E, Sjöberg B-M, Poole AM. Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. BMC Evol Biol. 2010:10:383–NaN.
Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015:112(51):15690–15695. https://doi.org/10.1073/pnas.1514974112.
Mardanov AV, Eldarov MA, Beletsky AV, Tanashchuk TN, Kishkovskaya SA, Ravin NV. Transcriptome profile of yeast strain used for biological wine aging revealed dynamic changes of gene expression in course of flor development. Front Microbiol. 2020:11:538. https://doi.org/10.3389/fmicb.2020.00538.
Markmann-Mulisch U, Wendeler E, Zobell O, Schween G, Steinbiss H-H, Reiss B. Differential requirements for RAD51 in *Physcomitrella patens* and *Arabidopsis thaliana* development and DNA damage repair. Plant Cell. 2007:19(10):3080–3089. https://doi.org/10.1105/tpc.107.054049.
Matsuura T, Miyai K, Trakulnaleamsai S, Yomo T, Shima Y, Miki S, Yamamoto K, Urabe I. Evolutionary molecular engineering by random elongation mutagenesis. Nat Biotechnol. 1999:17(1):58–61. https://doi.org/10.1038/5232.
McClune CJ, Laub MT. Constraints on the expansion of paralogous protein families. Curr Biol. 2020:30(10):R460–R464. https://doi.org/10.1016/j.cub.2020.02.075.
McDonald MJ, Wang W-C, Huang H-D, Leu J-Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biol. 2011:9(6):e1000622. https://doi.org/10.1371/journal.pbio.1000622.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020:37(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.
Mistry J, Bateman A, Finn RD. Predicting active site residue annotations in the Pfam database. BMC Bioinform. 2007:8(1). https://doi.org/10.1186/1471-2105-8-298.
Moutinho AF, Trancoso FF, Dutheil JY. The impact of protein architecture on adaptive evolution. Mol Biol Evol. 2019:36(9):2013–2028. https://doi.org/10.1093/molbev/msz134.
Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R-Y, van der Giezen M, Tielens AGM, Martin WF. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev: MMBR. 2012:76(2):444–539.
Nakjang S, Williams TA, Heinz E, Watson AK, Foster PG, Sendra KM, Heaps SE, Hirt RP, Martin Embley T. Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics. Genome Biol Evol. 2013:5(12):2285–2303. https://doi.org/10.1093/gbe/evt184.
Nenarokova A, Záhonová K, Krasilnikova M, Gahura O, McCulloch R, Zíková A, Yurchenko V, Lukeš J, Heitman J, Matlashewski G. Causes and effects of loss of classical nonhomologous end joining pathway in parasitic eukaryotes. mBio. 2019:10(4). https://doi.org/10.1128/mBio.01541-19.
Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, Chen T, Zhang X, Timmermans MCP, Beck J, Buckner B, et al. Global gene expression analysis of the shoot apical meristem of maize (*Zea mays* L.). Plant J. 2007:52(3):391–404. https://doi.org/10.1111/j.1365-313X.2007.03244.x.
Onuț-Brännström I, Stairs CW, Campos KIA, Thorén MH, Ettema TJG, Keeling PJ, Bass D, Burki F, Eme L. A mitosome with distinct metabolism in the uncultured protist parasite *Paramikrocytos canceri* (Rhizaria, Ascetosporea). Genome Biol Evol. 2023:15(3). https://doi.org/10.1093/gbe/evad022.
Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within the DNA damage response. Nat Rev Cancer. 2015:15(3):166–180. https://doi.org/10.1038/nrc3891.
Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF chimera—a visualization system for exploratory research and analysis. J Comput Chem. 2004:25(13):1605–1612. https://doi.org/10.1002/jcc.20084.
Phillips MA, Steenwyk JL, Shen XX, Rokas A. Examination of gene loss in the DNA mismatch repair pathway and its mutational consequences in a fungal phylum. Genome Biol Evol. 2021:13(10):10. https://doi.org/10.1093/gbe/evab219.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinformatics. 2020:70(1):e102. https://doi.org/10.1002/cpbi.102.

Robison AD, Sun S, Poyton MF, Johnson GA, Pellois J-P, Jungwirth P, Vazdar M, Cremer PS. Polyarginine interacts more strongly and cooperatively than polylysine with phospholipid bilayers. J Phys Chem B. 2016:120(35):9287–9296. https://doi.org/10.1021/acs.jpcb.6b05604.

Salas-Leiva DE, Tromer EC, Curtis BA, Jerlström-Hultqvist J, Kolisko M, Yi Z, Salas-Leiva JS, Gallot-Lavallée L, Williams SK, Kops GJPL, et al. Genomic analysis finds no evidence of canonical eukaryotic DNA processing complexes in a free-living protist. Nat Commun. 2021:12(1):6003. https://doi.org/10.1038/s41467-021-26077-2.

Samson RY, Bell SD. Archaeal DNA replication origins and recruitment of the MCM replicative helicase. Enzymes. 2016:39:169–190. https://doi.org/10.1016/bs.enz.2016.03.002.

Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011:27(6):863–864. https://doi.org/10.1093/bioinformatics/btr026.

Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014:30(14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Sekelsky JJ, Hollis KJ, Eimerl AI, Burtis KC, Hawley RS. Nucleotide excision repair endonuclease genes in Drosophila melanogaster. Mutat Res. 2000:459(3):219–247.

Sheppard EC, Morrish RB, Dillon MJ, Leyland R, Chahwan R. Epigenomic modifications mediating antibody maturation. Front Immunol. 2018:9:355. https://doi.org/10.3389/fimmu.2018.00355.

Shin J, French L, Xu T, Leonard G, Perron M, Pike GB, Richer L, Veillette S, Pausova Z, Paus T. Cell-specific gene-expression profiles and cortical thickness in the human brain. Cereb Cortex. 2018:28(9):3267–3277. https://doi.org/10.1093/cercor/bhx197.

Sneddon TP, Li P, Edmunds SC. GigaDB: announcing the GigaScience database. GigaScience. 2012:1(1):11. https://doi.org/10.1186/2047-217X-1-11.

So CR, Fears KP, Leary DH, Scancella JM, Wang Z, Liu JL, Orihuela B, Rittschof D, Spillmann CM, Wahl KJ. Sequence basis of barnacle cement nanostructure is defined by proteins with silk homology. Sci Rep. 2016:6(1):36219. https://doi.org/10.1038/srep36219.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014:30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Starokadomskyy P, Gemelli T, Rios JJ, Xing C, Wang RC, Li H, Pokatayev V, Dozmorov I, Khan S, Miyata N, et al. DNA polymerase-α regulates the activation of type I interferons through cytosolic RNA: DNA synthesis. Nat Immunol. 2016:17:495–504.

Steenwyk JL, Opulente DA, Kominek J, Shen XX, Zhou X, Labella AL, Bradley NP, Eichman BF, Čadež N, Libkind D, et al. Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol. 2019:17(5):e3000255. https://doi.org/10.1371/journal.pbio.3000255.

Steinfeld JB, Beláň O, Kwon Y, Terakawa T, Al-Zain A, Smith MJ, Crickard JB, Qi Z, Zhao W, Rothstein R, et al. Defining the influence of Rad51 and Dmc1 lineage-specific amino acids on genetic recombination. Genes Dev. 2019:33(17–18):1191–2398.

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021:49(D1):D605. https://doi.org/10.1093/nar/gkaa1074.

Taupin V, Méténier G, Vivarès CP, Prensier G. An improved procedure for Percoll gradient separation of sporogonial stages in *Encephalitozoon cuniculi* (Microsporidia). Parasitol Res. 2006:99(6):708–714. https://doi.org/10.1007/s00436-006-0231-y.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. Tissue-based map of the human proteome. Science. 2015:347(6220). https://doi.org/10.1126/science.1260419.

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024:42(2):243–489.

Vieira-da-Rocha JP, Passos-Silva DG, Mendes IC, Rocha EA, Gomes DA, Machado CR, McCulloch R. The DNA damage response is developmentally regulated in the African trypanosome. DNA Repair (Amst). 2019:73:78–90. https://doi.org/10.1016/j.dnarep.2018.11.005.

von Stechow L, Olsen JV. Proteomics insights into DNA damage response and translating this knowledge to clinical strategies. Proteomics. 2017:17(3-4):3–4. https://doi.org/10.1002/pmic.201600018.

Wang M, Kurland CG, Caetano-Anollés G. Reductive evolution of proteomes and protein structures. Proc Natl Acad Sci U S A. 2011:108(29):11954–11958. https://doi.org/10.1073/pnas.1017361108.

Warringer J, Blomberg A. Evolutionary constraints on yeast protein size. BMC Evol Biol. 2006:6(1):61. https://doi.org/10.1186/1471-2148-6-61.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, De Beer TAP, Rempfer C, Bordoli L, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018:46(W1):W296–W303. https://doi.org/10.1093/nar/gky427.

Williams BAP, Hirt RP, Lucocq JM, Embley TM. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. Nature. 2002:418(6900):865–869. https://doi.org/10.1038/nature00949.

Wiredu Boakye D, Jaroenlak P, Prachumwat A, Williams TA, Bateman KS, Itsathitphaisarn O, Sritunyalucksana K, Paszkiewicz KH, Moore KA, Stentiford GD, et al. Decay of the glycolytic pathway and adaptation to intranuclear parasitism within Enterocytozoonidae microsporidia. Environ Microbiol. 2017:19(5):2077–2089. https://doi.org/10.1111/1462-2920.13734.

Wong HJ, Mohamad-Fauzi N, Rizman-Idid M, Convey P, Alias SA. Protective mechanisms and responses of micro-fungi towards ultraviolet-induced cellular damage. Polar Sci. 2019:20:19–34. https://doi.org/10.1016/j.polar.2018.10.001.

Xu F, Jerlström-Hultqvist J, Einarsson E, Ástvaldsson Á, Svärd SG, Andersson JO. The genome of *Spironucleus salmonicida* highlights a fish pathogen adapted to fluctuating environments. PLoS Genet. 2014:10(2):e1004053. https://doi.org/10.1371/journal.pgen.1004053.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. Ensembl 2018. Nucleic Acids Res. 2018:46(D1):D754–D761. https://doi.org/10.1093/nar/gkx1098.

Zhu ZY, Karlin S. Clusters of charged residues in protein three-dimensional structures. Proc Natl Acad Sci U S A. 1996:93(16):8350–8355. https://doi.org/10.1073/pnas.93.16.8350.

Zuo L, Zhou T, Pannell BK, Ziegler AC, Best TM. Biological and physiological role of reactive oxygen species—the good, the bad and the ugly. Acta Physiol. 2015:214(3):329–348. https://doi.org/10.1111/apha.12515.

Associate editor: Courtney Stairs