The evolutionary landscape of the DNA Damage Response network: a computational approach

Aida Arcas Mantas

TESI DOCTORAL UPF / ANY 2013

DIRECTORA DE LA TESI

Dra. Ana María Rojas Mendoza

Computational Cell Biology Group
IMPPC – Institut de Medicina Predictiva i Personalitzada del Càncer

UNIVERSITAT POMPEU FABRA

IMPPC
Institut de Medicina Predictiva
i Personalitzada del Càncer

ACKNOWLEDGEMENTS

In this section I would like to express my sincere thanks to all the people who have helped me in any way since I started doing research in Bioinformatics, without their contribution and support it would not have been possible to write this PhD dissertation.

I am enormously grateful to my supervisor Dr. Ana M. Rojas for giving me the opportunity to pursue my PhD at the IMPPC (when I thought all chances for doing a PhD had passed), for supporting and guiding my work during the last three years, and for helping me in every way she could.
I would also like to acknowledge the members of my thesis committee Eduardo Eyras, Cedric Notredame and Marcus Buschbeck for critically reviewing my work and providing me with insightful remarks and suggestions.

In chronological order, I would like to thank Dr. Juan Antonio G. Ranea for giving me the chance to do my first research project in Bioinformatics and for his useful advise in computational biology research; and to Professor Christine Orengo, for all her support and help during my eight months stay in her lab at the London University College, in which I became interested in evolutionary biology and molecular evolution. I would like to thank as well my lab colleagues Sarah and Adrian for always lending a hand when was needed and for the good moments we shared in London.

Next, I must acknowledge Dr. Manuel José Gómez Rodríguez, without whom I would have never had the opportunity to work in the Centre for Astrobiolgy (CAB)(INTA-CSIC) in Madrid; I must thank him for teaching me about many bioinformatics tools, programming, extremophilic bacteria and metabolism.
At the CAB I was very fortunate to work in a multidisciplinary environment and to meet many great researchers. Among them I would like to thank Professor Juan Pérez-Mercader, for his support and interest in our work, and specially Dr. Víctor Parro, for his confidence in me and for making possible that I could stay longer in the CAB and finish my research project there.
Other people I am particularly grateful to are Paloma, for her good advice and insights on how a research center works; the IT guys Luis and Fernando, for always being available and willing to help; and of course, to my lab mate and good friend Noemí, one of the most kind persons I have ever met.
I would also like to mention my friends from “la ruta 7” of INTA: Paco (the ‘driving force’), Piluca, Andrés, Alberto and Pilarcita, with whom I have spent many hours on the road and had lots of fun.

At the IMPPC, I would like to acknowledge all the people who have made my stay in Barcelona a bit more bearable, principally Alba and Javi during the first year. Also, special thanks go to my lab colleagues Edu and Ida for their contribution to

this thesis project, and to Alonso, for always having interesting comments and ideas about this work.
I am also particularly grateful to Judith, Lorena and Víctor for creating the nice working environment in the bioinformatics area, for the interesting and never-ending conversations about all kinds of topics, and for the help (especially Víctor with Perl scripts) and support I have always received from them.

On a different level, I would like to express my gratitude to all those anonymous people that have made the easy access to information through the Internet possible, and to those music bands that have made the soundtrack of my life (in particular OMD) and have made the endless hours before the computer screen pass fast.

Also, I must thank Jorge for his technical help.

And last but not least, I would like to thank my parents and my sister for their constant support and for always encouraging me to surpass myself, and to the "pechus" for their help.

ABSTRACT

The DNA Damage response is a crucial signaling network that preserves genome integrity. This network is an ensemble of distinct but often overlapping sub-networks, where participating components exert different functions according to precise spatio-temporal frameworks. To understand how these sub-networks have been assembled and emerged along evolution, we have screened DDR components in 47 selected species covering the tree of life and analyzed their evolutionary and functional properties according to different gene ages and following a variety of classifications.

This is the first time a systematic analysis covers the DDR network's evolution as a whole. Our results indicate that most of the DDR components are ancestral genes, that all the subnetworks contain at least one representative protein traceable to Prokaryota, and that the ancestral core of the DDR machinery is mainly related to repair and is mostly built upon sensor and effector activities. Along evolution the enlargement of the network has occurred through the addition of new components that have evolved to interact and work together with the ancient ones, which may have increased the complexity of the DDR network in terms of fine-tuning and cross-talk to other pathways.

RESUMEN

La respuesta al daño en el ADN (DDR) es una red de señalización esencial que mantiene la integridad genética. Esta red es un conjunto de sub-redes distintas, pero a menudo solapantes, donde los componentes que participan desempeñan diversas funciones según marcos espacio-temporales precisos. Para comprender cómo estas sub-redes han surgido a lo largo de la evolución y cómo se han ido ensamblando, hemos buscado componentes de DDR en 47 especies que cubren el árbol de la vida, y hemos analizado sus propiedades evolutivas y funcionales según distintas edades de genes y siguiendo varias clasificaciones.

Esta es la primera vez que un análisis sistemático cubre la evolución global de la red de DDR. Nuestros resultados indican que la mayoría de los componentes de la DDR son genes antiguos, que todas las sub-redes contienen al menos un representante trazable hasta procariotas, y que el núcleo ancestral de la maquinaria de DDR está principalmente relacionado con reparación y se construyó sobre actividades de detección y efectores. A lo largo de la evolución, la ampliación de la red ha ocurrido a través de la adición de nuevos componentes que han evolucionado para interaccionar y funcionar junto a los antiguos, lo que puede haber incrementado la complejidad de la red de DDR en términos de precisión y de comunicación con otras redes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Cells are continuously at risk of DNA damage from multiple sources, both exogenous (e.g. ionizing radiation or ultraviolet rays, chemicals, etc) and endogenous (e.g. reactive oxygen species and DNA replication errors) [1,2]. To detect DNA lesions, signal their presence and promote repair of the damage, cells have evolved a sophisticated and intricate network of concerted pathways that overall constitutes the DNA damage response (DDR).

Perturbations to the network either by a deficient function of its components or by alterations on its regulation produce genomic instability, which can lead to disease.

Every living organism requires a proper and efficient system for genome protection, and this system has likely evolved to adapt to different life-styles.

Although various research works using different approaches has been done to delineate the structure of the DDR network, our understanding regarding how this network as a whole ever emerged is still far to be complete.

1.1 DNA DAMAGE AND REPAIR MECHANISMS

1.1.1 Brief history of DNA repair research

Long before the structure of DNA was determined and before DNA was identified as the fundamental repository of the genetic information of all known organisms (with the exception of RNA viruses), experiments by Hermann Muller showed in 1927 the mutagenic effects that environmental agents, such as X-rays, have on the genetic material of cells.

Later, in 1935, the first direct experimental evidence for DNA repair was obtained by Alexander Hollaender, who discovered that organisms have the ability to recover from the lethal effects of UV light and proposed the existence of cellular responses that momentarily arrest the growth of exposed cells, therefore enabling the repair of the damage before resuming growth.

It was not until almost a decade later that the term *DNA repair* was incorporated into the lexicon of molecular and cellular biology. [3].

Discoveries of diverse DNA repair pathways during the second half of the 20th century explained many of the early observations in mechanistic terms. Nevertheless, it was not until the 1990s, with the discovery of DNA damage-induced signal transduction pathways linking DNA damage to cell cycle arrest and apoptosis, and with the finding of the role of ataxia telangiectasia mutated (ATM) and ataxia telangiectasia mutated Rad3-related kinase (ATR) genes, that the full meaning of the DNA damage response began to be understood.

In this century, new discoveries such as the deep connection among DNA damage, aging and cancer [4], or the importance of post-translational modifications (PTMs) in the regulation of DDR pathways [5] have greatly expanded our views of how cells and tissues limit mutagenesis and tumorigenesis. Besides, over the past decade enormous progress has been made in the clarification of the workings and interconnections of repairs pathways, and in the spatio–temporal orchestration of DNA repair [6-8].

Some of the key findings in the field of DNA repair and DDR are summarized in Figure 1. Despite all these discoveries, we are still far away from encompassing the whole complexity of the DDR, but surely in the next years new incredible advances will be made in the field.

Nowadays, new elements involved in the DDR are being identified almost every week and recent discoveries are starting to unveil the role of miRNAs in the regulation of DDR genes [9-11]. Moreover, due to the development of new technologies such as genome-wide RNAi screening and next generation sequencing, the post-transcriptional regulation of the DDR by non-coding RNAs (ncRNAs) and RNA-binding proteins (RBPs) is beginning to be understood, adding a new layer of complexity to the DDR landscape [12-14].

Finally, thanks to our increasing knowledge of DNA damage responses and our understanding of their complexity and effects, opportunities for improving disease detection and management are arising.

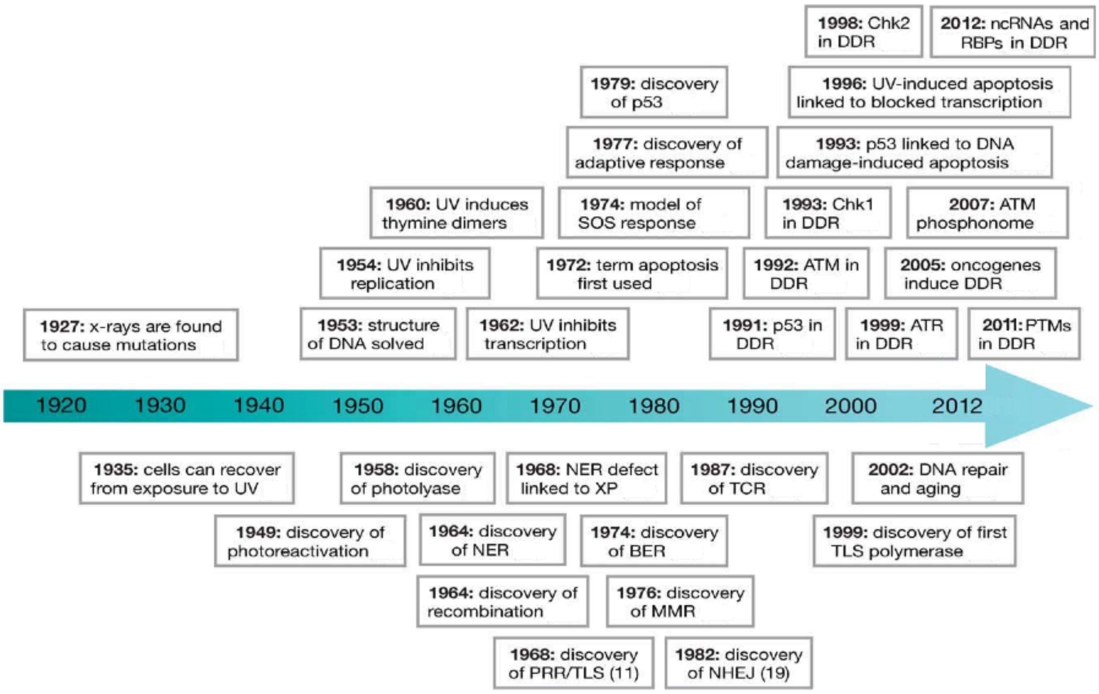

1998: Chk2 in DDR
2012: ncRNAs and RBPs in DDR
1979: discovery of p53
1996: UV-induced apoptosis linked to blocked transcription
1977: discovery of adaptive response
1993: p53 linked to DNA damage-induced apoptosis
1960: UV induces thymine dimers
1974: model of SOS response
1993: Chk1 in DDR
2007: ATM phosphonome
1954: UV inhibits replication
1972: term apoptosis first used
1992: ATM in DDR
2005: oncogenes induce DDR
1927: x-rays are found to cause mutations
1953: structure of DNA solved
1962: UV inhibits transcription
1991: p53 in DDR
1999: ATR in DDR
2011: PTMs in DDR
1920
1930
1940
1950
1960
1970
1980
1990
2000
2012
1935: cells can recover from exposure to UV
1958: discovery of photolyase
1968: NER defect linked to XP
1987: discovery of TCR
2002: DNA repair and aging
1949: discovery of photoreactivation
1964: discovery of NER
1974: discovery of BER
1999: discovery of first TLS polymerase
1964: discovery of recombination
1976: discovery of MMR
1968: discovery of PRR/TLS (11)
1982: discovery of NHEJ (19)

Figure 1. Timeline summarizing some of the key findings in the DNA repair and DDR fields. The discoveries listed on top are related to various effects of DNA damage on cellular functions and DNA damage signaling while the discoveries listed on the bottom are related to DNA repair (figure adapted from Ljungman M., 2010) [15].

1.1.2 Sources and types of DNA damage

Threats to DNA integrity come from multiple endogenous and exogenous sources. Regarding the former, there are three main sources: i) Spontaneous reactions (mainly hydrolysis) inherent to the chemical nature of DNA in an aqueous solution, which generate abasic sites and cause deamination [16,17] ii) Products of our own metabolism, such as reactive oxygen, nitrogen and carbonyl species [18], endogenous alkylating agents, estrogen and cholesterol metabolites, and lipid peroxidation products, all of which damage DNA. iii) Replication errors: replication defects can cause mismatches and replication fork collapse can result in strand breaks [19].

Regarding the exogenous sources, the main DNA damaging agents are chemicals and physical agents such as ionizing radiation or ultraviolet rays.

It has been estimated that reactive oxygen and nitrogen species alone generate more than 70 oxidative base and sugar products in DNA as well as different types of single-strand breaks (SSBs), while spontaneous base losses in nuclear DNA have been estimated to reach 10^4 per cell per day [16,17]. Together with other types of damage, the total number of DNA lesions that each of the ~10^{13} cells in the human body receives per day may be close to 10^5 [16].

Table 1 (below in section 1.1.3) summarizes some of the most common kinds of DNA damage and their sources, and a short description of the main types of DNA damage follows:

- Deamination

Deamination involves the loss of amino groups from DNA bases. All DNA bases but thymine (which does not have an amino group) undergo spontaneous deamination. Most types of deamination reactions produce a base that does not naturally occur in DNA (with the exception of deamination of 5-methylcytosine) and this fact facilitates the identification and excision of the deaminated base by DNA glycosylases **[20].**

The most frequent type of deamination event in cells is deamination of cytosine into uracil. In mammals this happens in about 100-500 bases per cell per day in spontaneous deamination reactions [2].

- Abasic sites, depurination and depyrimidination

An abasic site, also termed “apurinic or apyrimidinic” (AP) site, is formed when a base is lost from the DNA by cleavage of a N-glycosyl bond, leaving the sugar-phosphate chain intact [2]. Abasic sites can be produced by spontaneous depurination and depyrimidination reactions and are potentially mutagenic. Depurination reactions involve the loss of purine bases (adenine and guanine) from DNA. In these reactions, the N-glycosyl bond to deoxyribose is broken by hydrolysis, leaving the DNA’s sugar-phosphate chain intact and producing an abasic site, while depyrimidination imply the loss of pyrimidine bases (cytosine and thymine) from DNA [16]. Abasic sites can also be produced by reactive oxygen species (ROS) [21] and in intermediate steps of the base excision repair [22].

- Pyrimidine dimers

Pyrimidine dimers are mutagenic lesions formed from thymine or cytosine bases in DNA via photochemical reactions. These dimers alter the structure of the double helix and interfere with base pairing during DNA replication, consequently inhibiting polymerases and arresting replication [23].

Covalent linkages between adjacent pyrimidines in the same DNA strand characterize cyclobutane pyrimidine dimers (CPDs), which are the most frequent type of photoproduct produced when DNA is exposed to UV-B [16] or UV-C radiation. Thymine dimers are the type of CPD most frequently found in DNA. The formation of CPDs can also enhance the deamination of cytosine.

Other common UV product are 6,4-photoproducts, or 6,4 pyrimidine-pyrimidones, which occur at one-third the frequency of CPDs but are more mutagenic. [2].

- DNA strand breaks

Ionizing radiation (for example, from cosmic radiation or X-rays) can cause SSBs and double-strand breaks (DSBs) in the DNA double helix. If these breaks are not properly repaired, they can induce mutations and lead to widespread structural rearrangement of the genome.

Some strand breaks are generated in intermediate steps of natural occurring reactions. For example, the process of V(D)J recombination during B- and T-cell development is initiated by a DSB between two recombining variable-region gene segments and their flanking sequences [24]. Nevertheless, other strand breaks are a severe form of DNA damage and inhibit DNA replication, leading to the activation of the DNA repair machinery. This is the case when SSBs occur due to the oxidation of DNA bases by ROS, or when stalled DNA replication forks collapse and free double-stranded ends [25].

1.1.3 DNA repair mechanisms

The core of the cellular defense against DNA injuries is composed by diverse DNA repair mechanisms, each with their own damage specificity (Table 1, adapted from Giglia-Mari *et al.,* 2010). Collectively, they are capable of removing most lesions from the genome.

Table I. Sources of DNA lesions and corresponding repair pathways

Lesion	Cause	Repair pathway/process
CPD, 6-4PP(1)	Sunlight	Photoreactivation, NER
Bulky adducts(2)	Food, cigarette smoke	NER
Intrastrand crosslinks	Chemotherapy (e.g., Cis-Pt)	NER
8-oxo-dG(3)	ROS(4), respiration	BER
Thymineglycol(3)	ROS(4), respiration	BER
N$_7$-Alkyl-dG, N$_3$-Alkyl-dA	Food, pollutants	BER
O$_6$-Alkyl-dG	Food, pollutants	DR(5), BER?
5-methyl-dC	DNMT(6)	BER
Uracil, (Hypo)Xanthine	Spontaneous deamination	BER
Abasic site	Spontaneous hydrolysis	BER/ Trans-lesion bypass
Single-strand breaks	Ionizing radiation, ROS	Ligation, BER
Double-strand breaks	Ionizing radiation, ROS, V(D)J-rec	HR, NHEJ
Tyrosyl-3’ DNA(7)	Topo-I inhibition, ROS	SSBR
Mismatches	Replication errors	MMR
Small insertion/deletions	Replication slippage	MMR
Interstrand crosslinks	Chemotherapy	Fanconi anaemia pathway/ ICLR(8)/ HR?

1. CPD: cyclobutane pyrimidine dimer; 6-4 PP: 6-4 pyrimidine-pyrimidone photo-product.
2. A large group of chemicals conjugated to bases that cause DNA helix destabilization such as: Benzo(α)pyrene (a polycylic aromatic hydrocarbon); Aflatoxins (present in fungal food contaminations); and Nitrosamines (tobacco smoke).
3. A large group of different oxidation products affecting either the base or the phosphate-sugar backbone of which 8-oxo-dG is the most abundant.
4. ROS: reactive oxygen species, produced as side-product of respiration/metabolism and ionizing radiation. These species include superoxide, hydrogen peroxide, hydroxyl radicals and singlet oxygen.
5. DR: direct reversal.
6. DNMT: DNA methyltransferase, functions in epigenetic gene-expression control (e.g., at CpG islands).
7. Proteolytic degradation of conjugated Topo-I to 3’DNA termini creates tyrosyl-3’DNA bonds, resolved by TDP1.
8. ICLR: interstrand crosslink repair.

Depending on the nature of the DNA lesion [26], the cell type, the cell cycle phase in which the lesion is encountered [27], and if the DNA can be repaired after careful checking by the checkpoint pathways [28], different DNA repair systems can be utilized to restore the damaged DNA.

The existence of these pathways enables to avoid or minimize possible alterations of genome structures leading to loss of proliferative control or cell death, and therefore it ensures the accurate transmission of genetic information to the next generation.

DNA repair systems have received much attention [15,29-31], and the main ones are classified as direct reversal, mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), homologous recombination (HR) and non-homologous end-joining (NHEJ). Their specificities and functioning are explained below:

a) Direct reversal

The first DNA repair mechanism that is thought to have emerged in evolution (which was also the first to be discovered) is enzymatic photoreactivation, a process carried by photolyases, which selectively reverse UV-induced DNA damage [32].
Photolyases are not conserved in mammals, which have to rely on a more intricate mechanism to remove UV injuries: nucleotide excision repair [31].
Other simple solution that emerged in nature is the suicide enzyme O-6-methylguanine-DNA methyltransferase, which transfers the methyl group from a guanine in DNA to an internal cysteine residue in the enzyme, thereby inactivating itself [2].

b) Mismatch repair

The MMR system is a highly conserved pathway that elevates the fidelity of DNA replication and plays an important role in maintaining genetic stability [33]. Though MMR proteins are known to be involved in cellular responses and repair synthesis at the sites of various types of DNA damage [34,35], its best understood function is the elimination of mismatches arising during DNA replication, a process that has been extensively studied and best characterized in *E. coli*. In eukaryotes, mismatches produced during replication are recognized by the heterodimers MutSα (MSH2/MSH6), which binds base-base mismatches and small insertion-deletion loops, and MutSβ (MSH2/MSH3), which binds larger insertion-deletion loops. The heterodimer MutLα (MLH1/PMS2) is recruited by the MSH2 protein to form a ternary complex with one of the MutS complexes and promotes the repair process via its endonucleolytic activity, leading to an excision repair of the mismatch [36]. Additional proteins involved in this process may include the 5’–3’ double-stranded DNA exonuclease I (EXO1), helicase(s), replication protein-A (RPA), replication factor C (RFC), proliferating cell nuclear antigen (PCNA), and DNA polymerases α and β [37].
MSH6 has been reported to be involved in the repair of DSBs through a direct physical interaction with the NHEJ protein Ku70, upregulating the activity of the latter. The exact mechanism by which the mismatch repair protein interacts with and regulates Ku70 is still to be elucidated, though it is suggested that MSH6 could increase the binding

affinity of Ku70 for DNA, modulate the amount of Ku70/80 complexes or facilitate the recruitment of other important NHEJ proteins to bind the broken-DNA ends [38].

c) Base excision repair

Lesions caused by endogenous and exogenous reactive species generating small chemical alterations like alkylation, deamination, abasic or AP sites, SSBs and oxidization of DNA bases are corrected by BER through excision of the damaged base, incorporation of the correct nucleotide(s), and strand ligation [30,39].
BER is initiated by DNA glycosylases, which remove the damaged base, leaving an AP site with mutagenic potential. The AP site is then processed by AP endonucleases and the gap is subsequently processed by at least two BER sub-pathways: the short-patch (SP) BER and the long-patch (LP) BER. In SP-BER, the gap is filled inserting a single nucleotide and is performed by the DNA polymerase β, while in LP-BER two to about thirteen nucleotides are incorporated and is carried out by the DNA polymerase β and the replicative DNA polymerases ∂ or/and ε [40]. The pathway requires several DNA replication factors, including PCNA. Finally, the strand ligation is performed by the X-ray cross-complementing group 1 (XRCC1)/DNA ligase III complex in SP-BER, while in LP-BER the ligation step is carried out by DNA ligase I [41].

d) Nucleotide excision repair

The NER system, which recognizes bulky and helix-distorting base lesions, uses two sub-pathways that differ in the mechanism of lesion recognition: global genome repair (GG-NER), which scans the entire genome for distorted DNA and eliminates the damaged bases, and transcription-coupled repair (TC-NER), which specifically removes lesions that block transcription [29].
In TC-NER, damage sensing is performed by the stalled RNA polymerase II, and the Cockayne syndrome factors CSA and CSB play essential roles in TC-NER complex assembly [42]. In GG-NER, the XPC-RAD23B complex detects the lesions, opens the DNA locally and helps recruit transcription factor IIH (TFIIH). In the next steps, the two sub-pathways converge. The combined action of these complexes creates short stretches of single-stranded DNA around the lesion, and then the unwound DNA is stabilized by XPA and RPA. Sequentially, XPB and XPD, the two helicase subunits of TFIIH, bind and extend the single-stranded DNA around the damage site, allowing further NER factors to bind and form a pre-incision complex. The endonucleases XPF-ERCC1 and XPG then cleave on the 5’ and 3’ sides of the lesion, respectively, generating a patch of approximately 30 nucleotides that is subsequently filled in by DNA replication proteins, including RFC, PCNA, RPA, and several DNA polymerases. Finally, the gap is sealed by DNA ligases I or III [4,43,44].

e) Double-strand break repair

Among the different type of lesions, and because of their potential to provoke major chromosomal rearrangements, DSB are particularly hazardous and elicit a very robust cytotoxic response; therefore, efficient repair of DSBs is essential for genome stability and viability.

As explained above, DSBs can be caused by a variety of sources such as ionizing radiation or X-rays, certain chemicals and ROS. In addition, DSBs can be produced during V(D)J recombination and when the replication machinery collapses at replication blocks [29].

In eukaryotes, the main pathways involved in the repair of DSBs are homologous recombination (HR) and non-homologous end-joining (NHEJ).

NHEJ uses limited or no sequence homology to rejoin juxtaposed ends, while HR requires a homologous sister chromatid as a template to properly align and seal the broken DNA ends [29]. This availability of a homologous template makes HR an error-free repair system of the damaged DNA. On the contrary, NHEJ is considered an error-prone repair system since bases are generally deleted or inserted as part of the reparation process of this pathway.

Nevertheless and despite its mutagenic nature, NHEJ is responsible for repairing a major fraction of DSBs in higher eukaryotes, especially in mammals; whereas HR is the preferred pathway for DSB repair in bacteria [45,46] and lower eukaryotes [47,48].

Non-homologous end-joining

In modern eukaryotes, NHEJ consists of at least two distinct sub-pathways: the main classic end-joining pathway (C-NHEJ), and an alternative NHEJ (A-NHEJ) [49]. Little is known about the regulation and the elements involved in A-NHEJ, in contrast, at least seven proteins are known to be required for C-NHEJ in mammalian cells.

Ku70 and Ku80 form the ring-shaped heterodimer Ku, that recognizes and binds to DSBs protecting the DNA ends from resection [50]. DNA-dependent protein kinase catalytic subunit (DNA-PKcs) is then recruited to the DSB and activated by autophosphorylation. Once autophosphorylated, DNA-PKcs bridges the two proximal broken DNA ends [51]. Together, Ku and DNA-PKcs form the DNA-PK holoenzyme which, when end processing is required, binds to and phosphorylates *Artemis*, a downstream factor of the ATM signaling pathway, activating its nuclease function [52]. Besides *Artemis*, other end processing enzymes are recruited to the DBS. Finally, the end ligation step is mediated by a complex consisting of DNA Ligase 4 and X-ray repair cross complementing protein (XRCC4), whose activity is stimulated by the association of the XLF protein (XRCC4-like factor), also called *Cernunnos* or nonhomologous end-joining 1 (NHEJ1) [53].

Homologous recombination

HR is implicated in the repair of damaged replication forks and their re-initiation when stalled and collapsed. Besides, it is involved in centrosome stabilization [54], in the repair of telomeres through the action of the SMC5-SMC6 complex [55], and it is also responsible for accurate segregation of homologous chromosomes in meiosis [56]. Also, during meiosis a fundamental role of HR is to generate crossovers between homologous chromosomes, thereby contributing to genetic variation. Such events arise at the molecular level through either resolution of double Holliday junctions (HJs) by the SLX1-SLX4 complex [57] or through MUS81-EME1 dependent cleavage of HJs [58].

In addition, HR is used to repair interstrand DNA crosslinks, the repair of which involves the Fanconi anemia protein complex [59].

The HR pathway starts when the MRN (Mre11–Rad50–Nbs1) complex recognizes the DSB, binds there, and holds together and stabilizes the DNA ends [60]. The MRN complex also provides scaffolding for the CtBP-interacting protein (CtIP) nuclease, which catalyzes end resection at the break together with the Bloom syndrome protein (BLM) Helicase, DNA2 helicase/nuclease [61] and EXO1 [62,63]. EXO1 is a key mediator of DNA end resection and DSB repair and damage signaling decisions, since resection by this exonuclease facilitates a transition from ATM- to ATR-mediated cell cycle checkpoint signaling [64].

DSB end processing generates an extended region of ssDNA that is then bounded by RPA. Next, Rad51 and other factors are recruited to the DSB. Afterwards, the RPA-coated ssDNA is replaced with Rad51 in a process that involves Rad52, BRCA1 and other proteins implicated in the Fanconi anemia pathway such as BRCA2 and PALB2 (partner and localizer of BRCA2) [59]. The Rad51 nucleofilament promotes the search for the homologous duplex DNA in the undamaged sister chromatid and facilitates strand invasion into the homologous template. Finally, the action of DNA polymerases, nucleases, helicases, resolvase enzymes and other HR factors, mediate DNA ligation and resolution of repair intermediates to produce undamaged DNA molecules [65].

1.1.4 DNA Damage Tolerance

Persisting lesions not fixed by any of the repair mechanism will negatively affect DNA replication. At least two mechanisms to bypass DNA damage have evolved: translesion synthesis (TLS) and template switching. These processes do not eliminate lesions, but serve as a temporary solution to overcome lesion-stalled replication forks, which can lead to highly cytotoxic DSBs and thus require a quick response.

In TLS, upon lesion-induced replication blockage, the usual high-fidelity DNA polymerases (pol∂, ε or α) are transiently replaced with low-fidelity translesion polymerases (polζ-κ) that are able to synthesize DNA using a template strand encompassing a DNA lesion. Once the replication fork passes the site of the lesion, the low-fidelity DNA polymerases are replaced with the usual high-fidelity enzymes, which allows DNA synthesis to continue as normal [66]. Though TLS can circumvent lesion-induced replication stalling the low fidelity of the alternative polymerases usually causes enhanced mutagenesis [31].

In template switching, the DNA lesion is bypassed at the replication fork by leaving a gap in DNA synthesis opposite the lesion. Once the lesion has passed the replication fork, the single-strand gap is repaired utilizing template DNA on a sister chromatid, in a similar way to the process employed during HR [25].

1.1.5 Regulation of DNA repair throughout the cell cycle

Eukaryotic cells are equipped with a wide range of mechanisms to maintain genomic stability during the cell cycle (Figure 2). DNA lesions activate checkpoint pathways that regulate specific DNA-repair mechanisms in the different phases of the cell cycle. Recent studies have provided insights into the mechanisms that contribute to DNA repair in specific cell-cycle phases and have highlighted the machinery that controls cell-cycle progression or arrest [6,67,68]. Besides, it has been proposed that PTMs may tune the efficiency or the specificity of the repair machinery towards a certain type of lesion, often to facilitate repair in a specific cell-cycle phase [6].

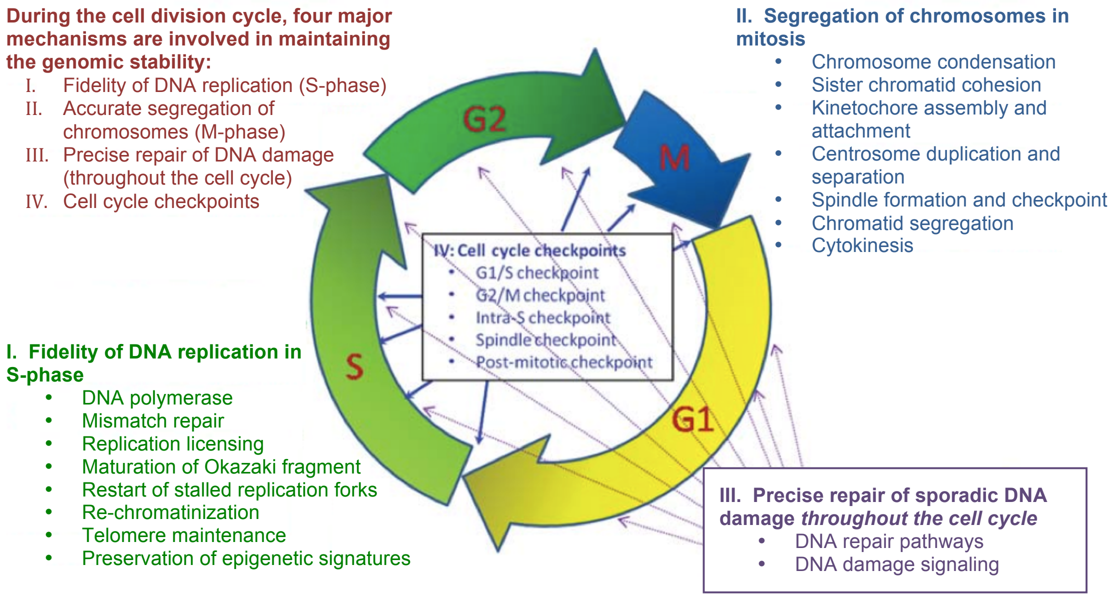

During the cell division cycle, four major mechanisms are involved in maintaining the genomic stability:
I. Fidelity of DNA replication (S-phase)
II. Accurate segregation of chromosomes (M-phase)
III. Precise repair of DNA damage (throughout the cell cycle)
IV. Cell cycle checkpoints
II. Segregation of chromosomes in mitosis
Chromosome condensation
Sister chromatid cohesion
Kinetochore assembly and attachment
Centrosome duplication and separation
Spindle formation and checkpoint
Chromatid segregation
Cytokinesis
G2
M
S
G1
IV: Cell cycle checkpoints
G1/S checkpoint
G2/M checkpoint
Intra-S checkpoint
Spindle checkpoint
Post-mitotic checkpoint
I. Fidelity of DNA replication in S-phase
DNA polymerase
Mismatch repair
Replication licensing
Maturation of Okazaki fragment
Restart of stalled replication forks
Re-chromatinization
Telomere maintenance
Preservation of epigenetic signatures
III. Precise repair of sporadic DNA damage throughout the cell cycle
DNA repair pathways
DNA damage signaling

Figure 2. Overview of the major mechanisms to maintain genomic stability during the cell cycle (figure from Z. Shen, 2011).

The driving force to select alternative repair pathways largely depends on the cell cycle status of the cell (Box1).

In the case of DSBs, although the exact mechanisms underlying DNA repair pathway choice and the precise role of the proteins involved in this process are still to be elucidated, there are a series of factors known to influence this selection, including the source and dose of the damage, the kind of lesion, the chromatin context of the DSB, the process of DSB resection, the cell type and the cell cycle phase [6,69-71].

NHEJ uses limited or no sequence homology to rejoin juxtaposed ends and acts throughout the cell cycle, being the major pathway in G1 [72], while HR is limited to the S, G2 and M phases, since it requires a homologous sister chromatid as a template to properly align and seal the broken DNA ends [29].

Apart from being controlled by DNA damage response signaling pathways, HR is also tightly influenced by a high cyclin-dependent kinase activity that triggers end resection, and which is present only in the S, G2 and M phases of the cell cycle [73].

	G1	S	G2-M
DSBs or SSBs	NHEJ	HR-mediated fork restart	HR-mediated repair
Mismatches		Mismatch repair	
Bulky lesions	NER	Template-switch-mediated damage bypass TLS-mediated damage bypass	

Box1: Main repair mechanisms used in the different cell-cycle phases.

1.2 THE DNA DAMAGE RESPONSE

1.2.1 A combined signaling and genome-maintenance network

Maintenance of genome integrity is a continuous task in all cells and to preserve cellular viability and prevent disease, a perfectly synchronized response to DNA damage is essential. Eukaryotic cells have an intricate genomic maintenance network formed by multiple repair pathways and diverse sensing, checkpoint, signal-transduction, and effector systems linked to replication, transcription, recombination, chromatin remodeling, and differentiation [74]. Besides, genome maintenance safeguards the integrity of mitochondrial DNA [75] and includes the complex telomere-processing machinery [76].

The DDR is a signal transduction cascade of proteins composed of sensors, mediators, transducers and effectors [77] (Figure 3). The DDR starts with sensing the damage at the break sites by a variety of sensor proteins such as the MRN complex [60,78], members of the poly(ADP-ribose) polymerase (PARP) family [79], the Ku heterodimer [50] or proteins of the phosphatidylinositol 3-kinase-like protein kinase family —ATM, ATR, and DNA-PK— [80]. These sensors recruit mediator proteins to the damaged site and activate transducer proteins. The transducers amplify the damage signal and activate a complex signalling transduction cascade that activates effector proteins and induces cell cycle arrest to allow time for lesion removal prior to replication or cell division. When the damage cannot be properly repaired or when too many injuries are encountered, replicative senescence is induced or apoptosis is activated in order to protect the organism from potentially harmful cells [29].
Effector proteins are directly phosphorylated by ATM/ATR or by kinases such as CHK1, CHK2 or MK2 [74], and they elicit a series of cellular responses.
Besides phosphorylation, the assembly of the DDR cascade depends on other posttranslational modifications like acetylation, ubiquitination, sumoylation, poly(ADP-ribosylation), methylation or neddylation [81]. These modifications are specifically recognized by a variety of proteins, many of which mediate the recruitment of other DDR factors to sites of DNA damage. This recruitment can be visualized by light microscopy as nuclear domains, or *foci*, which are highly dynamic structures subjected to a precise spatiotemporal regulation [82] and whose formation involves protein-protein interactions [5]. These *foci* have become a tool to evaluate the presence of certain proteins at DNA breaks.

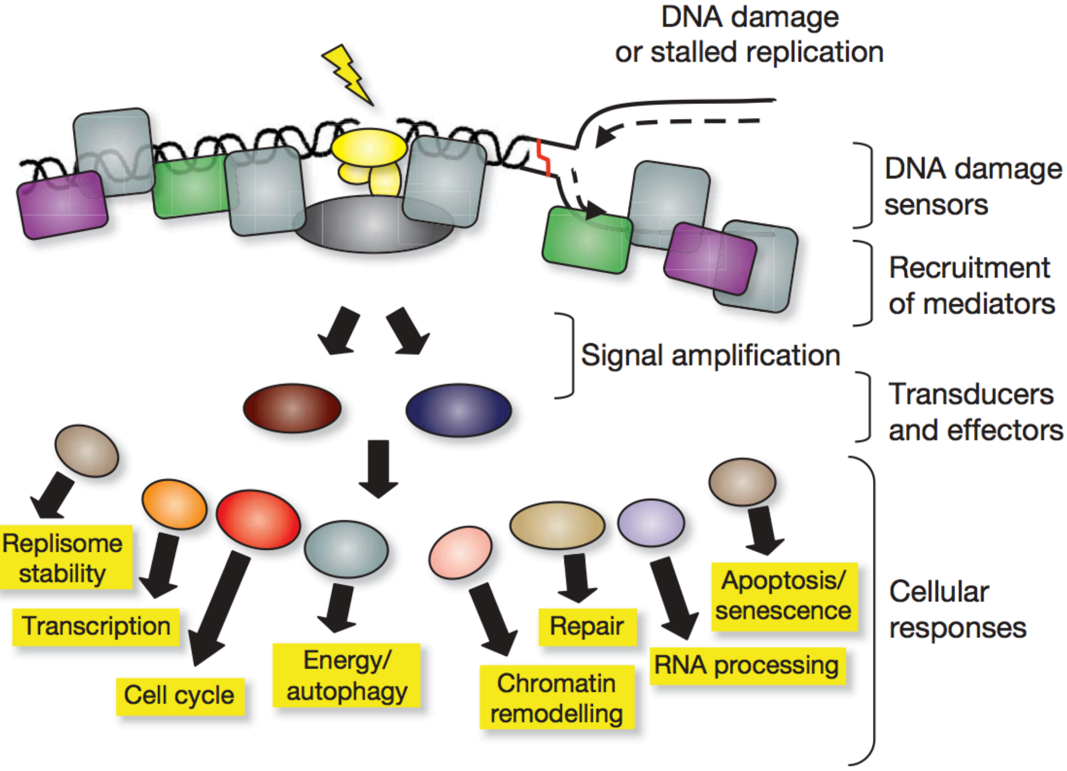


Figure 3. The DDR: a signal transduction cascade of proteins composed of sensors, mediators, transducers and effectors. DDR pathways contain four major components (some of which have overlapping functions): the "sensors" detect the damage and transmit signals to the "transducers," that convey the signals with the help of "mediators" to down-stream "effectors," which in turn execute the response (Image from Jackson and Bartek, 2009).

1.2.2 Chromatin and DDR

Recent research has provided insights into how chromatin responds to DNA damage and how cells mobilize large segments of chromatin to protect the genome against destabilizing effects posed by DNA lesions, and thus guard genome integrity.

DNA lesions can trigger histone alterations, nucleosome repositioning and changes in higher-order folding of the chromatin fibre. These modifications cause massive accretion of proteins in large segments of lesion-flanking chromatin that are visible as nuclear foci, the study of which has led to unravel the molecular pathways that reshape chromatin around DNA lesions.

DNA-damage-induced chromatin responses are promoted by PTMs of histones and histone-binding proteins, which generate modifications on the chromatin structure that need to be carefully tuned in space and time for proper damage signaling, DNA repair and modifications reversal after completion of DNA repair. For instance, while histone deacetylations and PAR-dependent events generally tend to silence or compact chromatin, most other modifications induce chromatin relaxation [83].

Some repair events are context-specific and determined by the status of the loci before DNA damage. For example, simple DSBs taking place in euchromatic DNA can be rapidly ligated by the NHEJ core proteins, while more complex DSBs and those occurring in heterochromatic regions of DNA activate the ATM kinase, which in turn triggers the DNA damage response, and the recruitment of HR elements [73].

1.2.3 The ATM/ATR pathway

Both ATM and ATR are key regulators of the DDR, coordinating cell cycle transitions, DNA repair, DNA replication and transcription, RNA splicing, metabolic signaling and apoptosis, among many other cellular activities [84,85]. One of the most recent canonical DDR pathway representations (see figure 4a and 4b) is the one described by Harper and Elledge [74].

DNA damaging agents (e.g. ionizing radiation) can cause DSBs and, as stated before, when these are located in heterochromatic regions of DNA, and when NHEJ fails, they lead to the activation of ATM. ATM then phosphorylates and inactivates KRAB-associated protein 1 (KAP-1, also known as TIF1-beta or E3 SUMO-protein ligase TRIM28), a heterochromatin-building protein, to cause local chromatin relaxation [86-88], enabling access of signaling and repair components to the site of damage. Besides, ATM is essential in the immediate response of cells to DSBs and the following switch to ATR activation after DNA end resection [74,89].

a) ATM

ATM is recruited to DSB sites by the MRN complex (see figure 4a), and there its activation is mediated by MRN in conjunction with other proteins such as 53BP1 **[90]** and the tat-interactive protein 60kDa (TIP60, also termed KAT5) acetyltransferase [91], which also modifies chromatin at sites of DNA damage [92].
Next, ATM phosphorylates the variant histone H2AX on Ser 139 to form γH2AX [93]. This event is accompanied by dephosphorylation of the neighbouring Tyr 142, a residue constitutively phosphorylated in the absence of damage [94]. γH2AX provides a high-affinity binding site for the MDC1 protein, which in turn orchestrates the recruitment of many additional factors to the damaged DNA leading to the generation of IR-induced foci (IRIF) [95]. MDC1 retains MRN, which further enhances ATM activation and γH2AX expansion [96]. Then the E3 ubiquitin-protein ligase RNF8 is recruited to the DSB, where it binds MDC1, recruits other downstream E3 ligases such as UBC13, RNF168, HERC2 and RAD18, and initiates a complex ubiquitylation cascade of histones H2A and H2AX at the DSB-flanking region, which causes chromatin restructuration [97].
Sequentially, ATM phosphorylates HERC2, which stimulates the interaction of the latter with RNF8. HERC2 is needed for RNF8 to promote UBC13-dependent poly-ubiquitylation of H2A-type histones [98]. Next, RNF168 assembles at the DSB, interacts with ubiquitylated H2A, and by targeting H2A and H2AX, further propagates the ubiquitylation of H2A and other targets at the DSB [99].
Consecutively, RAD18 is recruited in an RNF8/UBC13-dependent manner, where it directly binds to the ubiquitin chains at the sites of DNA breaks through its Zinc finger domain. In response to DSB, RAD18 binds to RAD51C, which allows the accumulation of RAD51C at DNA damage sites and thus facilitates RAD51 foci formation and HR repair [100].

Another protein whose localization to DNA damage foci is dependent on RNF8 and UBC13 is the PAX-interacting protein 1 (PTIP), which is involved in the recruitment of

53BP1 to these sites and acts as a local assembly platform for chromatin modulating activities [101]. Besides, PTIP is thought to promote ATM signaling in response to genotoxic stress through its ability to interact with 53BP1 [102] and has recently been found to interact also with γH2AX [103]. PTIP and 53BP1 induce chromatin remodeling at the damage sites, which promotes the association of ATM with chromatin and induces the phosphorylation of ATM substrates such as SMC1A (Structural maintenance of chromosomes protein 1A) [104,105], which is a structural component of the cohesin complex, involved in gene expression regulation, maintenance of genome stability, and in sister chromatid cohesion. [106]

Ubiquitylated histones at sites of DNA damage mark the spot for important downstream factors of the DDR such as 53BP1 and BRCA1. BRCA1 is recruited to IRIF though the BRCA1-A complex, composed of the RAP80/UIMC1 protein, which binds to ubiquitylated histones, and the proteins Abraxas/FAM175A, MERIT40/NBA1, the de-ubiquitinating enzyme BRCC36/BRCC3, BRE/BRCC45, BRCA1 and the BARD1 (BRCA1-associated ring domain protein 1) E3 ligase [107-109].

In addition to BRCC36, several de-ubiquitylating enzymes also function at DSBs, like USP11 (Ubiquitin carboxyl-terminal hydrolase 11), which interacts with BRCA2 [110] and also regulates the recruitment to IRIF of a subset of DSB repair proteins including RAD51 and 53BP1 [111].

Regarding sumoylation, DSB repair is promoted by PIAS1 and PIAS4, small ubiquitin-like modifier (SUMO) E3-ligases that are recruited to the damage sites, and are needed for the complete accretion of repair proteins to these locations [112]. PIAS4 acts earlier in the DDR cascade, where it influences RNF168 and the subsequent RNF168-dependent protein recruitment, while PIAS1 appears to induce RAP80 and BRCA1 accumulation. Besides, SUMO proteins also regulate BRCA1's ubiquitin ligase activity [113].

ATM activation leads to the phosphorylation of CHK2, which regulates cell cycle checkpoints, and p53, a nuclear transcription factor that induces cell-cycle arrest, senescence or apoptosis in response to DNA damage [114]. Both ATM and CHK2 regulate p53 by preventing its ubiquitination by the RING E3 ligase MDM2, which, along with MDMX (MDM4), is part of a multi-component E3-complex that targets p53 for proteasomal degradation. DNA damage also induces ATM-dependent phosphorylation of MDMX, which is then selectively bound and degraded by MDM2 preceding p53 accumulation and activation [115].

The transcription factor SOX4 is also required for the activation of p53 since it enhances its acetylation and interacts with and stabilizes p53 blocking its MDM2-mediated ubiquitination and degradation [116]. In addition, MDM2 controls degradation of hnRNP K (Heterogeneous nuclear ribonucleoprotein K), a p53 cofactor that plays key roles in coordinating transcriptional responses to DNA damage [117].

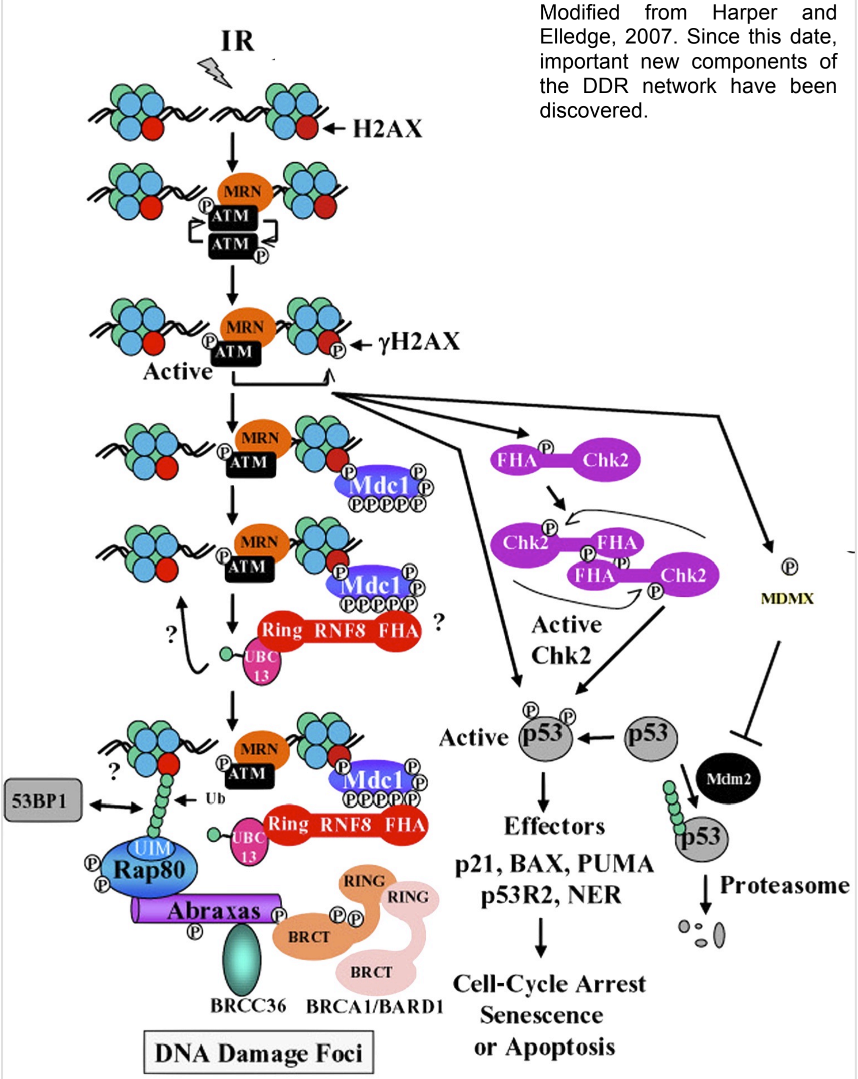


Figure 4a. DDR canonical pathway, ATM and DSBs. Modified from Harper and Elledge, 2007. Since this date, important new components of the DDR network have been discovered.

Moreover, ATM mediates phosphorylation of FBXO31 [118], a component of a SCF (SKP1-cullin-F-box) protein ligase complex that triggers the ubiquitination and subsequent degradation of cyclin D1 by the proteasome, resulting in G1 arrest after DNA damage [119].

b) ATR

While ATM is set in motion at DSBs, ATR is activated when ssDNA are generated at stalled replication forks (see figure 4b) or due to the processing of DSBs ends, and also responds to damage by ultraviolet light. Subsequently, the single strand DNA-binding protein replication protein A (RPA) binds the newly created ssDNA overhangs, and recruits ATR via ATR-interacting protein (ATRIP) to regulate the checkpoint response [120]. RPA also recruits SMARCAL1, an ATP-dependent annealing helicase involved in the replication stress response [121].

ATR activation depends on RAD17 loading of the PCNA-related 9-1-1 (RAD9, RAD1 and HUS1) complex onto DNA through a RAD9–RPA interaction. Then TopBP1 (DNA topoisomerase 2-binding protein 1) and Claspin are recruited to the site to be phosphorylated by ATR, which also phosphorylates RAD17 and the 9-1-1 complex. Consequently, RAD17 and Claspin, together with the TIM-Tipin complex promote ATR phosphorylation and activation of CHK1 and other kinases such as Tao and MK2 [7] which, in turn, phosphorylate effector proteins that control the cell cycle checkpoints, stabilize stalled forks, repair collapsed forks and prevent late origin firing [122].

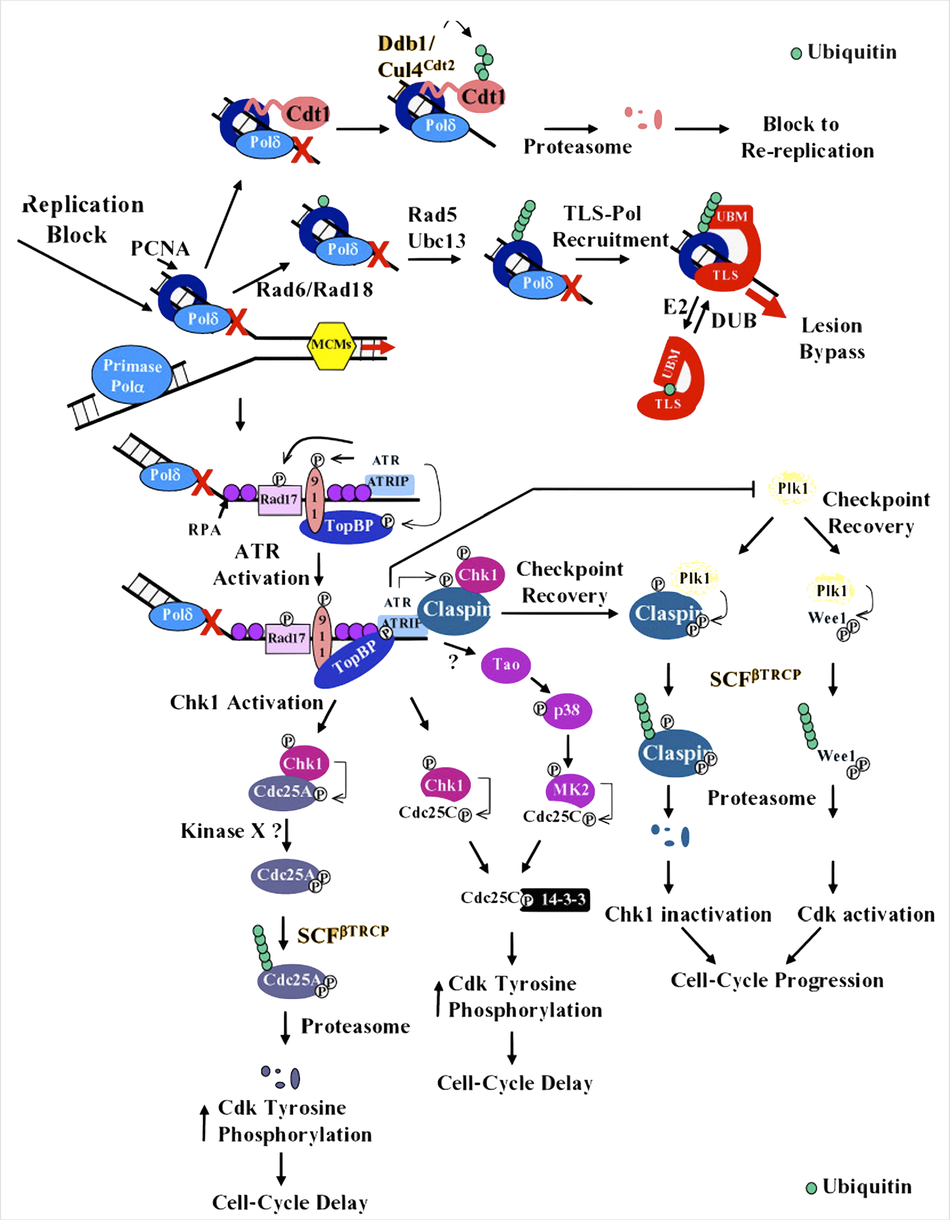


Figure 4b. DDR canonical pathway, ATR and replication block, from Harper and Elledge, 2007.

c) The ATM/ATR pathway and the cell cycle

Both the ATM/CHK2 and the ATR/CHK1 sub-pathways lead to the phosphorylation and inactivation of proteins of the CDC25 family of dual-specificity phosphatases (see figure 4a and 4b), which play an important role in driving dividing cells through the cell cycle [7].
CDC25C phosphorylation by checkpoint kinases leads to cytoplasmic sequestration of the dual-specificity phosphatase by 14-3-3 proteins [123], and phosphorylation of CDC25A by CHK1 leads to its ubiquitination by the SCFb-TRCP ubiquitin ligase and its subsequent degradation [124]. CDC25A regulates the G1/S transition by controlling CyclinE and CyclinA/Cdk2 activity, and also seems to play a role in facilitating the G2/M transition by activating CyclinB/Cdk1 [125].

In a normal G2/M transition, PLK1 (Polo-like kinase 1) phosphorylates WEE1 and Claspin, generating a phosphodegron (specific phosphorylated sequence of amino acids) that targets them for destruction via ubiquitination by SCFb-TRCP [124], leading to Cdk activation, reduced CHK1 signaling and cell-cycle progression [126]. PLK1 also promotes nuclear translocation of CDC25C [127] and inhibits CHK2 and 53BP1 [128].

When the DDR is triggered, on the one hand PLK1 is inhibited, which prevents the formation of the WEE1 phosphodegron and, on the other hand, CHK1 and CHK2 kinases regulate CDC25, WEE1 and p53, which eventually inactivate cyclin-dependent kinases, thus inhibiting cell-cycle progression [7].

As mentioned previously, an important role of the DDR is the inhibition of DNA replication during repair to prevent polymerases from encountering DNA damage. Part of this regulation occurs by targeting the DNA replication factor CDT1 for ubiquitin-mediated destruction by an SCF-like ubiquitin ligase composed of Ddb1, Cul4, RBX1 and Cdt2 [129].

After successful DNA repair, PLK1 and phosphatase Wip1 switch off the checkpoint by contributing to the activation of cyclin B/Cdk1 and by allowing checkpoint recovery, which leads to regain of the ability to exit the cell cycle arrest [130].

1.2.4 DDR and disease

Due to its essential role in safeguarding the genome, the DDR signalling pathway is crucial to preserve mammalian health [77]. Cells defective in DDR and DNA repair mechanisms normally display heightened sensitivity towards DNA-damaging agents. Besides, the majority of the mutations and large genomic alterations (loss of heterozygosity, amplifications, etc.) that are relevant to cancer originate from aberrant genome maintenance [4].

The DDR has gained much attention because of its involvement in cancer [131-133], aging-related pathologies and other diseases and complex disorders (see ST10 in Annex) such as Ataxia-telangiectasia (ATM deficiency) [134], Seckel syndrome (ATR deficiency) [135], Nijmegen breakage syndrome (caused by mutations in Nbs1) [136] or Cockayne syndrome (caused by mutations in the ERCC6 and ERCC8 genes) [137]. Moreover, alterations to the pathway generate genomic instability impairing the cell viability. This has produced much work in human [4,138,139], where detailed and extensive studies have been conducted in particular components of the DDR [90,120,140].

DNA break-associated proteins and the *foci* that they assemble into are of considerable medical importance, with defects in them being associated with various pathologies, particularly cancer (reviewed in [83]). Besides the widely acknowledged role of BRCA2, other proteins are involved in cancer development. For instance, the ALC1 chromatin-remodelling enzyme is frequently amplified in human hepatocellular carcinomas, raising the possibility that unscheduled chromatin relaxation contributes to the pathogenesis of this malignancy. Mutations of impact are for instance those affecting the RNF168 ubiquitin ligase first associated with RIDDLE syndrome [141], and homozygous deficiency of RNF168, that underlies a radiosensitivity syndrome that mimics ataxia-telangiectasia [142]. In this regard, it is notable that the immunodetection of γH2AX foci, which indirectly measure DSB formation and repair, is showing promise as a sensitive diagnostic tool to detect cancer cells and also monitor cancer progression and assess responses to treatment [143]. Moreover, the existence of many druggable protein targets in DNA break-associated events is providing exciting opportunities for developing new therapeutic agents that, by exploiting differences between normal cells and cancer cells, have the potential to markedly improve cancer management [144]. Besides, the possibility of using miRNAs as potential therapeutic candidates is beginning to be addressed [145]**.** On the contrary, little is known still about the maintenance machinery of the epigenome and its contribution to aging and cancer [4].

1.3 NETWORK STUDIES TO APPROACH EVOLUTION

1.3.1 Introduction: what has been done?

Since the complete sequencing of the first organism in 1995 [146], the rate of growth in available genomes has increased exponentially, and the pace to complement this data with functional studies has become unfeasible. As of November 2012, more than 18888 genome and metagenome sequencing projects had been developed, from which 3811 are completed genomes according to the Genomes OnLine Database [147].

Given the unfeasibility of experimentally characterizing all these data, it is necessary to use computational methods to analyze nucleic acid and protein sequences and structures, which has helped develop high confidence predictions regarding biological function directly from genome sequence.
To this purpose it is important to predict the proteins included in pathways; thus, many approaches have been developed to expand pathways in unknown organisms, like the use of phylogenetic profiling [148], protein domains analysis to increase the mapping of

proteins to pathways [149] or machine learning methods to infer functional relations in predicted protein networks [150].

The genomic revolution permitted the first glimpses of the architecture of regulatory networks and pathways. Combined with evolutionary information, the network perspective of biological processes leads to significant insights into how organismal systems have been shaped along evolution. Moreover, the birth of genomics permitted the first robust reconstructions of evolutionary relationships between organisms and also allowed the identification of the genomic correlates of main morphological transitions in evolution, such as the emergence of eukaryotes and the origins of multicellularity [151,152].

Concerning the evolution of networks, most of the early comparative genomics research was done in bacteria and was focused mainly in metabolism, in which the evolution of metabolic pathways and how the networks assemble were analyzed. Also, the specialization and evolution of enzymes and the role of HGT in the formation of gene clusters in operons involved in metabolism were investigated [153]**.**

In this regard, and more recently, several studies have been conducted about the evolutionary mechanisms that shape metabolic pathways and the evolution of metabolic network organization, in which a wide range of organisms, from prokaryotes to complex eukaryotes, have been compared [154].

Beyond metabolic pathways, other biological systems like the apoptosis network [155], the Ras switch genetic system [156,157], the insulin/TOR signal transduction pathway [158] or the cellular stress response [159] are examples of the analysis of the evolution of regulatory processes.

The evolutionary perspective is essential in comparative studies since all organisms have an evolutionary history and thus, analyzing the genetic similarities and differences among species allows us to better understand how and why these variations arose. Besides genomics, many fields of comparative biology including developmental biology, physiology or ecology have extensively used evolutionary information in their studies.

1.3.2 Homology-based extension of pathways

As aforementioned, determining the function of unknown proteins encoded by the DNA sequences produced in sequencing projects has become an important challenge. Functional annotation has been traditionally done using only sequence similarity, but below 25% similarity the "twilight zone" is reached (see Figure 5), and assigning evolutionary relationships becomes unfeasible because the determination of homology at this level of low identity is extremely difficult [160]. Further development of sequence structure techniques allowed lowering this threshold; in particular, the use profile-based identification of homologs due to the conservation of position-specific patterns important for the three-dimensional folding and function of proteins. Thus, utilizing structural information, homology-based detection can be successfully performed with sequence identity as low as 15-10% [161].

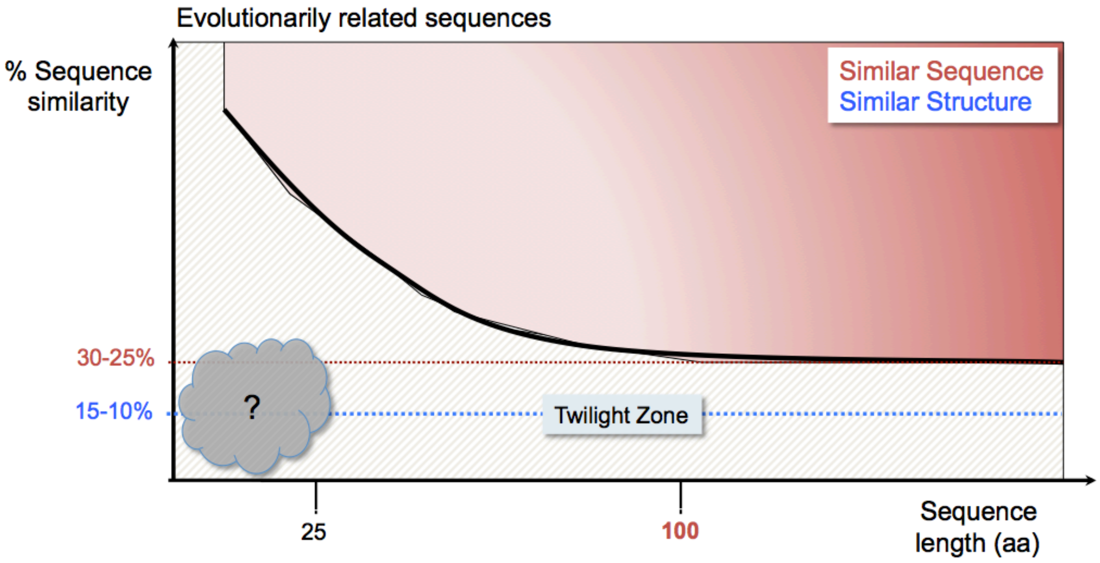


Figure 5. Pair-wise sequence similarity versus alignment length (modified from Chothia and Lesk, 1986; and Rost B. 1999). The "twilight zone" is reached when the proteins' sequence similarity is below 25%, or below 15-10% when structural information is also used. The length of the protein is also an essential parameter to establish potential homology; thus, at low sequence similarity and with short sequences, no evolutionary relationship can be determined.

Even so, prediction of protein function from homology-driven approaches presents certain problems. Though all homologous proteins should have a common ancestry and thus are expected to have similar three-dimensional structures and to perform the same or highly related functions, these proteins might evolve different functions due to sequence variation or context-dependent changes [162].

Recently, computational methods for inferring protein function are complementing the traditional sequence homology-based approaches with information on the context of a protein in cellular networks. These network-based functional inference techniques provide data on the proteins' function and their role in a given network, and also offer a better understanding of the function of uncharacterized proteins [163].

1.3.3 Methods for homology-based annotation of sequences

Different computational methods can be applied to extract biological information from proteins and to annotate unknown sequences. Some of the most commonly used methods use pair-wise comparisons and multiple sequence alignments; motif, profile and pattern searching; and structure prediction (fold assignment).
Most of these function-prediction methods rely on inferring relationships between proteins by transferring functional annotations from one to the other. An important challenge in this regard is deciphering the connection between the detected similarities (both in sequence and in structure) and the actual level of functional relatedness [164].

a) Sequence-based methods:

Initially, proteins identified in genome sequencing projects were habitually annotated by sequence homology inferred using pair-wise alignment tools such as BLAST [165]. Thus, sequence similarity has been used as surrogate of function. This method for homology-based annotation transfer has traditionally been the most widely used in computational function prediction, though it has been shown with technological developments that, in general terms, is not very accurate.

Another approach is to use partial information of proteins, using protein domains. Domains are the functional units and building blocks of proteins, and therefore, studying proteins at a domain level allows more accurate functional inference [166]. Besides, due to genome rearrangements during evolution, domains have duplicated, fused, recombined and have been inserted and depleted within sequences to produce proteins with novel structures and functions [167-169]. Consequently, domain analysis is useful for predicting the function of novel domain combinations that possibly gave raise to new protein functions.

In the available resources, a family of domains is represented as a multiple sequence alignment, which is then converted into a statistical family signature profile (for example in PROSITE [170] and NCBI-CDD [171]) or into a profile hidden Markov model (HMM) provided by the HMMER package [172], such as in InterPro [173], Pfam [174] and SMART [175]. These profiles capture position-specific information about how conserved each column of the alignment is, and which residues are likely to be in that position.

For most of them there is 3D information that is incorporated in the structural profile. Therefore, regardless of the algorithm used, the precision of these sequence-based methods is influenced by the type and amount of information on the particular protein family but, in general, they are fairly accurate.

b) Structure-based methods:

During evolution, the three-dimensional structure of homologous proteins usually remains more conserved than their sequence due to spatial and physical restraints [160]. Thus, similarities in protein structure can be more consistent than similarities in sequence for identifying distant homologs, which sometimes preserve a common function [176]. The two most complete structure-based family resources, CATH [177] and SCOP [178], classify domains into structural classes and evolutionary families. Besides, other structure-comparison methods such as HHPred [179] can identify sequences with structural similarities in the Protein Data Bank [180], which may be functionally related. However, in all cases, transferring function from one protein to another should be done with caution, since two proteins may have similar fold but different functions (i.e, the TIM-barrel scaffold) [164].

1.3.4 Limitations of function transfer by homology-based methods

The underlying principle behind homology-based annotation comes from the Neutral Theory of Molecular Evolution [181], which predicts that if two sequences have a high similarity at sequence level, then they have a common phylogenetic ancestor and, subsequently, they should have similar three-dimensional structure.

This theory also serves as the null model of molecular evolution and plays a central role in data analysis. According to this theory, most evolutionary change is invisible to natural selection and thus it is evolutionarily neutral. The outcome of neutral mutations is determined by random genetic drift, a stochastic process by which a neutral mutation will be lost to evolution, but sporadically by chance a neutral mutation can become the predominant variant in a population [182].

Sequence similarity, a mathematical concept used as a proxy for homology (an evolutionary concept), is frequently used to support the transfer of functional annotations from experimentally characterized proteins to new sequences lacking functional characterization. However, functional annotation via sequence similarity seems to have reached its limit since most of the newly identified proteins do not show significant sequence similarity with well-studied protein examples [183]. Besides, the power of homology-based annotation is being challenged due to the effects of gene duplications and domain shuffling events, which might lead to divergence of function.

Moreover, homology-based annotation transfer has led to error propagation even across human curated sequence databases. Recently, it was found that function prediction error (i.e., misannotation) is a significant issue in all databases but the manually curated database Swiss-Prot [184]. Morevoer, it must be pointed out that similarity in sequence, structure and function has only been verified for globular segments of proteins. For non-globular regions, similarity of sequence is not necessarily a result of divergent evolution from a common ancestor but the consequence of amino acid sequence bias. This has led to many proteins inheriting completely wrong function assignments from protein databases containing domain models with transmembrane regions and signal peptides, which are non-globular segments of proteins with a hydrophobic bias [185].

Although domain assignation methods have been extensively studied, still nowadays their accuracy to predict domain boundaries is not entirely satisfactory. Various methods provide reliable predictions if a structural template for the protein is available, but when this is not the case, the experimental annotation used to infer the function might refer to a different domain or region in the analyzed sequence, an thus the annotation by homology-based inference would be erroneous [186]. Besides, in given cases the availability of a structural template does not guarantee the identification of protein function [187].

Regardless all these caveats, there has been a boost in the number and variety of automated approaches for functional inference. These automated methods are based on different features, such as sequence identity (best bidirectional hits (BBHs), orthologs detection, etc.), sequence profiles, protein structure patterns, chromosomal location, expression profiles, protein-protein interaction data, phylogenetic information, and gene co-evolution **[163]**.

Although homology-based inference of function can produce misannotations, the frequency of such errors is fairly small compared to the number of correct inferences [185]. In addition, these homology errors can be alleviated by careful tree-based inference with extensive human input.

1.3.5 Homology, orthology and paralogy

It is very important to emphasize that homology refers to sequences that share a common ancestor. However, the term 'homology' is still often incorrectly used instead of 'similarity' in articles describing a comparison of protein or nucleic acid sequences [188].

Orthology and paralogy are key concepts of evolutionary genomics and reflect two different kinds of evolutionary relationships. Orthologs are defined as genes from different species that derive from a single gene in the last common ancestor of those species, while in-paralogs are genes that derive from a single gene that was duplicated within a genome after the speciation event [189] (see Figure 6).
As with homology, the term 'orthology' has been frequently misused in comparative genomics and specially in the fields of molecular and cell Biology, to indicate genes that are functionally equivalent across species, but without any reference to speciation events and many times without any common origin **[190].**

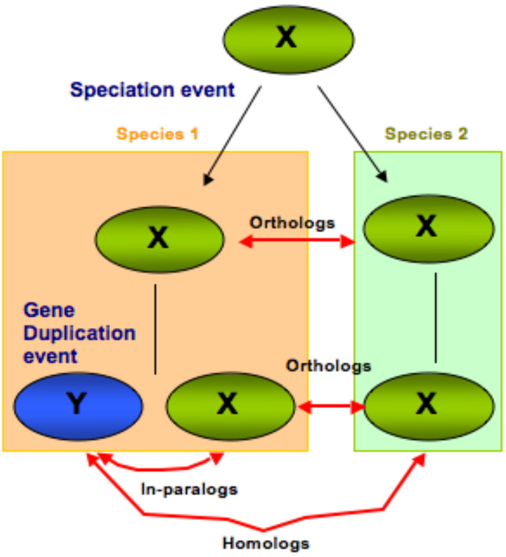


Figure 6. Diagram showing the hypothetical evolutionary history of a gene. An ancestral species split into two daughter species, each of whose genomes contains one copy of gen X. Genes X in the two species are orthologs. If, after the speciation event, gene X duplicates within species 1 and a new gene Y emerges, genes X and Y within the same species are in-paralogs, while they are homologs if we consider the two different species.

While orthologs in different species tend to retain identical or similar molecular and biological functions, paralogous proteins are likely to diverge along evolution to carry out different functions through sub-functionalization or neo-functionalization paths [191].

Nevertheless, functional conservation among orthologs should be inferred with caution as some orthologous genes can diverge functionally even among closely related organisms [192] and some paralogs could retain the original function.

Distinguishing between orthology and paralogy is critical for the construction of a reliable evolutionary classification of genes and to consistently annotate newly sequenced genomes. This distinction can be achieved by using sequence-similarity patterns, by analyzing the specific conservation of residues responsible for function in the family of orthologous proteins, or on the basis of the protein structure. However, sequence similarity of orthologs may decrease with divergence time, and this poses a problem when identifying orthologs in phylogenetically distant organisms. In spite of this, genome comparisons have shown that orthologous relationships with genes from taxonomically distant species can be established for most genes **[193].**

Automated approaches that infer orthology relationships from pair-wise sequence comparisons alone were the first to be developed. Although these methods perform reasonably well, they have numerous drawbacks that can lead to annotation errors or misinterpretation of data **[194]**. An example is when genes are lost after duplication events, which would lead to paralogs being identified as orthologs when using BBHs methods (see figure 7).

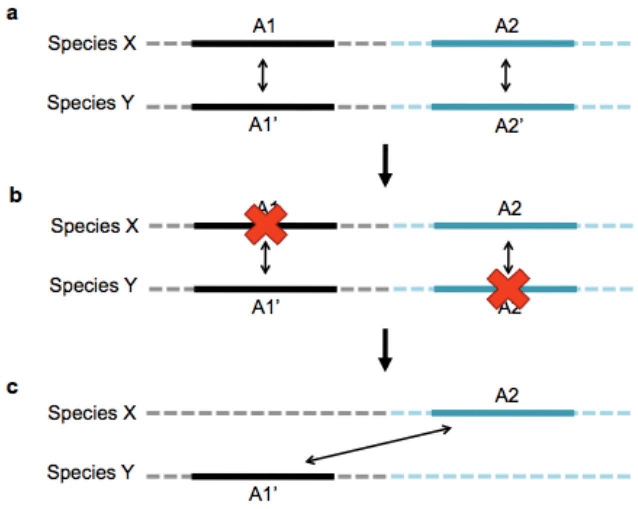


Figure 7. Diagram showing how differential gene loss after a duplication event can lead to the incorrect prediction of orthology. The speciation event occurs after the duplication of an ancient gene A, and thus, (**a**) species X contains genes A1 and A2, and species Y has genes A1’ and A2’. Genes 1 and 2 are in-paralogs within the same species, while A1 - A1’, and A2 - A2’ are orthologs between the two species. (**b**) If gene A1 is lost in species X and gene A2’ is lost in species Y, when using BBH methods to identify orthology relationships, the most similar genes between both species will be A1’ and A2 (**c**), and they will be considered as orthologs when they are actually paralogs.
Double-headed arrows indicate orthology relationships.

Another problem when identifying orthology is the existence of convergent evolution. Examples of this process are the antifreeze glycoproteins from the fishes *D. mawsoni* and *B. saida*, which show 69% sequence identity but are not homologous [195].

An additional difficulty in defining orthology relationships among proteins is that they frequently contain different domains that may have followed distinct evolutionary paths. These proteins can be generated by fusion processes and recombination between genes, and may lead to the acquisition of a new domain by a member of a given protein family after recombination with another family. These are represented by multidomain families, where the different domains should be considered as independent evolutionary units and the orthology relationships should be first established among the core domains and then extended to the newly acquired domains or the flanking regions [194].

1.3.6 Methods for the identification of orthology

The myriad of algorithms and methods available for the identification of orthology [196,197] can be classified into two main groups [198]:

a) Phylogenetic tree-based approaches

These methods are more precise and less prone to error than pair-wise heuristic approaches when used carefully, because they use information on the evolutionary history of the genes, but on the other hand, they demand large amounts of time and computing power, so the use of these methods is limited to single gene families or small datasets. Some examples of phylogenetic tree-based methods are HOPS (Orthostrapper/hierarchical grouping of orthologous and paralogous sequences) [199], RIO (Resampled inference of orthologs) [200], COCO-CL (COrrelation COefficient-based Clustering) [201] and MetaPhOrs (MetaPhylogenyBased Orthologs) [202].

b) Heuristic best-match methods

These methods are usually easy to automate and implement and are fast since they are BLAST-based. Besides, the BLAST score ranking they provide has proven to generally be a good statistical predictor of orthology at genome scale, especially the BBHs. The main drawbacks of these heuristic methods are that they do not use an evolutionary distance model and that they fail to detect differential gene loss (see figure 7). Some of the most widely used heuristic best-match methods are COG/KOG (Clusters of Orthologous Genes) [203], InParanoid [204], OrthoMCL [205] and DODO (Domain-based detection of orthologs) [206].

For evolutionary inference purposes, orthology relationships are generally best inferred by phylogenetic analyses [193,194].

1.4 PHYLOGENETIC APPROACHES TO COMPARATIVE GENOMICS

Phylogenetics is the discipline devoted to delineate the evolutionary relatedness among organisms or taxa through molecular sequencing data, and is considered to be the seed discipline that contributed to the development of Computational Biology [207]. Phylogenetic analyses are performed by a variety of automatic methods and algorithms in order to reconstruct a phylogenetic tree representing the evolutionary history and relationships among the sequences and species involved [194].

Comparative biological analysis can be carried out only in the context of a phylogeny. Phylogenetic approaches permit different types of comparative analyses, including detection of domain shuffling and horizontal gene transfer, speciation and duplication events, reconstruction of the evolutionary diversification of gene families, assessment of gene orthology and paralogy relationships, tracing of evolutionary variation in protein function at the amino acid level, and prediction of structure-function relationships [162]. Limitations are that they require a good coverage of genes or species, and sequence quality is paramount.

To date, nucleotide or preferably amino acid sequences are still the most used data type for phylogenetic reconstruction. In most phylogenetic methods, sequence alignments are extensively used to construct and refine phylogenetic trees to classify the evolutionary relationships between homologous genes from genomes of divergent species.

1.4.1 Phylogenetic methods

The most commonly used methods for phylogeny reconstruction from sequence data are: Parsimony, Neighbour Joining, and probabilistic-based methods (Maximum Likelihood (ML) and Bayesian inference (BI)). The main difference between the probabilistic-based methods is that ML generates one tree while BI creates thousands; also ML is computationally much demanding than BI.

All methods depend upon an implicit or explicit mathematical model describing the evolution of the sequences from the species included in the study (see Table 2 for a comparison of the methods and main characteristics of each of them). Although different methods may identify different topologies as optimal, the disparities among these topologies usually involve poorly resolved groupings.

The most frequent approach for phylogenetic analysis is generally a two-step process: first, the input DNA or preferably protein sequences are aligned with a multiple sequence alignment (MSA) program, such as MAFFT [208], T-Coffee [209] or the original ClustalW [210] that has been replaced by the former ones.

This step is critical as the quality of the alignment is the most crucial step in phylogenetic reconstruction [211}], and frequently neglected.

Then, the phylogeny is inferred from the alignment using phylogenetic tools such as Mr. Bayes [212], PHYLYP (Phylogeny Inference Package) (www.phylip.com/) or MEGA (Molecular Evolutionary Genetic Analysis) **[213].** Most phylogenetic reconstruction methods assume a fixed alignment of the input sequences, which is known to have impact on the accuracy of the inferred phylogeny.

According to a methods comparison study, Bayesian trees estimated from protein sequences alignments are the most accurate, followed by Maximum Likelihood trees calculated from DNA sequences and way less Parsimony trees estimated from protein sequences **[214].** In these cases, it is important to select a correct evolution model for the proteins to be analyzed. Software as ProtTest [215] allows inferring the best evolution model prior to running ML trees.

MrBayes is a program for Bayesian inference of phylogenies from DNA and protein sequences, and morphological characters [216]. It assumes a prior distribution of tree topologies and uses Markov Chain Monte Carlo (MCMC) methods to search tree space and infer the posterior distribution of topologies. The program outputs posterior distribution estimates of trees and parameters. It can use different models of sequence evolution, and allows for rate variation among sites and for multiple-chain Metropolis-coupled Markov Chain Monte Carlo (MC3) runs for more extensive search. Besides, it can spread jobs over a cluster of computers using the MPI message-passing interface implementation [217].

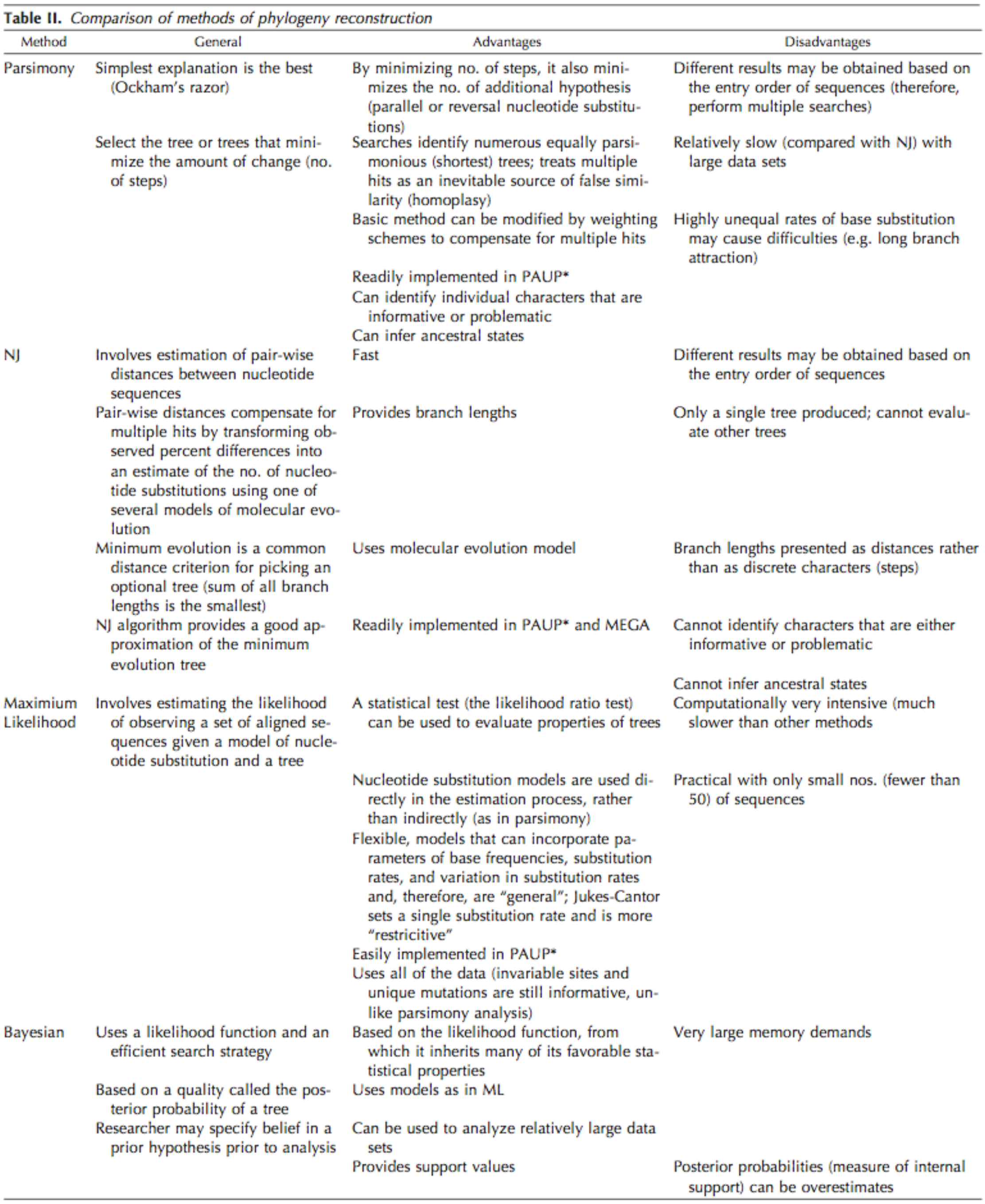


Table II. *Comparison of methods of phylogeny reconstruction*

Method	General	Advantages	Disadvantages
Parsimony	Simplest explanation is the best (Ockham's razor)	By minimizing no. of steps, it also minimizes the no. of additional hypothesis (parallel or reversal nucleotide substitutions)	Different results may be obtained based on the entry order of sequences (therefore, perform multiple searches)
	Select the tree or trees that minimize the amount of change (no. of steps)	Searches identify numerous equally parsimonious (shortest) trees; treats multiple hits as an inevitable source of false similarity (homoplasy)	Relatively slow (compared with NJ) with large data sets
		Basic method can be modified by weighting schemes to compensate for multiple hits	Highly unequal rates of base substitution may cause difficulties (e.g. long branch attraction)
		Readily implemented in PAUP*	
		Can identify individual characters that are informative or problematic	
		Can infer ancestral states	
NJ	Involves estimation of pair-wise distances between nucleotide sequences	Fast	Different results may be obtained based on the entry order of sequences
	Pair-wise distances compensate for multiple hits by transforming observed percent differences into an estimate of the no. of nucleotide substitutions using one of several models of molecular evolution	Provides branch lengths	Only a single tree produced; cannot evaluate other trees
	Minimum evolution is a common distance criterion for picking an optional tree (sum of all branch lengths is the smallest)	Uses molecular evolution model	Branch lengths presented as distances rather than as discrete characters (steps)
	NJ algorithm provides a good approximation of the minimum evolution tree	Readily implemented in PAUP* and MEGA	Cannot identify characters that are either informative or problematic
			Cannot infer ancestral states
Maximium Likelihood	Involves estimating the likelihood of observing a set of aligned sequences given a model of nucleotide substitution and a tree	A statistical test (the likelihood ratio test) can be used to evaluate properties of trees	Computationally very intensive (much slower than other methods
		Nucleotide substitution models are used directly in the estimation process, rather than indirectly (as in parsimony)	Practical with only small nos. (fewer than 50) of sequences
		Flexible, models that can incorporate parameters of base frequencies, substitution rates, and variation in substitution rates and, therefore, are "general"; Jukes-Cantor sets a single substitution rate and is more "restricitive"	
		Easily implemented in PAUP*	
		Uses all of the data (invariable sites and unique mutations are still informative, unlike parsimony analysis)	
Bayesian	Uses a likelihood function and an efficient search strategy	Based on the likelihood function, from which it inherits many of its favorable statistical properties	Very large memory demands
	Based on a quality called the posterior probability of a tree	Uses models as in ML	
	Researcher may specify belief in a prior hypothesis prior to analysis	Can be used to analyze relatively large data sets	
		Provides support values	Posterior probabilities (measure of internal support) can be overestimates

Table 2. Summary of the main methods of phylogeny reconstruction (from [218]).

Robustness and reliability

Regardless of the method utilized to construct the tree, a numerical assessment on the reliability of the grouping should be provided in non-probabilistic based methods. The most commonly used method for this is "**phylogenetic bootstrapping",** which simulates obtaining new data on the relationships among a group of sequences by

resampling with replacement the same set of characters and performing a new phylogenetic analysis. This is done thousands of times and typically a further majority rule consensus tree is constructed for the resulting trees. The frequency with which specific groupings appear on the majority rule tree gives a measure of their support by the sequence data **[182]**.
Bootstrap values are conservative measures of phylogenetic accuracy. Values of 80% or more are considered as indicators of strong support, corresponding to "true" clades in experimental phylogenies [219]. In contrast, probabilistic inference allows to directly sample probabilities. Though different phylogenetic methods may yield dissimilar optimal topologies, the variations normally involve poorly supported clades, since those strongly supported usually appear in topologies independently of the method of phylogenetic inference used **[218].**

1.4.2 Limitations of phylogenetics

As aforementioned, the large demands of time and computing power needed to generate reliable trees have traditionally limited the use of phylogenetics to single gene families or datasets of moderate size. Moreover, phylogenetic trees are difficult to automate for genome scale data, and the topology of the tree is strongly dependent on the tree building method chosen. Besides, in some occasions, pair-wise comparison approaches have outperformed more complex algorithms that use sophisticated tree reconstruction and reconciliation approaches [198].

Generating a phylogenetic tree involves a estimation of divergence among the characteristics shared by the species being compared. In molecular studies, a crucial problem is producing a good MSA, especially in studies of genes from divergent taxa. Although alignment of nucleotide or amino acid sequences should be a major consideration, yet it remains one of the most complicated and badly understood aspects of molecular data analysis. Researchers should revise the MSA automatically generated and also modify the default settings in the phylogenetic programs to adapt the analysis to the type of data examined **[220].**

Sometimes, different genes of the same species evolve at distinct paces or speed. This is another prediction of the Neutral Theory [181], which suggests the existence of a molecular clock, where the different families evolve at distinct rates. Knowledge of how genes evolve and at what rate they do it is essential for understanding gene function across species or within gene families.

One of the most challenging facets of reconstructing evolutionary relationships using comparative genomic analyses and phylogenetics is the existence of extensive HGT events among organisms, which can produce unexpected phylogenetic tree topologies for some genes [221], for instance the chymotrypsin from fungi [222].

Other complication is when the analyzed sequences are involved in recombination events and they evolve under positive selection; thus, an increase in polymorphisms in some lineages and not others might blur the correlation between genetic similarity and evolutionary relatedness [223]. Besides, gene duplication followed by gene loss may produce alternative phylogenetic topologies, which may then be confused with HGT **[224].**

1.4.3 Incongruences between phylogenies: gene- vs. species-trees

The phylogenetic gene-trees constructed do not always perfectly reproduce the taxonomy tree, or in other words, the evolutionary tree that represents the historical relationships between the species being analyzed, as we are comparing sequences at our present time.

Common sources of incongruence between gene-trees and species-trees are:

- The taxonomic tree is an artificial tree prone to mistakes.

- Mixing paralogous and orthologous sequences. Only orthologous sequences can be used to infer taxonomic relationships, since paralogous sequences trace the history of gene duplication events within species.

- HGT events. These can make two sequences similar not because the organisms share a recent common ancestor, but because the species have horizontally acquired a genomic segment from another distantly related species, which is particularly frequent in bacteria, and also in eukaryotes. In these cases, phylogenetic trees with incorrect topologies and wrong branch lengths can be generated **[225].**

- Unequal amounts of divergence in different lineages. This can occur when DNA sequences evolve very rapidly in one lineage but not in another (i.e. amphibians and reptiles versus mammals **[226]**).

- Back mutations. The probability of multiple substitutions on the same site producing undetectable "aminoacid saturation" augments as time since the divergence of two taxa increases, and this can result in homoplasies.

- Convergent evolution. This process, that occurs in short stretches of sequences (for example the catalytic triad of serine-proteases with subtilisin in prokaryotes and the chymotrypsin clan in eukaryotes), or the presence of linear motifs that can make two sequences reach a high identity when sequence length is not accounted for.

1.4.4 The concept of gene ages

Recent work has provided evidence of the existence of a universal model of evolution along different groups of genes with associated evolutionary-based ages **[227]**. This universal distribution implies a steady-state process, with equal distributions of evolutionary rates among genes that are gained and genes that are lost.
A gene is considered to belong to a certain "gene age" (for example plant-specific genes in *A. thaliana*), if there are no detectable homologs of the encoded protein outside the given taxon. Thus, genes can be divided into different age classes according to their time of emergence, ranging from very ancient (already present in proteobacteria) to very modern (detected only in mammals).
Gene evolution can be described by the sequence evolution rate and by the proclivity of a gene to be lost or preserved during evolution. Generally, those genes that are indispensable for the organisms will be conserved in all lineages, as long as there is no substitute gene for their function.

According to previous studies [228,229], ancient genes usually encode large proteins with functions that are common to a broad range of cells. Also, these old genes are evolutionarily conserved, having strong selection pressure, slow sequence evolution and little propensity to be lost; their expression levels are high, and generally have numerous physical and genetic interactions. On the other hand, modern genes usually possess the opposite characteristics and are mainly involved in lineage-specific processes.

This simple model of genome evolution provides a novel and an amenable framework to systematically analyze the DDR network globally.

1.5 DDR AND EVOLUTION

The comparison of biological networks in different organisms is of particular interest in the fields of evolutionary and systems biology. Such comparisons eventually help us to understand the forces influencing the evolution of biological pathways and systems.

In the case of the study of DDR and DNA repair proteins, the first comparisons between species revealed that organisms possess multiple repair pathways, that the repertoire of repair mechanisms frequently vary between species and that differences in specificity are found in almost all types of repair [230-233].
Comparative studies of repair genes and DDR can be used to infer the evolutionary history of species, to understand the similarities and differences found among them, and to shed light into the evolutionary history of DNA repair processes and pathways.

Evolutionary studies have many other potential uses in the study of DDR and DNA repair including the prediction of functions for conserved uncharacterized genes [234-236] and domains [237]; the characterization of genes that are part of multigene families [238,239], the identification of motifs or domains conserved among homologs [240,241]; and the study of structure-function relationships of repair genes [242,243].

The boundary between evolution and repair is of high interest due to the influence that repair mechanisms and pathways have in evolutionary patterns. Since repair processes influence mutation rates and patterns, differences in repair mechanisms can lead to different mutational rates and evolutionary capabilities within and between species. Therefore, there is a complex interplay between the need for fidelity of transmission of genetic information to the offspring and the need for evolvability (that is, the ability of an organism to generate adaptive genetic diversity, and thereby evolve through natural selection)[244].

During evolution, organisms with high levels of genetic variation have had better chances to survive sudden environmental changes by random variations in their genome. However, as organisms evolved more complex genomes, genomic instability became mostly detrimental and the need for systems safeguarding the integrity of DNA increased [15].

1.5.1 Previous computational approaches / State of the art

DNA repair proteins and pathways have been studied in quite a few species, mainly Bacteria. However, the ecological and evolutionary diversity of such studies has been limited [231,245] and a systematic approach addressing this issue in a larger evolutionary scale has not been conducted.

The usual procedure to approach a given pathway is to focus in a particular model organism. Much work has been done in the popular components of DDR using yeast and flies. For instance, homologous proteins containing domains such as BRCT repeats, Tudor domains or kinase domains have been found to perform the same functions in different species, like the case of checkpoint kinases [246]. Less studied are proteins involved in the aforementioned post-translational modifications (reviewed in [5]).
In humans, examples of protein components of DDR that have received much attentions are BRCA1 [247], BRCA2 [248], ATM [249,250] and ATR [135,251], due to the great impact that their alterations produce in the pathway and their effect on disease [30,74,77,252] and aging [253]. Therefore, it has been proposed that targeting specific components of the pathway may be useful to address disease [254].
In alternative organisms, the focus has been directed to human important orthologs, for instance p53 in several mammals [255], CHK1/2 in *Neurospora crassa* [256,257] and ATM in plants [258].
Some work has been conducted analyzing DDR-related pathways, where the focus has been made on specific sub-modules or sub-networks in animals involving popular proteins like p53 and its interactors (exemplified by the p53-MDM2 interaction [259]). Alternative network approaches make use of evolutionary information [260] to establish their topological constraints, while others are based on protein-protein interactions only [261], or focus in checkpoint proteins [262].

From an evolutionary point of view, systematic analyses of DDR-related proteins have been conducted only in specific parts of the process as exemplified by the chromatin modifiers genes [263], where their evolutionary landscape has been compared in human and hundreds of eukaryotic genomes. Network dynamics approaches have been also conducted in yeast, where the effect of given perturbations affect the DDR transcription network [264]. Although these works are helping to understand the structure of the network, our understanding regarding how this network as a whole ever emerged is still far to be complete.

To understand how such a delicate network has been efficiently assembled since life emerged on earth, it is necessary to interrogate this question using a wide evolutionary framework that implies screening in very deep nodes of the species tree of life which is far to be resolved [265].

2 OBJECTIVES

There is one commonality that transcends all the differences among living organisms: for a species to survive, its cells need to replicate faithfully.
Therefore, if having a damage detection system to prevent genomic instability is universal, **we hypothesise** that there must be an essential core of components common to all living organisms, which expanded along evolution according to particular needs to suit specific organisms demands originated by different life-styles.

In this work we aim to delineate the emergence and evolution of the DDR pathway, and to understand its functional implications.

To this purpose we have established 4 specific objectives:

a) Establishing a consensus set of DDR components

By identifying well known DDR components from four model organisms in a set of species covering the whole tree of life and including the widest variability of phyla available, we aim to establish the “core machinery” of the DDR pathway.

b) Infer the age of the network comparing gene-content and gene-trees

Focusing on the human DDR network, we aim to clarify how the different DDR pathways have been shaped along evolution. In this work we try to reconstruct the ancient network by tracing the presence of DDR orthologs along evolution, and by determining the appearance of novel components along evolutionary lineages. To this purpose we will use information from species phylogenetic profiles and gene-trees.

c) Creating a curated pathway of the human DDR network

So far there is a lack of a formal representation of the human DDR network. In the pathways repositories only partial networks can be found, especially the repair fraction (Reactome, Kegg, etc). However, there are intricate relationships and overlaps among the different sub-networks.
By using the data available in the literature, we aim to manually reconstruct the human DDR network, illustrating how components of the sensing part are also participating in related pathways like cell cycle checkpoints and DNA repair.

d) Domain-based analyses of the DDR components

Domains are the functional units and building blocks of proteins. In this work we study the DDR proteins at a domain level to increase the functional inference, to analyze the conservation of domains in DDR orthologs from different species, and to determine the acquisition of novel functions due to diverse domain architectures reflecting differences at the species level. Also, we intend and to identify whether there are domains enriched in DDR-related functions.

3 METHODS

3.1 OVERVIEW

In this work we have:

1. Manually compiled from literature DNA repair and DDR components in four model organisms: *Escherichia coli*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Homo sapiens.*
2. Searched for DDR orthologs in 43 additional species covering the whole species tree used in this study.
3. Studied the evolution of the DDR based on gene-content and gene-trees.
4. Analyzed the functional repertoire of these genes using different classification systems.
5. Identified evolutionary time-points with potential for establishment of post-translational modifications activities.

An overview of this flow is depicted in Figure 8.

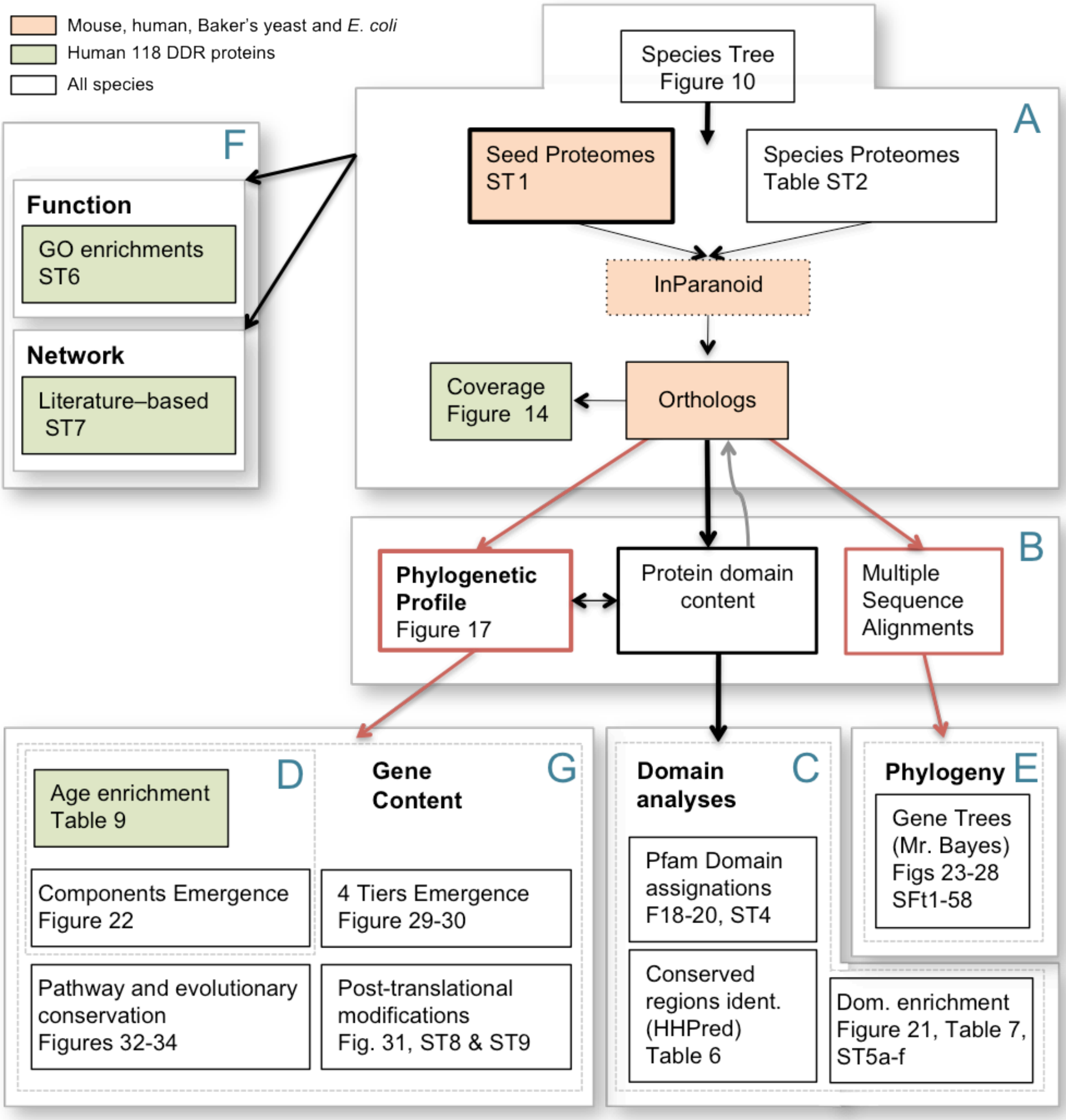


Figure 8. Flow chart depicting the main steps followed in this work and the data analysis performed (divided in letter panels A to G).

A collection of Perl scripts was written to, from the InParanoid output, automatically generate the phylogenetic profiles, determine the protein domain content of the orthologous sequences and to generate the MSAs.

3.2 DATASETS AND GENOME SOURCES (Figure 8, Panel A)

- Literature mining:

We have manually checked the literature to compile a comprehensive list of DDR components in four model organisms where DNA repair and DDR have been extensively studied. The four seed data and the number of proteins selected from literature are the following (see Box 2):

Species	DDR proteins selected	Code
H. sapiens	118	Hsa-118
E. coli	46	Eco-46
A. thaliana	122	Ath-122
S. cerevisiae	91	Sce-91

Box2: Seed species and number of DDR proteins selected from literature.

Information about the DDR proteins selected from literature for each of the four seed organisms can be found in Table ST1 Annex.

- Genomes:

To trace the presence of DDR orthologs along evolution, sequence datasets for 47 proteomes were downloaded from the available databases (for a full list of organisms, data sources and characteristics such as completeness, quality, etc., see Table ST2 Annex), which represent complete and incomplete proteomes and include both predicted and confirmed peptide sequences. When a particular proteome was available from different databases, the coverage was compared (Figure 13, Results 4.3) and the version containing the highest number of human DDR orthologs was chosen.
These datasets include 8 eubacteria, 3 archaea and 36 eukaryotes ranging from *Cryptophyta* to mammals.

- Species tree:

Organisms were grouped on the basis of previously defined phylogenetic studies [266,267] (see Figure 10 for the species phylogenetic tree) always accounting for the polytomy at the Eukarya tree [268]. The species divergence time values have been extracted from http://www.treetime.org [265] where consensus estimates from literature have been used due to the fact that no expert dates are available for all the species.

3.2.1 Sequence selection

For the eukaryotic seed DDR proteins we selected the Uniprot's definition of 'canonical peptide sequence' [269]. To reduce redundancy, the UniProtKB/ Swiss-Prot describes in a single entry all the protein products encoded by one gene in a given species. For each entry, a 'canonical sequence' is chosen based on the following criteria: i) it is the most prevalent, ii) it is the most similar to orthologous sequences found in other species, iii) due to its length or amino acid composition, it allows the clearest description of domains, isoforms, polymorphisms, post-translational modifications, etc. and iv) in the absence of any information, the longest sequence is chosen.

3.2.2 Selection of organisms

Deep branching organisms from the three kingdoms of life were selected to obtain the widest variability of *phyla*. Nevertheless, this was not always possible since many *phyla* lack completely sequenced representatives. Also, model organisms have been preferentially chosen, as well as those whose genomes are completely sequenced and well characterized.

To incorporate information regarding different lifestyles we have included free-living organisms, as well as endosymbionts (3) and parasites/pathogens (8) in the tree. Besides, species inhabiting diverse environments (aquatic, terrestrial, extreme ecosystems, etc.) and with distinct types of metabolism (heterotrophic or autotrophic) have been selected.

Here follows a brief summary of the characteristics of those species selected in this study that have interesting features or distinctive lifestyles:

a) Bacteria worthy of note

- *Gemmata obscuriglobus* and *Pirellula staleyi* (Taxonomy: *Planctomycetes*; *Planctomycetacia; Planctomycetales*) [270,271]

The Planctomycetales order forms an independent, monophyletic phylum of the Bacteria kingdom. It consists of four genera: Planctomyces, Gemmata, Isosphaera, and Pirellula. These organisms have a life cycle consisting of a motile swarmer stage and an aggregate-forming sessile stage. They are quite abundant in terrestrial freshwater environments and marine habitats, where they catalyze essential transformations in global carbon and nitrogen cycles.

Both bacteria present unique combinations of morphological and structural properties, such as budding replication, a lack of peptidoglycan in their cell wall and a membrane-bound DNA-containing nucleoid resembling the eukaryotic nucleus, for which these species represent an exception of prokaryote/eukaryote dichotomy and are thus important in the understanding of the evolutionary implications of compartmentalization on major molecular processes in the cell [272].

- *Deinococcus radiodurans* (*Deinococcus-Thermus; Deinococci; Deinococcales*)
This organism can tolerate high levels of chemical, oxidative, UV, and ionizing radiation-induced damage to the cell's DNA, which it efficiently repairs. The resistance to radiation may reflect its resistance to desiccation [273] and oxidative stress [274], which also causes DNA damage. This organism carries multiple copies of many DNA repair genes, suggesting a robust system for dealing with DNA damage [275].

- *Bacillus subtilis* (*Firmicutes; Bacillales; Bacillaceae*)
It is one of the better-characterized bacterial organisms and is a model system for cell differentiation and development. This soil bacterium can divide asymmetrically, producing an endospore at times of nutritional stress that is resistant to environmental factors such as heat, acid, and salt, and which can persist in the environment for long periods of time. The sporulation process is complex and involves the coordinated regulation of hundreds of genes [276].

b) The endosymbionts

- *Buchnera aphidicola* (*Proteobacteria; Gammaproteobacteria; Enterobacteriales*)
This bacterium is the primary endosymbiont of the pea aphid, *Acyrthosiphon pisum*. Almost all aphids contain maternally transmitted bacteriocyte cells that house the *Buchnera*, which provide the essential nutrients the host lacks [277]. Due to the symbiotic relationship with aphids for millions of years, *Buchnera* have lost the genes to produce lipopolysaccharides for the outer membrane, and those required for anaerobic respiration, synthesis of amino sugars, fatty acids, phospholipids, and complex carbohydrates. This makes for an obligate endosymbiont relationship between host and *Buchnera*, and has also resulted in one of the smallest and most genetically stable known genomes of any living organism [278].
Buchnera is believed to have had a free-living ancestor similar to modern Enterobacteriaceae, such as *E. coli*. Thus, comparative genomic studies can shed light on the evolutionary mechanisms of intracellular endosymbiosis as well as the different underlying molecular basis between organisms with parasitic behaviour and symbionts.

- *Guillardia theta* (*Cryptophyta; Pyrenomonadales; Geminigeraceae*)
This organism is a flagellate, unicellular alga that consists of a flagellate host cell, complete with mitochondria and nucleus, surrounding another cell with reduced cytoplasm that contains a plastid and a residual nucleus called nucleomorph, which were acquired through secondary endosymbiosis by engulfing and retaining a red alga [279]. The *G. theta* nucleomorph consists of 3 chromosomes with a genome size of 551 kilobases [280]

- *Bigelowiella natans* (*Rhizaria; Cercozoa; Chlorarachniophyceae*)
As *G. theta*, this amoeboflagellate cercozoan obtained its chloroplast by engulfing a photosynthetic eukaryote by secondary endosymbiosis. The host cell also contains the nucleus and the cytoplasm of the engulfed alga though in reduced form. *B. natans* is useful in studying evolution of the chloroplast and is used for comparative analyses [281].

The nuclear genomes of the cryptophyte *G. theta* and the chlorarachniophyte *B. natans* have been recently sequenced [282]. Both genomes have more than 21,000 protein genes, but in this study we only include the proteome of the nucleomorphs, since these would represent the more ancient genomic contributor.

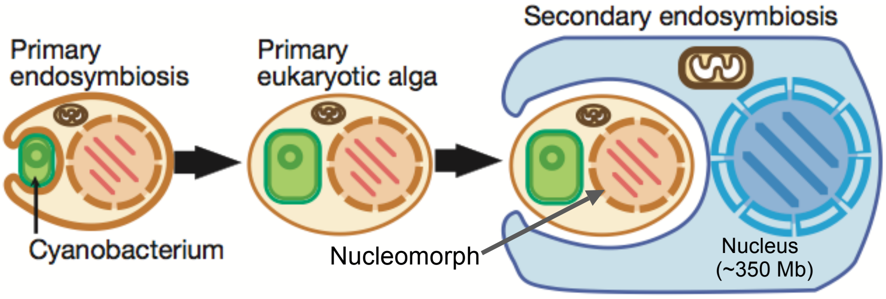


Figure 9. Secondary endosymbiosis process in Cryptomonads and chlorarachniophytes (Modified from Douglas S. *et al.* 2001). These eukaryotes acquired plastids by secondary endosymbiosis whereby a eukaryotic phagotroph engulfed and retained another plastid-containing eukaryote that was a descendant of the primary endosymbiotic event.

Because of their reduced size and cell simplification, the minute nucleomorphs make an important model system to study genome and cell function and help in the understanding of the more complex chromosomes of typical nuclei.

c) The parasites/pathogens

- *Mycoplasma genitalium* (Bacteria; *Tenericutes; Mollicutes; Mycoplasmataceae*)

This small parasitic bacterium lives on the ciliated epithelial cells of the primate genital and respiratory tracts. Infection proceeds through attachment of the bacteria to the host cell via adhesins and subsequent invasion, which can result in prolonged intracellular persistence that may cause lethality.

The genome of *M. genitalium* consists of 521 genes (482 protein encoding genes), being the free-living bacteria with the smallest known genome [283], and thus was the organism of choice in The Minimal Genome Project to find the smallest set of genetic material necessary to sustain life.

Along evolution, mycoplasmas seem to have lost many genes involved in metabolism and biosynthesis, resulting in the requirement of a full spectrum of substrates and factors taken up from the host. Besides, they lack a number of genes involved in cell division, heat shock response, regulatory genes, the two-component signal transduction systems and most transcription factors [284].

- *Batrachochytrium dendrobatidis* (Fungi; *Chytridiomycota; Chytridiomycetes*)

This recently emerging pathogen, the first sequenced representative of the *Chytridiomycota* phylum, is considered as a primary causative agent of amphibian declines in populations all over the world. This fungus contains proteolytic enzymes and esterases that help it digest amphibian cells and use amphibian skin as a nutrient source. *B. dendrobatidis* zoospores invade the top layers of the skin cells of the host, forms a cyst underneath the surface, initiates the reproductive portion of its life cycle and causes thickening of the keratinized layer [285]. The amphibians infected with these zoospores are shown to die from cardiac arrest.

- *Encephalitozoon cuniculi* (Fungi; *Microsporidia; Unikaryonidae*)

This single-celled obligate intracellular parasite infects various mammals, among them humans, where it causes a variety of conditions affecting the nervous system and respiratory and digestive tracts.

Microsporidia are atypical fungi that are thought to have lost mitochondria during evolution. *E. cuniculi* has one of the smallest known eukaryotic genomes (2.9 Mb), which is organized in 11 chromosomes and has approximately 2,000 predicted protein-encoding genes [286]

- *Cryptococcus neoformans* (Fungi; *Dikarya*; *Basidiomycota*)

This facultative intracellular pathogen [287] is found worldwide, and frequently in soil contaminated by bird excrement. *C. neoformans* are the causal agents of most human and animal cryptococcosis, which is acquired via inhalation of haploid yeast or basidiospores from the environment. These infections usually occur in immunocompromised hosts and mainly consist of a lung infection, though fungal meningitis and encephalitis, especially as a secondary infection for AIDS patients, are often caused by this fungus, making it particularly dangerous [288].

- *Cryptosporidium parvum* (*Alveolata*; *Apicomplexa; Coccidia*)

This protozoal species is an obligate intracellular parasite that has a complex life cycle, with multiple asexual and sexual developmental stages. It is a causal agent of cryptosporidiosis, a parasitic disease of the intestinal tract in mammalians, which consists of acute diarrhea. Infection is caused by ingestion of sporulated oocysts transmitted by the faecal-oral route. This alveolate has emerged as a very important pathogen worldwide due to its morbidity and mortatility in AIDS patients.

C. parvum has a compact genome and is one of the few organisms without transposable elements. Unlike other apicomplexans, it has no genes in its plastids and possesses a degenerate mitochondrion that has lost its genome [289].

- *Plasmodium falciparum* (*Alveolata; Apicomplexa; Aconoidasida*).

This protozoan parasite is the causal agent of human malaria. This parasite has a very complex life cycle, involving vertebrate and invertebrate hosts. Infective forms (sporozoites) are transmitted to the human host by the female Anopheles mosquito. The disease is caused by those parasite stages that multiply asexually in red blood cells [290]. Malaria is a devastating parasitic disease that infects 300 million people and kills up to three million people per year.

The entire genome of this organism has over 5,300 genes described [291].

- *Trypanosoma brucei* (*Euglenozoa; Kinetoplastida; Trypanosomatidae*)

This ubiquitous unicellular flagellated protozoan is the causal agent of African sleeping sickness that is transmitted by the tse-tse fly. African trypanosomiasis is a zoonosis and both cattle and wild game can act as reservoirs of human infective trypanosomes. Due to the large difference between its hosts the trypanosome undergoes complex changes during its life cycle to facilitate its survival in the insect gut and the mammalian bloodstream.

The incidence of this disease in humans is considerable (up to 500,000 cases per year) and in most cases is fatal if left untreated.

T. brucei has several large chromosomes that contain most genes, while the small chromosomes it possesses carry genes involved in antigenic variation [292].

- *Schistosoma japonicum* (*Metazoa; Platyhelminthes; Trematoda*)
This organism is a parasitic flatworm with a complex life cycle with various differentiated stages. It causes human schistosomiasis, which affects approximately 210 million people in 76 countries, is a cause of serious morbidity and is estimated to account for more than 250 thousand deaths per year, mainly in China and the Philippines. Besides, it also infects at least 30 species of mammals.
The *S. japonicum* genome consists of 7 autosomes and 2 sex chromosomes. Its genome is 397Mb in size and encodes at least 13,469 genes [293].

3.3 DISTRIBUTION OF DDR COMPONENTS IN AGE GROUPS (Figure 8, Panel A)

To establish important evolutionary points it is necessary to define groups of genes that share the same “age”. In this work, “age” indicates in what precise evolutionary point of a given species tree a gene is present. Following this scheme, we have defined 11 age groups in the 47 species tree ranging from proteobacteria to human. Thus, from more ancient to more modern age-groups, group 1 includes homologs present in the main three supra-kingdoms (along the 47 proteomes); group 2 contains genes present in most eukaryotes (except those organisms with particular life-styles, see Discussion), but absent in prokaryotes*;* group 3 includes proteins found conserved from plants; group 4 includes one *Unikonta* (*Amoebozoa)* representative; group 5 points to conservation in the *Opisthokonta* (*Fungi* and *Metazoa*) split; group 6 from *Metazoa* (being *Placozoa*, the most primitive animals in our study); while group 7 from *Radiata* includes one cnidarian species to represent different body plan symmetry; group 8 from *Bilateria*; group 9 includes *Chordata*, group 10 includes *Vertebrata;* and finally group 11 contains the mammalians (Figure 10, red boxes). This distribution is similar to previous classifications in ages [227]; therefore it is amenable to conduct comparative analyses.

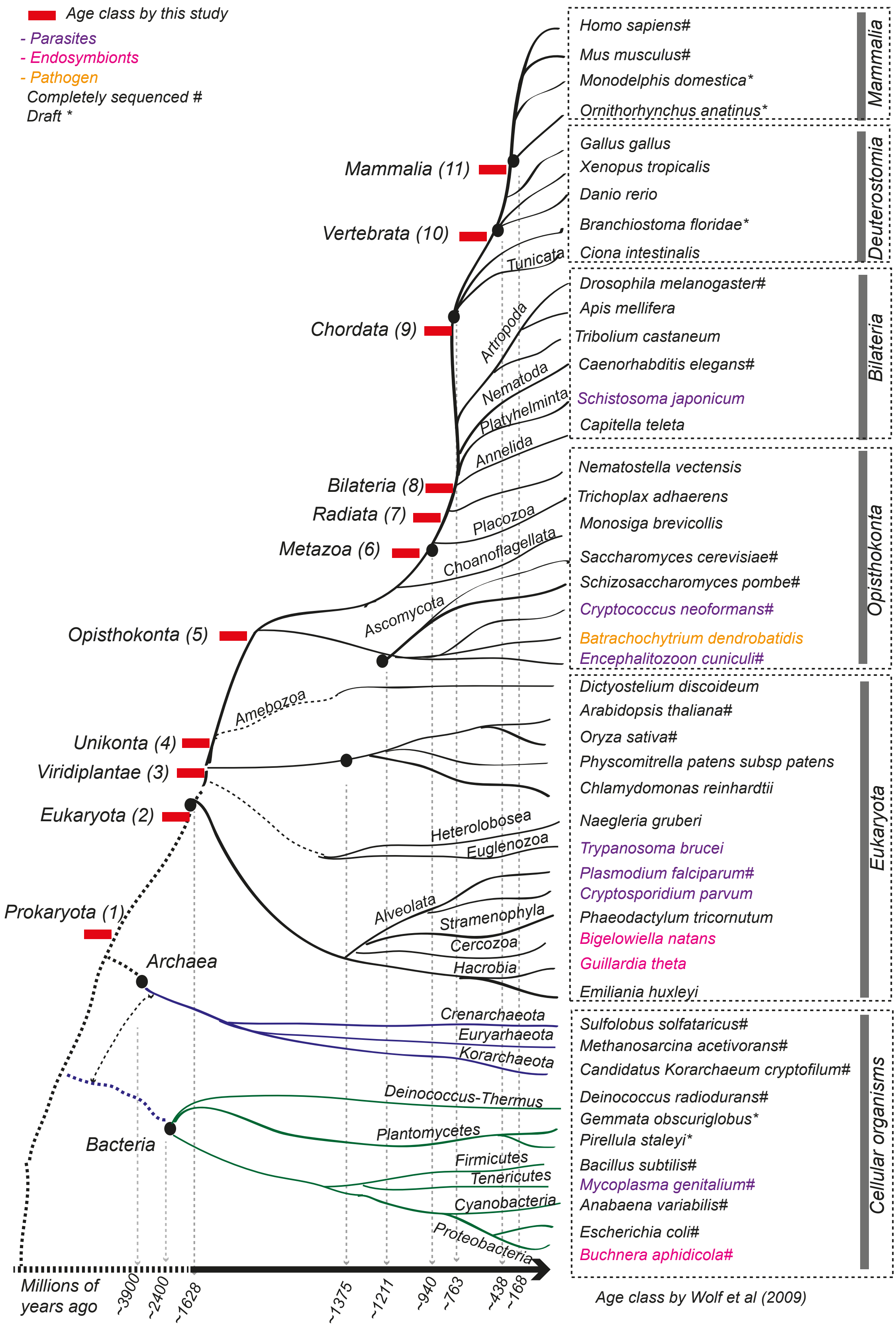


Figure 10. Species tree and age-groups. The red boxes show the age-groups in this study. On the right, the 47 species have been divided in wider phylogenetic groups. The dashed boxes are ages used elsewhere (Wolf *et al.* [227]).

3.4 IDENTIFICATION OF ORTHOLOGS (Figure 8, Panel A)

Orthologs are defined as genes from different species that derive from a single gene in the last common ancestor of those species, while in-paralogs are genes that derive from a single gene that was duplicated within a genome after the speciation event [189]. We developed a computational pipeline (Figure 8) to systematically identify orthologs using Inparanoid [294], an automatic method that uses pair-wise similarity scores, calculated using NCBI-Blast [165], between two proteomes for constructing orthology clusters. These clusters are seeded with a two-way best pair-wise match of orthologous sequences, after which an algorithm for adding in-paralogs is applied. The basic assumption is that sequences from the same species that are more similar to the main ortholog than to any sequence from other species are in-paralogs belonging to the same group of orthologs. Each member of the cluster receives an in-paralog score, which reflects the relative distance to the seed in-paralog. The confidence that the original seed-ortholog pair contains true orthologs is estimated by sampling how often the pair is found as reciprocally best matches by a bootstrapping procedure. Bootstrap values are generated by counting how many times the seed-pair of genes are each others best match in a sampling with replacement procedure that is applied to the original Blast alignment.

Each of the four seed datasets (*H. sapiens, E. coli, A. thaliana and S. cerevisiae*) were used as a query list to find DDR orthologs in the 47 proteomes included in this study (ST2 Annex), which cover the whole tree of life.

We run InParanoid using the default and also modified parameters to:

a) Avoid obtaining too many in-paralogs with very low similarity to the main ortholog in distantly related organisms. To this purpose we set:

$conf_cutoff = 0.25 (raising it from the default value of 0.05).

b) Obtain hits that share common domains in sequences that have non-conserved regions:

$segment_coverage_cutoff = 0.2 (instead of the original 0.25)

Then the matching segments must cover at least 20% of the longer sequence, but always forcing the total matched area to be longer than 40% of the longer sequence:

$seq_overlap_cutoff = 0.4 (lowered from the default 0.5).

This should avoid clustering sequences that share only short domains.

Regarding the matrices, in all cases BLOSUM45 was used to compare prokaryotes, BLOSUM62 when comparing eukaryotic proteomes, and BLOSUM80 for orthologs within metazoa. All InParanoid blastalls were run with -e 0.01 to set the threshold e-value to 0.01.

We compared both strategies (Results 4.3, Figure 14). Besides, in some particular cases we manually checked the hits to increase the number of potential orthologs.

3.4.1 The human dataset: Expanding the repertoire of DDR proteins

We expanded the repertoire of DDR proteins for the 47 species and collected the orthologous sequences for the Hsa-118 set detected using the four different seeds (henceforth "**Ortho-DDR**"); in this way we could alleviate the effects of using a specific seed.

3.4.2 Multiple sequence alignments (Figure 8, Panel B)

The orthologous proteins were aligned using T-COFFEE [209] and MAFFT [208] to manually confirm the quality of the relationships, and the sequence coverage of the Hsa-118 orthologs was checked (Results 4.3.2 Figure 16).

3.4.3 Evolutionary conservation of orthologs

To assess if the sets of proteins involved in DDR are more evolutionarily conserved than other groups of proteins in the selected species, random sets of 100 proteins were sampled from both the Hsa-118 DDR set and *H. sapiens* proteome (excluding the 118 DDR proteins). Then their orthologs were identified in all the species and the results were compared using z-scores.

Using Z-scores to assess the significance of random distributions:
The statistical significance of the results can be measured by estimating the z-score for the different sets of proteins.
Each species has a collection of orthologs values in the different sets of 100 sampled proteins. If the distance values for all the sets show a normal distribution, the mean (*μ*) and standard deviation (*σ*) can be calculated for each set of orthologous proteins from each species, and z-scores can be calculated for each ortholog value (*x*) within each set: $z = \frac{x - \mu}{\sigma}$. *Z* is a normalized parameter that can be used for comparing different pairs of sets and their distributions.

3.5 PROTEIN DOMAIN IDENTIFICATION (Figure 8, Panel C)

In order to establish the domain repertoire of our target proteins and analyze its domain organization we have used the Pfam database (release 24.0) [174]. Pfam is a large and widely used database of protein domains and families. It contains curated multiple sequence alignments for each family, as well as associated HMMs for finding these domains in new sequences.
We have used the Pfam database since it has a wider coverage than other domain databases (Pfam v24.0 contains a total of 11,912 domain families versus 809 in the SMART database v6.0 or 5,608 in InterPro v28.0).
To identify domains we have used the improved version of HMMER (HMMER 3.0) [172,295] that uses profiles derived from high quality multiple sequence alignments taking into account structural features. These profiles are models representing the statistical

formalization of the multiple sequence alignments, which are useful to determine domain boundaries.

The orthologous sequences obtained were checked for consistency in their domain architectures (Figure 18 and 19 in Results 4.4.2, and ST4 in Annex), and some additional proteins found using alternative seeds with no clear annotations to DDR were identified (ST3, Annex).

3.5.1 *In-silico* determination of domain architectures in DDR proteins

Some DDR proteins have different isoforms, which may result in variations in domain composition of the orthologous proteins in the different species analyzed. Though there are efforts to establish the different repertoire of the protein isoforms and splice variants in a variety of species (like the ENCODE Project in human) [296], most of this information is available only for some model organisms and there are little data for a systematic analysis. As these data are incomplete or are not available for most of the species we have analyzed, we have built synthetic proteins for each set of DDR orthologs to try to represent the widest possible domain composition for a given gene and to facilitate further comparisons.

In-silico synthetic architecture domains were computed for each analyzed DDR protein in the 47 proteomes. The synthetic proteins were built to collapse all possible domain architectures in each group of DDR orthologous proteins (see example figure 11).

Figure 11: *In-silico* domain architecture, BLM protein example.

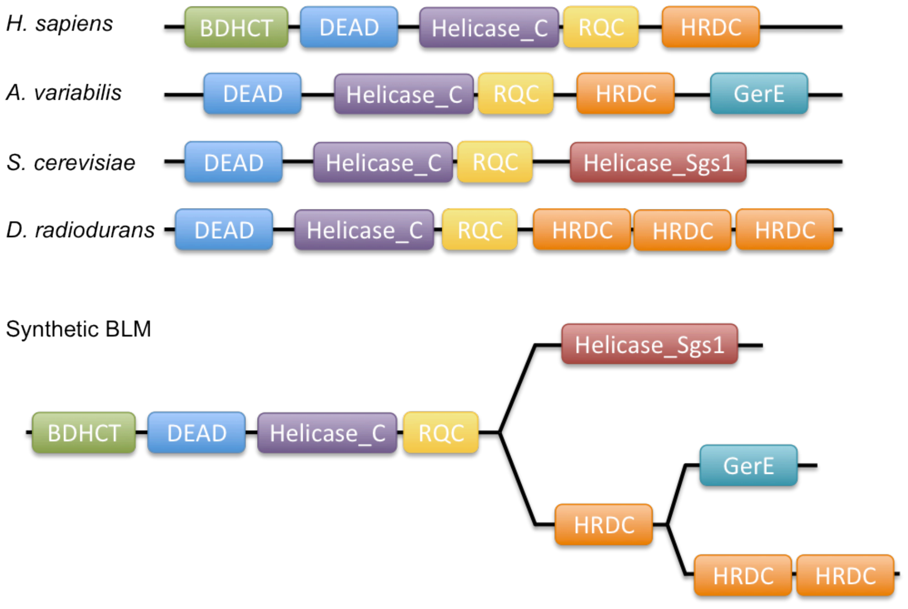


Figure 11. Example of the construction of a synthetic domain architecture protein. The domains detected in the BLM orthologs in the different species are included in the synthetic protein. In this particular case, the Helicase_Sgs1 and the HRDC domains are excluding, and may have an equivalent role since they are homologous domains belonging to the same HRDC-like domain clan (CL0426).

3.5.2 Conserved regions in proteins without detected Pfam domains

For those DDR proteins in which no Pfam domains were identified, we conducted extensive searches looking for distant relatives based on conserved regions. In this regard, multiple sequence alignments of orthologs were used as input for an in-house pipeline that uses HHPred [179] to detect and annotate the conserved regions in the sequences. The HHPred program is part of the open-source HH-suite software package, and it is mainly used for homology detection and structure prediction by HMM-HMM comparison.

3.5.3 Protein domain enrichment

The domain composition of the proteomes of *H. sapiens* and some model organisms (*E. coli, A. thaliana, S. cerevisiae, S. pombe, C. elegans, D. melanogaster*) belonging to different phylogenetic groups was analyzed, and later compared to the results obtained for the DDR proteins to determine whether the DDR proteins are enriched in given domains.
Even though they are completely sequenced model organisms, some proteomes contain repeated sequences, protein fragments, cDNA sequences and different isoforms; consequently, the proteomes were filtered to discard these sequences and when there were proteins with different isoforms, only the canonical ones were considered.

To do this analysis we compared the domain composition of the DDR proteins (including the seed proteins extracted from bibliography plus those orthologs detected when using other seed species) and the domains included in the model species' filtered proteomes (Results 4.4.5, Box3). For the statistics, we performed a Fisher's exact test and a multiple testing Bonferroni correction to normalize the data.

3.6 PHYLOGENETIC PROFILES OF PROTEINS AND DOMAINS (Figure 8, Panels B/D)

We constructed phylogenetic profiles [148] of the **Ortho-DDR** proteins identified. The profiles can be formalized as binary matrices of presence/absence of identified orthologs. These profiles are based on the assumption that proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion and tend to be coincidently present or absent within different genomes. Consequently, phylogenetic profiles can be used to delineate the correlated evolution of proteins [148].

To analyze the profiles in the context of evolution, we used as a reference the given phylogeny (Section 3.3, Figure 10) described previously and conducted different analyses:

3.6.1 Clustering of protein and domain profiles (Figure 8, Panel B)

Hierarchical clustering of the phylogenetic profiles was done using the open source software Cluster 3.0 [297]. Euclidean distance was used for the similarity metric with average linkage as the clustering method, which is suitable to cluster our data [298] and has been successfully used in previous studies [297,299].
Hierarchical methods are useful for representing protein sequence family relationships. These clustering algorithms partition the objects into a tree of nodes, where each node represents a cluster. Linkage is the criterion by which the clustering algorithm determines distance between two clusters. *Average linkage* takes the mean distance between all pairs of objects of two clusters, which makes it more computationally expensive than other methods, but is the most robust linkage method since it avoids the chaining problem of *single linkage* (which forces clusters together due to single objects being close to each other) and does not give special weight to outliers as *complete linkage*.
Finally, the clustering is illustrated by appending a tree showing sequence relations, with branch lengths reflecting profile similarity. We used Java Tree View to visualize the trees of clusters [300].

The same method was used to cluster the domains phylogenetic profiles in the **Ortho-DDR** dataset of proteins.

3.6.2 Gene-content (Figure 8, Panel D)

We have established 11 age groups to reflect important evolutionary points given our phylogenetic tree (Section 3.3, Figure 10, red rectangles). To analyze the evolution of gene content in the species in our study, we used the Count package [301] that contains different algorithms.
In particular, we used Wagner and Dollo parsimony to analyze the profiles of the **Ortho-DDR** dataset. Both algorithms aim to reconstruct the evolutionary history of proteins from given phylogenetic profiles.

- The Dollo parsimony assumes a single event of emergence per family (because gaining a gene is more rare than losing it), and the presence-absence pattern is explained by lineage-specific losses. This method leads to a considerable simplification of evolutionary analysis and provides for unambiguous reconstruction of evolutionary scenarios.

- The Wagner parsimony allows multiple gain and loss events and assumes that all character states are reversible with similar rates of transitions. It penalizes the loss and gain of individual family members, and infers the history with the minimum penalty.

These assumptions are strongly influenced by the quality of the profiles in the sense of true losses (see Discussion).

3.7 ENRICHMENT ANALYSES FOR GENE AGES (Figure 8, Panel D)

As aforementioned, the genes present in a given species emerged at a variety of evolutionary times, and several works have suggested that the context of a gene's origin can provide information about its cellular functions, regulation, interactions of the encoded proteins and adaptability [148].

To determine whether the *H. sapiens* dataset (*Hsa*-118) is enriched in certain ages, we have used *ProteinHistorian* [302] and calculated enrichments using five different methods (Jaccard, Multiparanoid, Panther7, OthoMCL and Naive Ensemble (Nens)) and two different ancestral family reconstruction algorithms (Wagner and Dollo parsimony) to account for expected differences according to different phylogenies and datasets. In all cases p-values were Bonferroni corrected (Results 4.6, Table 9).

Given an input set of proteins of interest, its phylogenetic distribution can be compared to that of a relevant background set. As different definitions of protein "age" may suit distinct contexts, diverse strategies for estimating ages from phylogenetic patterns using databases of evolutionary relationships can be utilized. However, age estimations are for eukaryotic proteins only, since prokaryotic proteins may have been affected by HGT events, which would complicate the inference of the evolutionary tree in prokaryotes.

ProteinHistorian makes use of several sets of protein family predictions from the Princeton Protein Orthology Database (PPOD) [303]. PPOD provides family predictions for all proteins in the genomes of the GO Reference Genome Project [304], which are made with MultiParanoid [305], OrthoMCL [205], Nens clustering-based consensus of the MultiParanoid and OrthoMCL predictions, and PPOD's own Jaccard clustering-based approach.

Regarding the pre-computed databases, the OrthoMCL, MultiParanoid and Nens contain families of predicted orthologs, while the Jaccard clustering produces larger families of more distantly related protein sequences. The Panther database is based on an OrthoMCL clustering of all proteins in the 48 species present in v7.0 of the PANTHER classification system. Ages are provided for all proteins in the eukaryotic species (32 in the present day) included in the PANTHER database [306].

3.8 PHYLOGENETIC TREES ANALYSIS (Figure 8, Panel E)

We built a phylogenetic pipeline where multiple alignments of the orthologous sequences were used as input for probabilistic-based phylogeny [216]. The probabilistic phylogeny was run using the MPI implementation of Mr. Bayes [217]. Alignments were manually checked to identify potentially conflictive regions. Only proteins and families with representatives in ancient eukaryotes were analyzed (SF1, Annex). Each job was run in 8 independent Markov chains for more than twenty-five thousand hours in a cluster that consists of 64 cores with an average of 8Gb/core of ram each. Every tree

was sampled for at least 5 million Markov generations and we discarded 25% of the generated trees to ensure convergence was reached. In a first approach, in-paralogs and paralogs as defined by Ensembl COMPARA were included to assure the correct ortholog was selected. In a second approach, phylogenies were run only with the orthologous sequences.
A total of 65 gene trees were generated, which were visualized with iTOL [307] and further analyzed for consistency with the species tree.

3.9 FUNCTIONAL CLASSIFICATION OF GENES AND DOMAINS (Figure 8, Panel F)

a) GO assignment

Next, we used the available server (DAVID, http://david.abcc.ncifcrf.gov/) [308] to calculate functional enrichment using GO terms [309] in three main categories: *Biological process*, *Cellular component* and *Molecular function* (ST6, Annex).

b) A broader classification

Alternatively, we have used a broader classification of the genes and the domains contained in these into a 4 tiers classification: "*Effectors*", "*Sensors*", "*Transducers*" and "*Mediators*". It should be noted that the same gene could belong to more than one tier (ST7, Annex).
In this schema, sensors and effectors would represent the extremes of a given directed pathway, while alternative functions can be incorporated to increase the complexity of the network by addition of proteins (or functions) belonging to the remaining classes. Consequently, "Mediators" will usually form complexes to recruit additional proteins acting as docking platforms (i.e. containing for instance protein-protein interaction domains, or phospho-peptides binding domains), and "Transducers" would trigger alternative signaling pathways (i.e.: kinases involved in checkpoints) and would therefore create complicated crossing-roads connecting different pathways. On the other hand, this assignment does not preclude functional overlapping in proteins of DDR, where some proteins are involved in more than one repair pathway or might switch their functions. For instance, MSH6 senses damage and acts as a repairing protein in MMR, but also acts regulating Ku70 in the NHEJ pathway [38].

3.10 COMPONENTS INVOLVED IN PTMs (Figure 8, Panel G)

We established pairs of target-modifier in the human dataset (Hsa-118), where *targets* are proteins post-translationally modified in DDR events, and *modifiers* are those proteins from the same dataset performing the modification. The PTMs considered here are phosphorylation, sumoylation, ubiquitination, acetylation, de-ubiquitination and de-acetylation; for which precise experimentally confirmed information was found for a fraction of the DDR proteins (see Table 11; ST8, Annex). We next checked whether a given pair of interactors was conserved along the evolutionary scale.

3.11 MAPPING EVOLUTION INTO THE HUMAN NETWORK (Figure 8, Panel G)

The DDR network encompasses a variety of processes and signals, including repair, sensing and activation/resuming of cell cycle check points. Recent work has focused on elucidating the dynamic properties of the network via PTMs, and even when there is some consensus, still some of its dynamic properties remain largely unknown.

Most DDR components and dynamic processes have extensively been studied in human, and therefore we focus on this species. Subsequently, we have collected the available literature and we have classified the overall network in 3 sub-networks:

- **General damage repair** sub-network, which contains proteins involved in general repair pathways (BER, NER, NHEJ, etc).
- **Replicative stress** sub-network, which includes proteins involved in the sensing and repair of damage at the replication fork or SSBs.
- **Double Strand breaks** damage sub-network contains proteins involved in sensing the damage at DSBs when NHEJ fails, and ATM-based takes over.

It should be mentioned that some sub-networks as ICL, or meiosis-specific DDR components, etc. have not been included for the sake of broadness.

Illustrations with associated published references are depicted on figures 32-34, and extensive explanations regarding each step are available in sections 1.1.3 and 1.2.3 of the Introduction.

4 RESULTS

4.1 DDR COMPONENTS BY LITERATURE (Figure 8, Panel A)

We manually extracted from literature DDR components from *Homo sapiens* (118 proteins, *Hsa*-118), *Arabidopsis thaliana* (122 proteins, *Ath*-122), *Saccharomyces cerevisiae* (91 proteins, *Sce*-91), and *Escherichia coli* (46 proteins, *Eco*-46) (Table ST1 Annex), and we calculated the overlap among these literature-based four datasets (Figure 12A), which is very low (MLH1, MSH6, RAD51 and SMC1A) and constitutes the common core of literature-based DDR.

4.2 HOMOLOGY-BASED EXTENSION OF DDR COMPONENTS FROM LITERATURE IN SEED SPECIES (Figure 8, Panel A)

To alleviate the effect of biases due to publication trends or research interests focusing on specific organisms and/or particular pathways, we conducted a systematic screening for orthologous proteins for each seed dataset along the complete proteomes of our model organisms in a "four model *versus* four model" approach to identify potential related proteins. In this regard, we identified a common set of proteins with a potential role in DDR, as well as lineage-specific proteins (Figure 12 B).

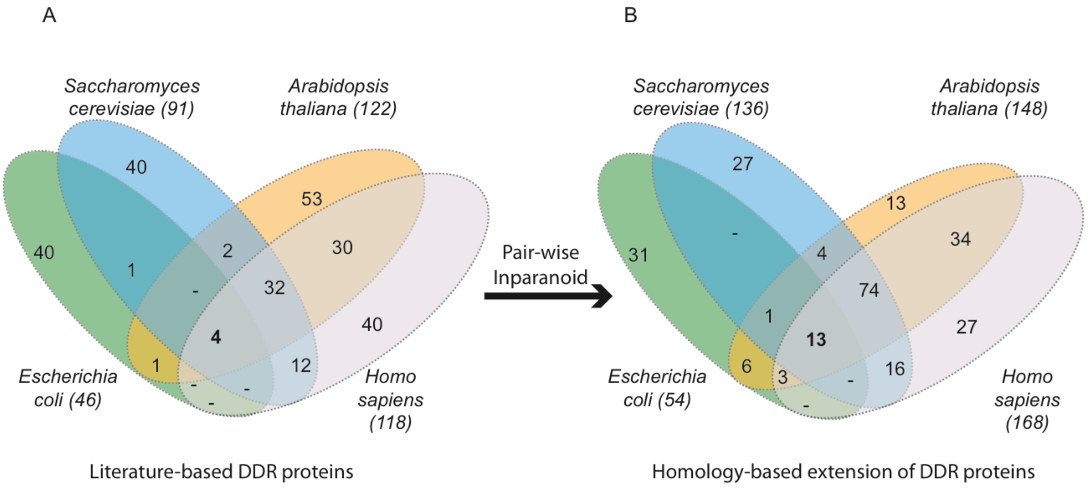


Figure 12. Overlap of DDR elements. Venn diagram showing the overlap among the literature-based DDR components from the four selected seed species (A), and the overlap among the same proteins plus the orthologs detected by InParanoid using the different seeds (B). The 13 proteins common to the four seed species are: BLM, CLPX, DPO2, DPO4, ERCC2, LON, MLH1, MSH3, MSH6, NTH, RAD51, SMC1A and UNG.

4.3 ORTHOLOGS IDENTIFICATION: EXTENSION OF THE PATHWAY USING HOMOLOGY-BASED INFORMATION IN 47 SPECIES (Figure 8, Panel A)

We selected the source proteomes according to the number of orthologs retrieved (Figure 13).

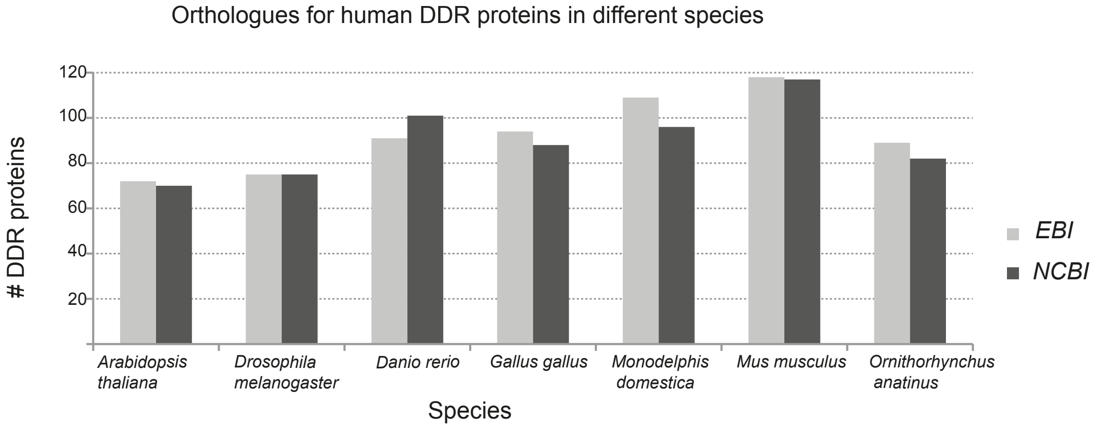


Figure 13. Comparison of the Hsa-118 DDR orthologs set detected in the proteomes downloaded from the EBI and NCBI of various species.

By modifying the InParanoid parameters (see Methods 3.4), we could detect a higher number of orthologs in the various species (Figure 14).

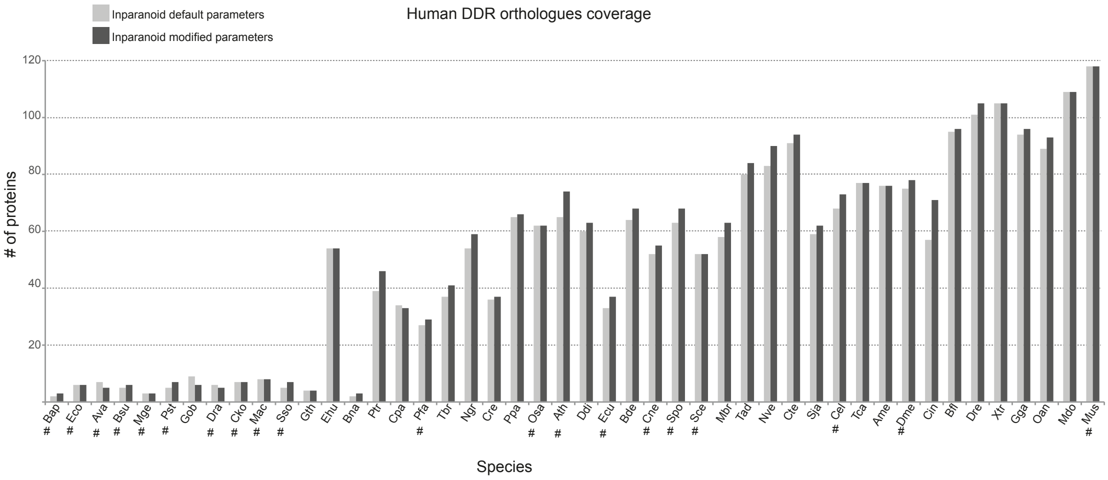


Figure 14. Human DDR orthologs coverage: orthologs detected in 46 species by InParanoid with the default and modified parameters. The # marks those species with completely sequenced genomes.

With these customized parameters, more than 2400 orthologous sequences were identified by InParanoid for the Hsa-118 set. These were manually checked and some best bidirectional hits and low confidence orthologs that had not passed the threshold were added when clear homology was detected manually. Also, we included hits that did not comply with the criteria settled up by means of length and score, but that were

BBHs, contained specific DDR domains found only in given proteins (i.e. RFA3, with the Rep_fac-A_3 domain) / shared the domain architecture of the corresponding orthologs, and aligned consistently in the MSA of these orthologous sequences. On the other hand, those orthologs detected by InParanoid that misaligned in the multiple alignments of orthologs and that lacked the characteristic domains of a given DDR protein were taken out from the sets. Table 3 below includes some examples of added and discarded orthologs when using the Hsa-118 set as seed:

Examples of added orthologs		Examples of discarded orthologs		
Proteins	Species	Proteins	Species	Reason for removal
1433E	Mbr, Cel	CHK1	Cko	1
ATM	Sce, Cel, Cin	ERCC2 (XPD)	Eco, Pst, Gob, Cko, Sso	1, 2
ATR	Osa, Ecu, Sja, Cel	EXO1	Ptr	1, 2
BLM	Sce	F175A (Abraxas)	Bde, Tad, Ame	1, 2
ERCC1	Bde	FANCM	Mac	1, 2
ERCC5	Nve, Dre	FBX31	Cre	1, 2
MLH1	Dra	HNRPK	Ath	1, 2
MMS21 (NSE2)	Oan	MDC1	Ehu, Bde, Cte	1, 2
NBN	Ppa, Osa	MDM2	Ehu, Ngr, Ath, Nve, Cte	1, 2
PALB2	Dre, Xtr	MDM4	Ehu, Ath	1, 2
PARP2	Mbr	MTA2	Mbr	1, 2
PLK1	Cne	PRKDC	Osa, Ath	1, 2, 3
RAD17	Mbr	RAD17	Ngr	1, 2
RAD50	Mac	RAD23B	Ame	1, 2
RAD9	Ehu, Ngr	RNF8	Ath	1, 2
RFA3	Ehu, Ptr, Cpa, Pfa, Tbr, Ddi, Ecu, Bde, Cne	SIR1	Bsu, Ppa	1
TOPB1	Mbr	TAOK1	Ehu	1
WEE1	Ddi	UIMC1 (Rap80)	Cte, Oan	1, 2

Table 3. Examples of added orthologs and discarded hits from InParanoid when using the Hsa-118 set as seed. Reasons for removal codes: (1) Inconsistent alignment in MSA of orthologs, (2) differences in domain architecture/only promiscuous domains shared, (3) experimental evidence of not being the corresponding ortholog (according to bibliography, the detected orthologs were actually mTOR).

As seen in Figure 14, the number of orthologous DDR proteins detected tends to increase as the organisms are more phylogenetically related to *H. sapiens*. Few orthologs were detected in parasites (*C. parvum, P. falciparum, E. cuniculi* and *S. japonicum*), endosymbionts (*B. aphidicola*) and especially in the small nucleomorph genomes of the secondary endosymbionts (*G. theta* and *B. natans*).

4.3.1 Evolutionary conservation of orthologs

As explained in Methods 3.4.3, to assess if the components of the DDR network are more evolutionarily conserved than other proteins, random groups of 100 proteins were sampled from the *Hsa*-118 DDR set and the *H. sapiens* proteome (excluding the 118 DDR proteins), their orthologs were identified in all the species used in this study, and the results were compared using z-scores.

Figure 15 shows, for the different species, the percentage of orthologous proteins involved in DDR and of those chosen randomly in the whole human proteome. The results indicate that in all organisms except prokaryotes (with the exception of the bacterium *M. genitalium*) and the nucleomorph proteomes of *G. theta* and *B. natans* (probably due to the reduced number of orthologs detected in all sets), there are significant differences (p-value < 0.01) regarding the number of orthologs detected in the sets of randomly chosen proteins and the set of DDR proteins. This suggests that the DDR components are more conserved along evolution than other proteins.

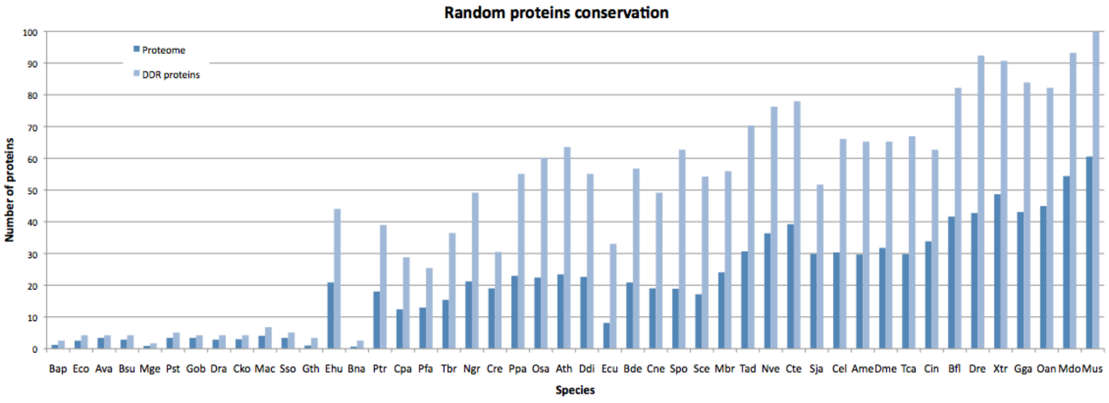


Figure 15. Random proteins conservation: orthologs detected by InParanoid in 46 species considering random sets of 100 proteins from the Hsa-118 set and the whole *H. sapiens* proteome.

4.3.2 Sequence coverage of the human orthologs

An analysis of the Hsa-118 DDR orthologs length coverage was performed to check the consistency of the orthology relationships and to analyze whether there is a trend in the sequences length distributions. The length of every sequence of the DDR orthologs (2538 sequences) was compared to the length of the corresponding human protein. Thus, the sequences were divided into four groups according to the coverage percentage, as seen in Figure 16.

The results show that 53% of the 2538 orthologs identified by InParanoid have a coverage equal or over 90%, and if we consider a coverage equal or over 80%, the sequences constitute nearly 70% of all the orthologs, which reinforce the reliability of the orthology predictions by InParanoid. Besides, if we take into account the phylogenetic distribution of the orthologs, we see that the phylogenetic groups (Prokaryotes, Ancient Eukaryotes, Plants, Fungi and *Metazoa*) are quite equally distributed among the different coverage percentage groups (except the prokaryotic sequences, which fall mainly in the ‘50 < % < 79.99’ group, likely due to the large phylogenetic distance between these organisms and human).

On the other hand, those orthologs detected by InParanoid with coverage under 50% constitute only 9% of the proteins, and correspond mainly to distant orthologs and to incomplete sequences in species whose proteomes are in early draft versions.

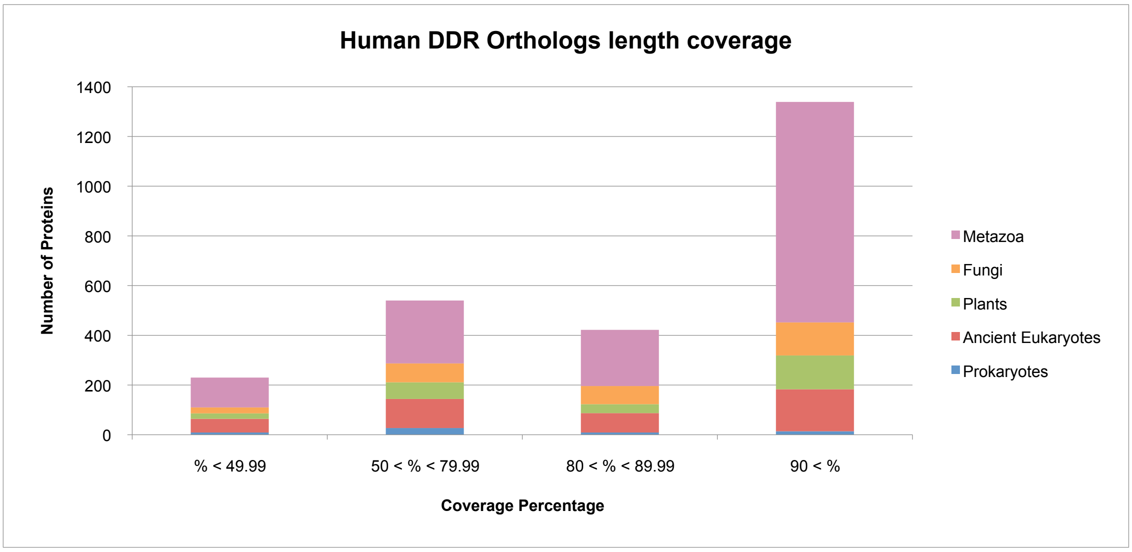


Figure 16. Length coverage of human DDR orthologs. The lengths of the orthologous sequences identified for each of the 118-Hsa proteins were compared and divided into four groups according to the coverage percentage referred to the length of the human proteins. Orthologs were coloured according to taxonomy of the species: *metazoa* in pink; fungi in orange; plants in green; ancient eukaryotes in red and prokaryotes in blue.

Regarding the sequences length distributions, a total of 308 orthologs were at least a 10% longer than the corresponding human protein. Among these, in 45 orthologs the sequence was at least 1.5 times longer than the human protein, being SLX1 the most numerous in this group since it was found in 5 species, followed by ERCC1 and MUS81 (each found in three organisms). These 45 orthologs were mostly from *D. discoideum* (7 proteins), followed by *C. neoformans*, *D. melanogaster, M. brevicollis* and *P. falciparum*, with 4 proteins each.

If we consider those orthologs with a sequence between 1.1 and 1.5 times longer than the corresponding human protein, MUS81, RBX1 and PLK1 are the most frequent, while the species where these longer proteins are found are mainly fungi, plants and again, *D. discoideum*, presenting 20 proteins longer than the human equivalent. The fact that this organism has a total of 27 proteins (out of the 65 DDR orthologs detected in this species) notably longer than their human counterparts could be because this genome is a draft assembly and the protein boundaries may have not been correctly predicted, or maybe it is due to this organism having longer genes than most of the other species.

On the other hand, 884 orthologs were at least a 10% shorter than the corresponding human protein. Of these, in 185 orthologs the sequence was below half the length of the human protein, being these proteins mainly in *O. anatinus* (16 proteins)*, N. vectensis* (14 proteins) and *C. intestinalis* (11 proteins), which is not surprising since the genomes of these three organisms are in the first draft versions and contain many incomplete and wrongly predicted proteins. Regarding the proteins, the most numerous among these 185 orthologs were ERCC3 (in 36 species), BLM (in 12) and FANCM (in 10 organisms). These three proteins are among the most ancient in the DDR network for they are found in prokaryotes. Besides, BLM has suffered important domain shuffling events, which may have contributed to the variations found in its length in the different organisms. Regarding the ATP-dependent helicases ERCC3 and FANCM, they both have the same domain composition (ResIII – Helicase_C) and are members of Family 17; the human ERCC3 sequence is notably longer than the ones found in most species, while FANCM seems to have increased its length along evolution, probably to be able to interact with an increasing number of proteins.

Although we use species that are distantly related in phylogenetic terms, and even though we have included a large number of incomplete proteomes in this study (being many of these in the first draft versions), most of the orthologs detected have a high percentage of coverage, which shows that our orthologs are quite reliable.

4.3.3 Expanding the search

Considering that using one species as seed would induce a bias towards an increased detection of orthologs in those species more phylogenetically related to the seed organism, and keeping in mind that there could be lineage specific domain insertions and/or losses in some orthologs, we decided to expand the original datasets by collapsing the four seed datasets and thus retrieve orthologs that might have skipped detection using only one species as seed. Therefore, for Hsa-118 we found 2453 orthologs, 498 for Eco-46, 2461 for Ath-122 and 1595 for Sce-91 (data not shown).

Besides, using four different seeds might allow us to detect proteins poorly described in one seed but well characterized in the other seed species. For instance, more than 50 orthologs of proteins not included in the original seed dataset of *Homo sapiens* (*Hsa*-118) because they are not published frequently in recent reviews, were detected in human after our searches using the DDR seeds from alternative organisms, suggesting that some of this hits could have a more important role in the network (Table ST3 Annex). Moreover, 13 proteins are common to the four DDR seeds (Figure 12b). From these, BLM, MLH1, MSH6, RAD51, SMC1A, TOP3A, present in the Hsa-118 seed, have well described functions in DDR, while other proteins with potential functions in human DDR or poorly characterized as DDR-related in human are: CPLX (mitochondrial ATPase), DPO2 (DNA polymerase delta, replication), DPO4 (DNA polymerase kappa, repair), LON protease (mitochondrial DNA replication), NTH (endonuclease III-like protein 1, oxidative damage and spontaneous mutagenic lesions), and UNG (Uracil DNA glycosylase, repairs misincorporation of dUMP residues by DNA polymerase).
For the 118 proteins involved in DDR in human, a total of 2656 orthologs were identified when gathering the orthologous sequences retrieved with the four different seeds, which overall constitutes the **Ortho-DDR** set.

4.3.4 Clustering of protein profiles (Figure 8, Panel B)

As a result of the clustering sorted according to the phylogeny (Figure 10), blocks of stable proteins were obtained (Figure 17).

The most conserved and widely distributed proteins among the organisms are the HR proteins RAD51, BLM and TOP3A, the post-replicative DNA mismatch repair proteins MSH3 and MLH1, and the NER helicase ERCC3 (Figure 17, lightest orange box, at the bottom).

While these proteins are homogeneously distributed along the three kingdoms of life, others appear to be specific to particular lineages. For example, PCNA (an auxiliary protein of DNA polymerase delta involved in the control of DNA replication) and RAD50 (MRN complex unit that binds to DNA ends at DSBs holding them in close proximity) are common to the *Archaea* and *Eukarya* kingdoms while SMC1A (involved in chromosome cohesion during cell cycle and in DNA repair) is detected in Bacteria and Eukaryotes. RFA (a ssDNA binding protein complex) is found only in eukaryotes, and other proteins like RNF168 (an E3 ubiquitin-protein ligase required for accumulation of repair proteins to sites of DNA damage) are specific of *Chordata.*

An important step in our evolutionary timeline is represented by the *Opisthokonta*, which points to the evolutionary split of animals and fungi. Generally speaking, the fungal species have incorporated novel lineage specific proteins and have suffered extensive gene losses [227]. In this regard, our analyses confirm this trend, where extensive losses (blue boxes in Figure 17) are found for important proteins, with only 65 orthologs detected in *S. cerevisiae* and 75 orthologs in *S. pombe*, out of the 118 human proteins studied here. Overall, *S. pombe* contains more identified orthologs than *S. cerevisiae*.

It should be noted that the use of parasites and pathogens with small genomes, as well as the simplified nucleomorph genomes of *G. theta* and *B. natans*, has expectedly produced large absences in the taxonomic distribution of some proteins in the cluster, most likely due to their particular biological requirements (parasitic lifestyle, etc). Also, there are few scattered dots that might be caused by noise in the experiment due to false positives in the detection of orthologs, and some missing proteins that will surely be false negatives, especially in *O. anatinus* and *M. domestica*, due to the incompleteness of their proteomes.

See section 4.5.2 a) for more explanations on the results.

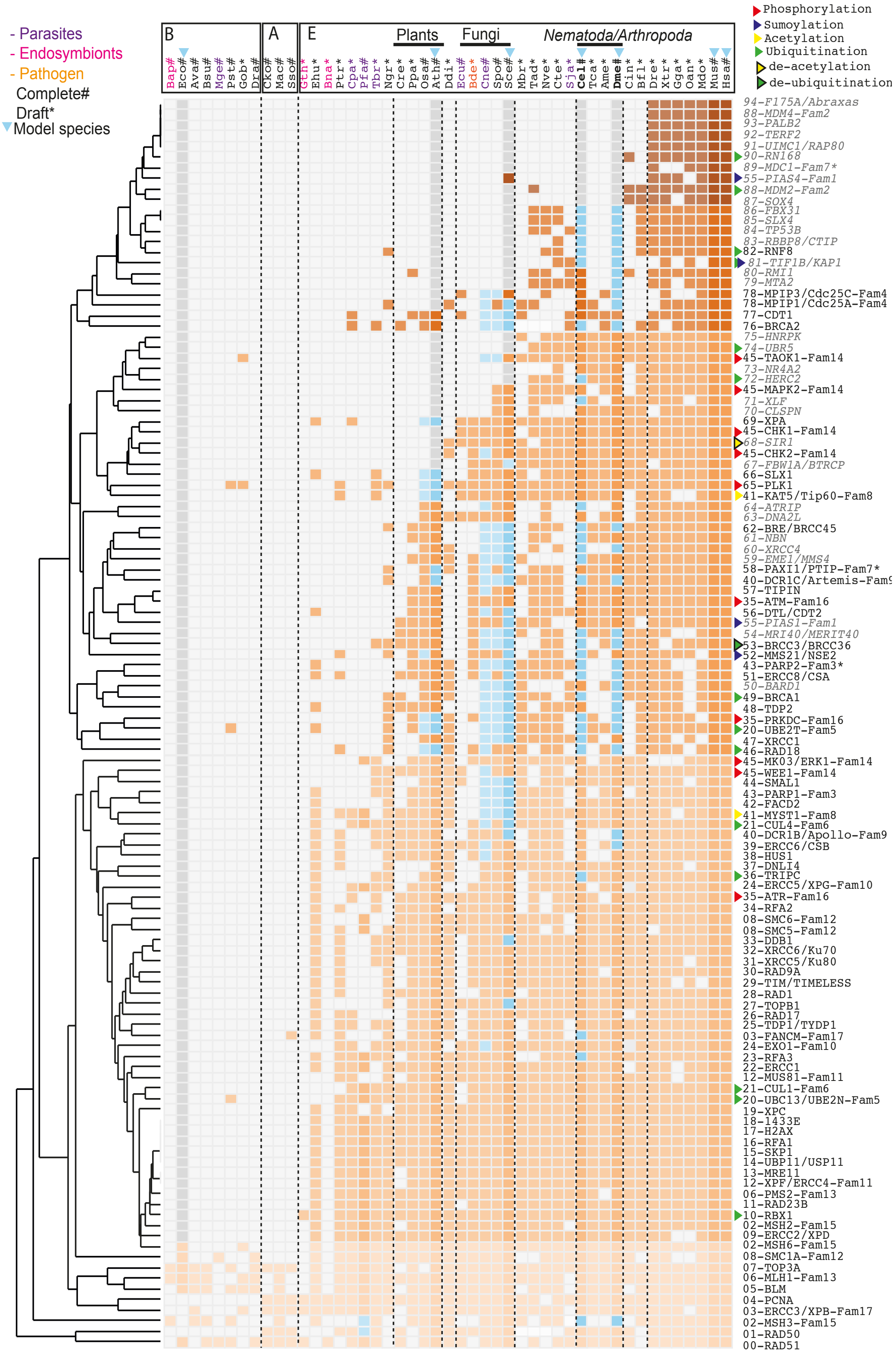


Figure 17. Cluster of the 118 Human DDR proteins orthologs according to the phylogeny of the 47 species analyzed in this study ordered by the evolutionary tree. Horizontal axis: species, being B, Bacteria; A, Archaea and E, Eukaryota. Vertical axis: DDR proteins with their family codes (see Table ST1) and PTMs (coloured triangles) they produce. Squares have been coloured from light orange to brown showing different blocks of orthologs conservation. Squares corresponding to model species (blue triangles) have slightly darker colours. For species whose genomes have been completely sequenced, gene losses are represented as blue squares, while grey squares indicate absences (due to proteome incompleteness or gene loss, which are indistinguishable).

4.4 DDR PROTEINS DOMAIN ANALYSIS (Figure 8, Panel C)

4.4.1 Domain identification

In our domain detection analyses we used an independent e-value (which shows the significance of the sequence in the whole database search, if this was the only domain that had been identified) threshold of 0.05, a total of 163 non-overlapping different domain families were detected in the 2656 human DDR orthologous proteins, being the most abundant the BRCT repeats, Helicase_C, Pkinase and AAA domains (data not shown). On the contrary, for 6 of the Hsa-118 proteins (ATRIP, CLSPN, F175A, FACD2, MRI40 and PALB2), no Pfam domains were detected.
Alternative thresholds were tested, but those more restrictive than 0.05 produced the loss of true positives, while a more relaxed threshold resulted in the identification of numerous overlapping domains and many false positives (data not shown).

4.4.2 Domain conservation and protein domain architectures

The protein domain content is consistent for the Hsa-118 orthologs of most families (Table ST4 Annex), since 90 proteins (76.3 % of the 118) (1433E, BRCA1**, BRCC3, BRE, CDT1, CHK1, CHK2, CUL1, CUL4, DCR1B, DCR1C, DDB1*, DNA2L, DTL, EME1, ERCC1, ERCC2, ERCC3, ERCC5, ERCC6, ERCC8, EXO1, FBX31, H2AX, HERC2, HUS1, KAT5, MAPK2, MDC1, MDM2, MDM4, MK03, MRE11, MSH2, MSH3, MTA2, MUS81, MYST1, NBN, NR4A2, NSE2, PAXI1, PCNA, PIAS4, PMS2, PRKDC, RAD1, RAD17, RAD18, RAD9, RBBP8, RBX1, RD23B, RFA1, RFA2, RFA3, RMI1, RN168, RNF8, SIRT1, SKP1, SLX1, SLX4, SMC1A, SMC5, SMC6, SOX4, TAOK1, TDP1*, TERF2, TIF1B, TIM, TOP3A*, TOPB1, TP53B, TRIPC*, UBE2N, UBE2T, UBP11*, UBR5, UIMC1, WEE1, XLF, XPA, XPC, XPF, XRCC1, XRCC4, XRCC5 and XRCC6) had a conserved domain architecture in all orthologs (*see Figure 18 for more information). However, we observed lineage specific domain insertions and/or losses in some orthologs of 20 families (Figure 19, Table ST4 Annex). In this regard, plants have suffered extensive domain shuffling in Lineage-Specific Expansions (LSEs), where orthologs have additional domains (see Discussion). Examples are the PHD domain in PIAS1, the SAP domains in PARP2, the zf-CCHC domain in TIPIN, or the zf-RanBP domain in TPD2, a tyrosyl-DNA phosphodiesterase that can remove a variety of covalent adducts from DNA through hydrolysis of a 5'-phosphodiester bond. Most of these domain combinations seem to be specific of plants since we have not detected these architectures in other species.

For those proteins that contained different domain distributions in certain orthologs, an *in-silico* synthetic architecture was built to facilitate comparisons.

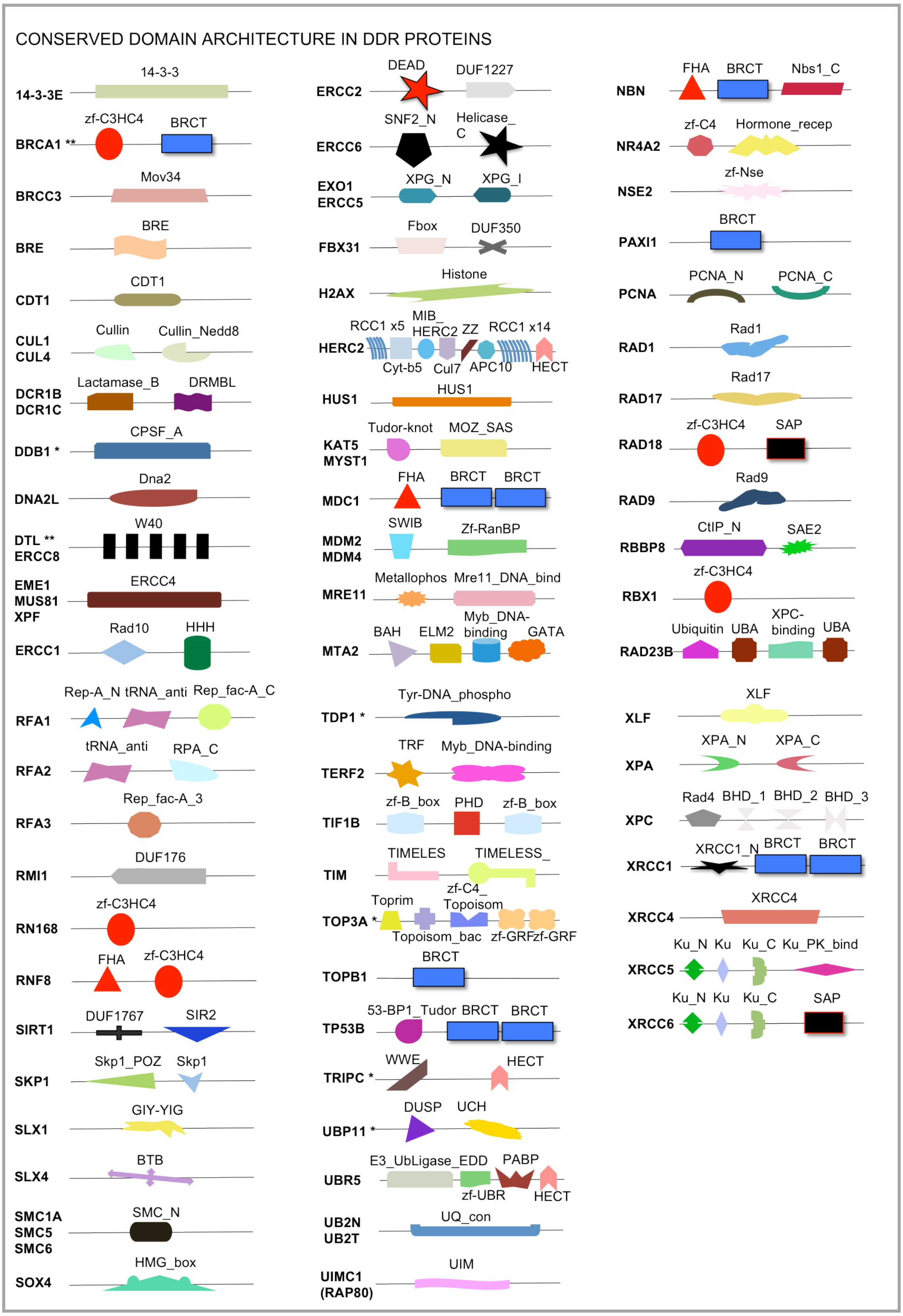


Figure 18. Schematic representation of Hsa-118 DDR proteins with conserved domain architecture in all orthologs. The exact position of the domains in the sequence and the relative length of the proteins have not been taken into account for this representation. Note: *Some orthologs presented slightly different domain architecture, most probably due to incorrect gene prediction or domains with bad scores. **The number of BRCT repeats varies in different species.

Interesting examples of domain variation among orthologs are the cases of two highly conserved HR proteins, BLM and RAD51.

For instance, beyond the DEAD-Helicase_C-RQC-HRDC core architecture of domains, the BLM protein in upper eukaryotes has a BDHCT domain (found in Bloom's syndrome DEAD helicase subfamily) in the N-terminal region of the sequences. In the *S. cerevisiae* ortholog, a Helicase_Sgs1 domain is found instead of the conserved HRDC, and these domains may have an equivalent role because they are homologous domains belonging to the same HRDC-like domain clan (CL0426). The C-terminal is the region with more variations in BLM; for example, in *D. radiodurans* the sequence has evolved to confront the types and amounts of DNA damage, having three HRDC repeats that increase the efficiency of the helicase activity, and in *A. variabilis* a GerE domain was identified. This DNA-binding, helix-turn-helix domain, present in transcription regulators of LuxR family of response regulators, is involved in quorum-sensing control of luminescence.

The particular case of BLM illustrates fairly well the acquisition of novel functions due to diverse protein domain architectures reflecting substantial differences at the species level likely due to divergence.

In spite of the apparent enormous variation of domain architecture in the Rad51 orthologs, a deeper analysis indicates the differences are not so large. The RecA domain found in the bacterial proteins is a homolog of the Rad51 domain found in the archaea and eukaryotes orthologs, being both domains members of the AAA clan (CL0023); and the Cdd1 domain (expressed as part of the pathogenicity locus operon in some bacteria) present in the archaeal sequences is homologous to the Helix-hairpin-helix motif (HHH) found in the eukaryotic orthologs, being these two domains part of the HHH clan (CL0198). These results might point towards an emergence of the eukaryotic Rad51 caused by a combined evolution of bacterial and archaeal sequences.

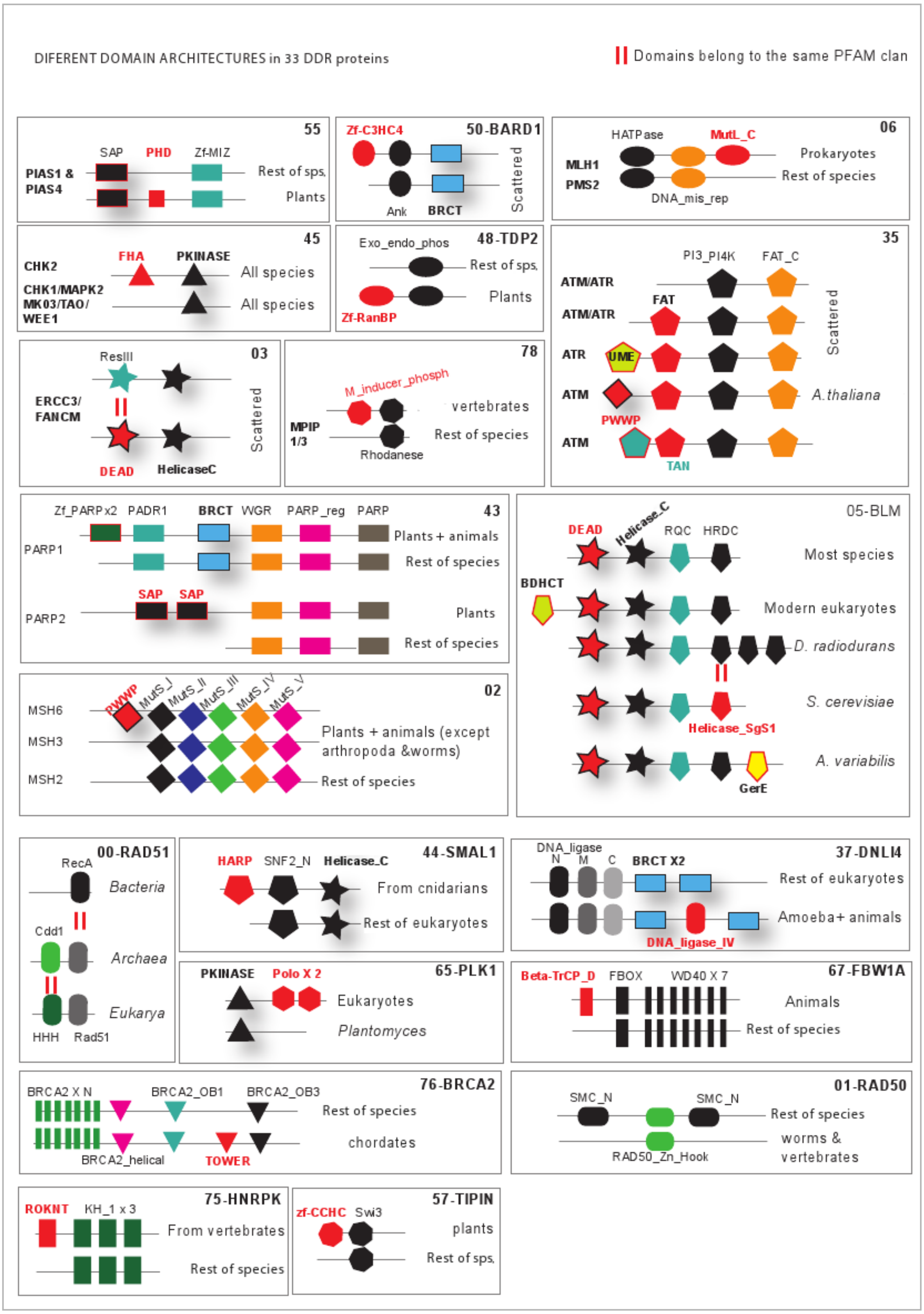


Figure 19. Schematic representation of Hsa-118 DDR proteins with variations in domain architecture in the orthologs. The exact position of the domains in the sequence and the relative length of the proteins have not been taken into account for this representation. Shadowed shapes indicate shuffling, and the || pipe indicates that both domains belong to the same clan.

4.4.3 Clustering of protein domain profiles

Protein domain phylogenetic profiles were clustered and six distinguishable blocks were obtained (figures 20a and b). Domains widely distributed and found in the three kingdoms of life have been coloured in magenta (bottom); those detected in archaea and eukaryotes, and in few bacteria are in the red block; while in the dark orange block are clustered the domains found in eukaryotes and in few prokaryotes. Both in the red and dark orange blocks, there are a series of domains in bold and highlighted in black boxes, which are those found in eukaryotic proteins but that have also been detected in the Planctomycetes representatives included in this study. Interestingly, the DNA_ligase_A_C and DNA_ligase_A_N (domains found in the eukaryotic DNA ligase 4 (LIG4), involved in NHEJ) have been detected in G. obscuriglobus, while in P. staleyi, the domains detected have been the UQ_con (UBE2N/T), PI3_PI4_kinase (found in PIKK kinases), 14-3-3 (in 14-3-3 proteins, involved in multiple signaling pathways) and PHD finger motif (found in Transcription intermediary factor 1-beta TIF1B (KAP1/ TRIM28) and plants PIAS1, and many proteins involved in chromatin-mediated gene regulation).

The light orange block contains domains present in few prokaryotes and scattered among eukaryotic species with some absences in whole phyla (i.e. PADR1 (a domain of unknown function found in PARP1) in fungi, or Replication fork protection component Swi3 (TIPIN) and Timeless (TIM) in *Apicomplexa*) (figure 20b). The yellow block includes domains detected mainly in bacteria but absent in most of the other organisms; this absence is likely due to the existence in eukaryotes of homologous domains performing similar functions (domains in green font colour). Finally, the domains found only in modern organisms, and those scattered among few ancient eukaryotes, and some plants and fungi, are coloured in blue (top). In this blue block, important absences of domains in *C. elegans* and *arthropoda* can be seen, such as the example of Telomere-length maintenance and DNA damage repair (TAN), a motif found in PIKK kinases which is essential for telomere length maintenance and ATM action in response to DNA damage. Other important missing domains in *C. elegans*, most *arthropoda* and *C. intestinalis*, are Nbs1_C from NBN, 53-BP1_Tudor (TP53B) and DNA_ligase_IV (DNLI4) (see red coloured fonts in figure 20b).

The results show that there are domains widely distributed among all organisms, while others seem to be specific of phyla or restricted to certain groups of species. Examples of conserved domains are DEAD, Helicase_C, HATPase_C (which have been detected in all 47 species), or MutS and MutL_C domains, SMC_N and Pkinase. On the other hand, the BDHCT and ROKNT domains are specific of *chraniata* (from *D. rerio* to *H. sapiens*) and there are other domains such as RecA, Cdd1, GerE or Topo_zn_Ribbon, which are present in prokaryotes and very few eukaryotes. Table 4 contains information about these domains and the DDR proteins where they are found.

Widely distributed domains	**Function**	**Protein**
DEAD	DEAD/DEADH bocx helicase. Unwinds nucleic acids.	BLM, FANCM
Helicase_C	Helicase conserved C-terminal domain	BLM, ERCC3, ERCC6, FANCM, SMAL1
HTPAse	Found in several ATP-binding proteins, like histidine kinase, DNA girase B, topoisomerases or heat shock protein SHP90.	MLH1, PMS2
MutS	Found in proteins of the MutS family	MSH2, MSH3, MSH6
MutL_C	Found in proteins of the MutL familiy. The domain is involved in proteins dimerisation.	MLH1, PMS2
SMC_N	Found at the N terminus of SMC (structural maintenance of chromosomes) proteins, which are essential for successful chromosome transmission during replication and segregation of the genome.	Rad50, SMC1A, SMC5, SMC6
Pkinase	Contains the catalytic function of protein kinases.	CHK1, CHK2, MAPK2, MK03, TAOK1, PLK1, WEE1
Narrowly distributed domains	**Function**	**Protein**
BDHCT	C-terminal domain in Bloom's syndrome DEAD helicase subfamily	BLM (*chordata*)
ROKNT	Found at the N-terminus of RNP K-like proteins that also contains KH domains	HNRPK (*chordata*)
RecA	Catalyses an ATP-dependent DNA strand-exchange reaction that is the central step in the repair of dsDNA breaks by homologous recombination	Rad51 (*Bacteria*)
Cdd1	Cdd1 protein is expressed as part of the pathogenity locus operon in different orders of bacteria	Rad51 (*Archaea*)
GerE	DNA-binding, present in transcription regulators of the LuxR family of response regulators, involved in quorum-sensing control of luminescence.	BLM (*A. variabilis*)
Topo_zn_Ribbon	C-terminal zinc-ribbon-like domain found in bacterial topoisomerase I (type IA) enzymes. This domain is still considered to be a member of the zinc-ribbon superfamily despite not being able to bind zinc.	TOP3A (*B. aphidicola*)

Table 4. Examples of widely and narrowly distributed DDR domains, their functions and proteins containing these domains.

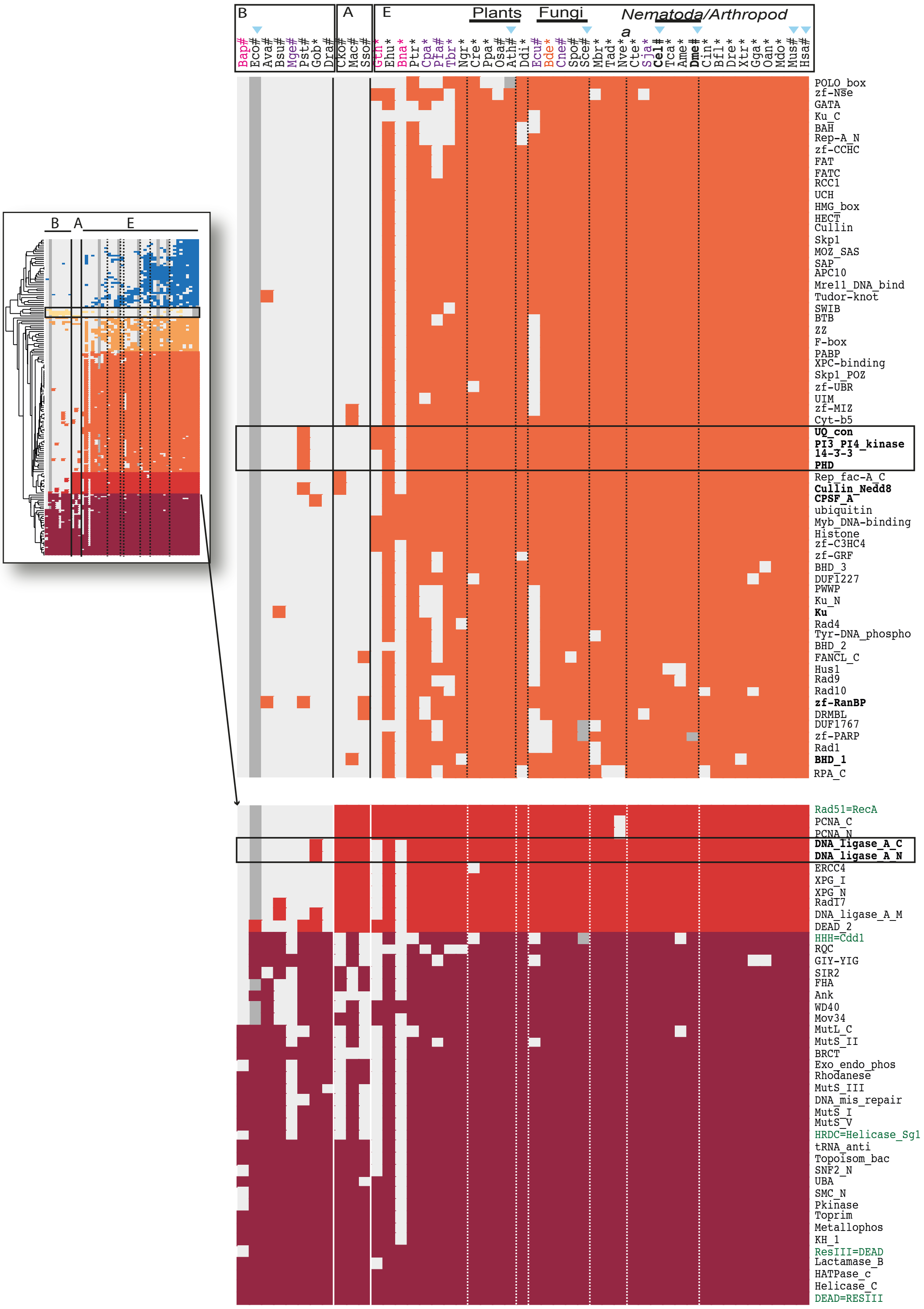


Figure 20a. Cluster of the 163 different domains detected in the 118 Human DDR proteins orthologs according to the phylogeny of the 47 species analyzed in this study ordered by the evolutionary tree. Horizontal axis: species, being B, Bacteria; A, Archaea and E, Eukaryota. Vertical axis: DDR domains. Six distinguishable blocks were obtained, from magenta (bottom) to blue (top). In this figure, the magenta, red and dark orange blocks have been zoomed in. The domains in green font colour with the = symbol are domains belonging to the same clan, and the domains in bold and highlighted in black boxes are those typical of eukaryotic proteins but that have been detected in the *Planctomycetes* representatives.

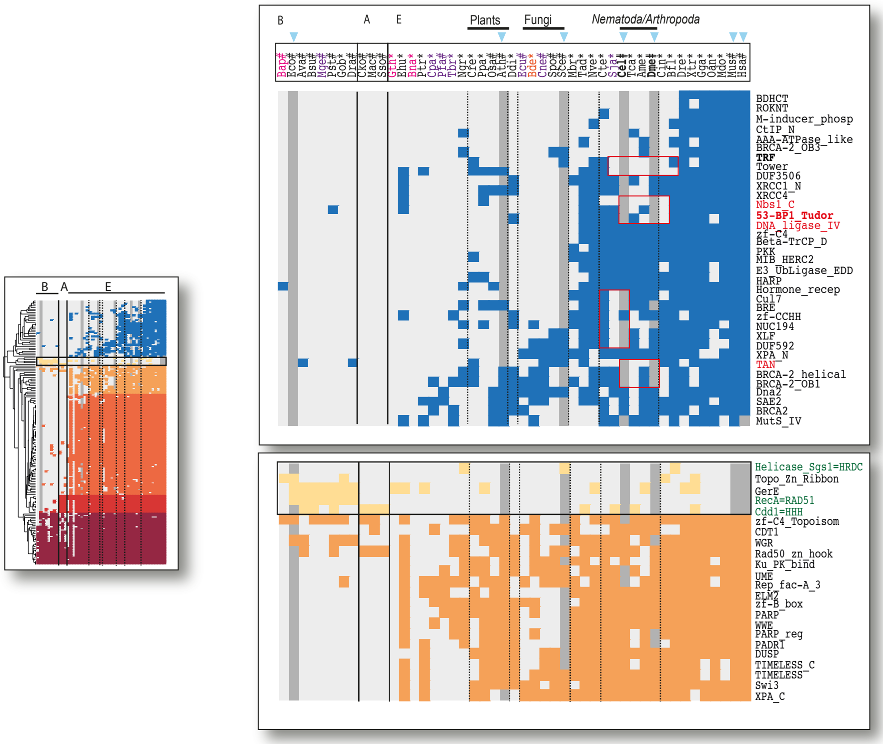


Figure 20b. Cluster of the 163 different domains detected in the 118 Human DDR proteins orthologs according to the phylogeny of the 47 species analyzed in this study ordered by the evolutionary tree. Horizontal axis: species, being B, Bacteria; A, Archaea and E, Eukaryota. Vertical axis: DDR domains. Six distinguishable blocks were obtained, from magenta (bottom) to blue (top). In this figure, the light orange, yellow and blue blocks have been zoomed in. The domains in green font colour with the = symbol are domains belonging to the same clan, and the domains in red font which have also been highlighted with red boxes are those with important functions and present in essential proteins but that are missing in some organisms (*C. elegans, C. intestinalis*, etc) and in certain phyla, mainly *arthropoda*.

4.4.4 Identification of conserved uncharacterized regions

From the human 118 DDR proteins, there are 6 (ATRIP, CLSPN, F175A, FACD2, MRI40 and PALB2) where no Pfam domains were detected using our methodology (see Methods 3.5.2 and Results 4.4.1).

From the remaining 112 proteins, 5 had equal or less than 10% of their length covered by Pfam domains, while 18 had equal or more than 80% of their sequence length covered by detected domains (see Table 5 below).

Protein length coverage by Pfam domains			
10% =<	>= 80%		
MDC1	1433E	PCNA	SMC5
RN168	BRE	RAD1	SMC6
SLX4	CUL1	RAD51	UBE2N
UIMC1	CUL4A	RFA3	XRCC4
XPF	HUS1	SKP1	XRCC5
	MSH2	SMC1A	XRCC6

Table 5. Human DDR proteins length coverage by Pfam domains. Those proteins where the domain coverage was equal or below 10% of the sequence, and those where the coverage was equal or over 80% are shown.

For those 6 proteins where no Pfam domains were detected, we analyzed the MSA of the corresponding orthologs to detect conserved regions in the sequences and further run profile-profile methods (HHPRED) to detect distant similarities. Hits with probability over 75% and e-value below 0.05 were found for all proteins but FACD2 (see Table 6).

Protein	Fragment start (Human protein)	Fragment end (Human protein)	Database	ID and annotation	Probability	E-value	Score	Identity %
ATRIP	114	225	Pf25	PF09798 LCD1: DNA damage checkpoint protein	96.47	1.20E-05	45.61	15
CLSPN	1043	1143	Pf25	PF09444 MRC1: MRC1-like domain	92.53	0.0058	31.52	31
F175A	1	180	pdb70	2o95_A 26S proteasome non-ATPase regulatory subunit 7	96.21	5.80E-05	46.65	17
F175A	1	180	Pf25	PF01398 Mov34: Mov34/MPN/PAD-1 family	87.51	0.044	29.92	24
F175A	1	180	smrt_18f	smart00232 JAB_MPN JAB/MPN domain.	83.69	0.0064	28.08	23
MRI40	87	197	hhr	d1atza_ c.62.1.1 (A:) von Willebrand factor A3 domain, vWA3	75.88	0.043	25.52	7
MRI40	87	197	pdb70	2x5n_A SPRPN10, 26S proteasome regulatory subunit RPN10	97.18	9.90E-07	53.61	21
MRI40	87	197	Pf25	PF04056 Ssl1: Ssl1-like	95.19	0.00033	39.8	19
MRI40	87	197	scop70	d1jeyb2 c.62.1.4 (B:6-241) Ku80 subunit N-terminal domain	96.2	3.30E-05	44.93	17
MRI40	87	197	smrt_18f	smart00327 VWA von Willebrand factor (vWF) type A domain	90.39	0.0011	31.42	19
PALB2	854	1184	pdb70	2w18_A PALB2, fancn, partner and localizer of BRCA2	100	0	813.98	61
PALB2	854	1184	scop70	d1tbga_ b.69.4.1 (A:) beta1-subunit of the signal-transducing G protein heterotrimer	96.56	9.30E-06	50	18

Table 6. HHPred results for the human DDR proteins where Pfam domains had not been detected.

In the ATRIP protein, a fragment similar to the LCD1 domain found in the *S. cerevisiae* and *S. pombe* checkpoint kinases is detected. In yeast, LCD1 is necessary for CHK1p activation in response to DNA damage and is also required for efficient DNA damage-induced phosphorylation of Rad9p and activation of Rad53p in response to DNA damage or DNA replication blocks [310]. This results show this region is highly conserved among different organisms and is involved in ATRIP's function.

In Claspin, there is a hit to a MRC1-like domain. This putative domain is the most conserved region in mediator of replication checkpoint protein 1, which is required for Rad3-dependent activation of the checkpoint kinase Cds1 in response to replication fork arrest. This domain is detected in the Claspin yeast orthologs in our study.

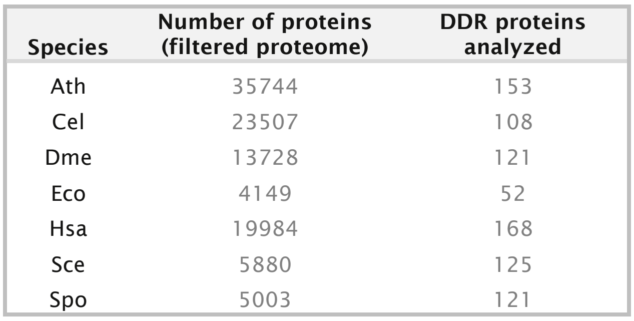
In F175A/Abraxas, which is a central scaffold protein that assembles the various components of the BRCA1-A complex and mediates the recruitment of BRCA1, there is a region similar to the MPN (JAB1/Mov34) domain. This domain is a widespread protein module found in archaea, bacteria, and eukaryotes. In the latter, the domain is found in subunits of various multiprotein complexes, including the proteasome, and might be involved in the removal of the polyubiquitin chain from substrate proteins [311]. Another component of the BRCA1-A complex, the MRI40/BABAM1/NBA1 protein, which is required for the complex integrity, has a region similar to the VWA domain (found in proteins involved in transcription, DNA repair, ribosomal and membrane transport, and the proteasome). The presence of these hits reveals similarities between the structure of the 26S proteasome (responsible for ubiquitin-dependent protein degradation) and the BRCA1-A complex. Curiously, in MERIT40, the same region has a hit with high probability and low e-value to the N-terminal domain of the Ku80 protein.

Finally, in PALB2/FANCN, the most conserved fragment is the region involved in the interaction with RAD51 and BRCA2.

Our results show that, though no Pfam domains have been detected in few proteins, these have conserved regions with hits to domains known to be involved in DDR processes.

4.4.5 Domain enrichment

The enrichment analysis shows that, for the 168 human proteins analysed (see Methods 3.5 and Box3), the BRCT repeats (involved in protein-protein interaction, and in DNA and Poly(ADP-ribose) binding), and the Rad51, Helicase_C and AAA (ATPase family Associated with diverse cellular Activities) domains are highly enriched in human DDR proteins (see Table 7).



Species	Number of proteins (filtered proteome)	DDR proteins analyzed
Ath	35744	153
Cel	23507	108
Dme	13728	121
Eco	4149	52
Hsa	19984	168
Sce	5880	125
Spo	5003	121

Box3: model species used for the domain enrichment analysis, number of proteins in the proteomes (after being filtered) and number of DDR proteins analyzed.

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
BRCT	PF00533	11	21	3.49E-18	7.50E-16
Rad51	PF08423	4	7	1.65E-07	3.56E-05
Helicase_C	PF00271	9	107	3.00E-07	6.44E-05
AAA	PF00004	7	56	4.29E-07	9.23E-05
DRMBL	PF07522	3	3	5.84E-07	1.25E-04
MutS_I	PF01624	3	3	5.84E-07	1.25E-04
Rep_fac_C	PF08542	3	3	5.84E-07	1.25E-04
RQC	PF09382	3	3	5.84E-07	1.25E-04
UQ_con	PF00179	6	39	8.37E-07	1.80E-04
ERCC4	PF02732	3	4	2.32E-06	4.99E-04
HHH	PF00633	3	4	2.32E-06	4.99E-04
MutS_II	PF05188	3	4	2.32E-06	4.99E-04
MutS_III	PF05192	3	5	5.76E-06	1.24E-03
MutS_V	PF00488	3	5	5.76E-06	1.24E-03
FAT	PF02259	3	6	1.15E-05	2.46E-03
FATC	PF02260	3	7	1.99E-05	4.28E-03
SMC_N	PF02463	3	8	3.17E-05	6.81E-03

Table 7. DDR domains enriched in *H. sapiens* (only those with a Bonferroni adjusted p-value < 0.01 are shown). Though the DDR proteins figures are low, the numbers were Bonferroni corrected to normalize the data.

Figure 21 shows the overlap among enriched domains in the four seed organisms. Interestingly, the number of enriched domains shared considering a Bonferroni adjusted p-value < 0.05 is higher between *A. thaliana* and human than between the latter and *S. cerevisiae* (even though yeast and *H. sapiens* are phylogenetically more related*)*, which could be due to gene losses.

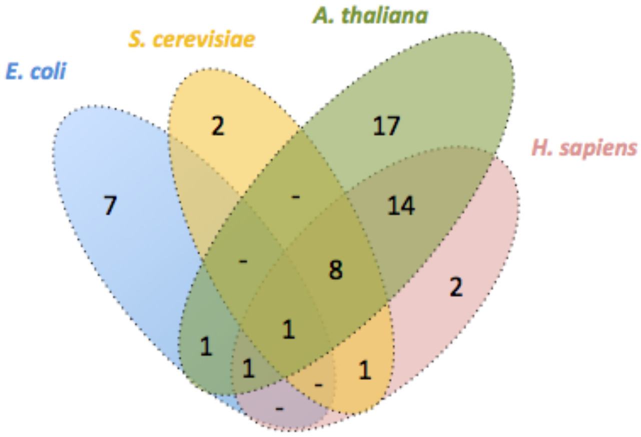


Figure 21. Overlap of DDR domains enriched in the four seed datasets in this study: *E. coli, S. cerevisiae, A. thaliana* and *H. sapiens* (only those domains with a Bonferroni adjusted p-value < 0.05 were considered). The only domain enriched in the four species is the SMC_N domain.

The domain enrichment analysis was also performed in other model species: *E. coli*, *A. thaliana, S. cerevisiae, S. pombe, C. elegans and D. melanogaster* (see tables ST5a-f, Annex). *A. thaliana* seems to be the species with more enriched domains when considering a Bonferroni adjusted p-value below 0.05 (see table ST5a, Annex), followed by human.

Considering a Bonferroni adjusted p-value below 0.05, the only domain enriched in all the seven species is the SMC_N, which is an ancient domain found in the SMC family of proteins and in some RAD50 orthologs, and is widely distributed among the proteomes analyzed.

In the eukaryotic species, the AAA domain, which is found in proteins involved in DNA replication, signal transduction, regulation of gene expression and many other processes, is enriched. This domain is not enriched in *E. coli* but it is found in the Lon protease, a protein that binds DNA and is required for cellular homeostasis and for survival from DNA damage and developmental changes induced by stress.
The BRCT domain is highly enriched in all species, as well as the Rad51 (RecA in *E. coli*), MutS (I to V) domains, the RAD17 from the Cell cycle checkpoint protein RAD17 (but in *E. coli*), the Helicase_C, the HhH-GPD superfamily base excision DNA repair protein, which contains a diverse range of structurally related DNA repair proteins, etc.

In general terms, most of the domains found to be enriched in the DDR proteins bind DNA, are from ancient origin, and are widely distributed in all the 47 organisms in this study.

4.4.6 Domain distribution in functional categories

As explained in Methods 3.9 b), the Pfam domains identified in the Hsa-118 DDR proteins were grouped according to a 4 tier-classification of sensors, mediators, transducers and effectors.
The results show that sensors and effectors are the tiers with more specific domains, even though sensors are the third most populated class (see Results 4.8). Examples of domains found only in sensors are the MutS, the MutL and DNA_mis_repair found in the MMR proteins MLH1 and PMS2, and the Ku and PARP related domains. Domains specific of mediators are the Histone, UIM (Ubiquitin interaction motif) from Rap80, Tower (which is essential for appropriate binding of BRCA2 to DNA [312]), BRE from BRCC45 o Swi3 from Tipin. Transducers are the functional class with less exclusive domains, being some examples the POLO_box, or the UQ_con from the Ubiquitin-conjugating enzymes UBE2N and UBE2T. Finally, effectors were the tier including the biggest number of specific domains, such as the XRCC4 found in EME1, MUS81 and XPF, or the SWIB domain contained in the MDM2 and MDM4 checkpoint proteins.
No age trend was found for the distribution of domains in the different functional classes.

On the other hand, two domains were found in all four tiers: BRCT and zf-C3HC4, which are among the most abundant domains in the DDR proteins, and are also enriched domains in the *H. sapiens* DDR proteins. Both domains are of ancient origin, since the BRCT was present in bacteria, and zf-C3HC4 seemed to emerge in ancient eukaryotes for it was detected in the hypothetical protein GTHECHR2167 from the *G. theta* nucleomorph.

4.5 EMERGENCE OF THE DDR PATHWAY USING GENE-CONTENT BASED METHODS (Figure 8, Panel D)

Homo sapiens is the most extensively studied species in terms of DDR functional data. Moreover, as it is the "youngest" species in our species-tree (Figure 10), the Hsa-118 dataset has been used as the reference to address how these components have emerged along evolution. Although this strategy presents certain issues (see discussion), it is a plausible approach to our purposes.

4.5.1 Classification of DDR components into protein families

We have classified the Hsa-118 set in protein families since some of them are homologous and have the same domain architecture (or arrangement of protein domain content). Thus, we have divided the 118 proteins in 95 subfamilies (Table ST1 Annex, right columns), where each family can either be single (79 families) or multigene (16 families, Table 8), containing the latter genes that are homologs between each other (i.e. PIAS1 and PIAS4).

Family number	Multi-gene	Proteins	Function
1	55	PIAS1, PIAS4	E3 SUMO-protein ligases
2	88	MDM2, MDM4	Both proteins inhibit p53-mediated cell cycle arrest. MDM4 inhibits degradation of MDM2
3	43	PARP1, PARP2	Poly [ADP-ribose] polymerases
4	78	CDC25A (MPIP1), CDC25C (MPIP3)	M-phase inducer phosphatases
5	20	UBE2N (UBC13), UBE2T	Ubiquitin-conjugating enzyme E2 proteins
6	21	CUL1, CUL4	Core components of cullin-RING-based E3 ubiquitin-protein ligase complexes that mediate ubiquitination and subsequent proteasomal degradation of target proteins.
7*	89, 58	MDC1, PAXIP1 (PTIP)	
8	41	KAT5, MYST1	Histone acetyltransferases
9	40	DCR1B (Apollo), DCR1C (Artemis)	5'-3' exonucleases
10	24	ERCC5, EXO1	ERCC5 is a single-stranded DNA endonuclease involved in DNA excision repair, while EXO1 is a 5'->3' double-stranded DNA exonuclease
11	12	XPF, MUS81, EME1	Endonucleases, XPF involved in NER, and MUS81 and EME1 in cleavage of Holliday junctions
12	08	SMC5, SMC6, SMC1	Structural maintenance of chromosomes proteins
13	06	MLH1, PMS2	Mismatch repair proteins
14	45	CHK1, MAPK2, TAOK1, WEE1, MK03, CHK2 (+ FHA domain)	Kinases involved in cell-cycle checkpoints
15	02	MSH2, MSH3, MSH6	Mismatch repair proteins
16	35	ATM, ATR, PRKDC	PIK-related kinases involved in DNA damage sensing
17	03	ERCC3, FANCM	ATP-dependent helicases

Tale 8. Multigene families, proteins that comprise them and their function. *: Family7: although because of auomatic detection Ensemble COMPARA assigns PTIP (PAXI1) and MDC1 to the same family, besides sharing the promiscuous BRCT domain, there is no detectable sequence similarity between these two proteins, and therefore we do not consider them as members of the same family.

We also manually checked for consistency in the protein domain content to identify variations and common domain architectures as a quality check to assess or not the orthology relationship (Figures 18 and 19, Table ST4 Annex).

Also, although three groups of proteins (PARP1/PARP2, ATM-ATR-PRKDC, and CHK1/CHK2) have different domain architectures, they are usually considered homologous sequences because they share homology at certain domains and regions, in particular in the domain involved in the function (i.e.: kinase).

For the gene content-based analyses, each protein was considered as an independent hit to build the presence/absence matrix, therefore sequence similarities within the dataset were not taken into account.

4.5.2 Phylogenetic profiles

The **Ortho-DDR** set was used to build a phylogenetic profile indicating whether a protein is either present or absent in the 47 proteomes screened.

a) Hierarchical clustering

The hierarchical clustering of the Hsa-118 profile (ordered according to our given species-tree) produced five distinguishable and stable blocks of proteins (sequential scale of oranges from ancient to modern species, Figure 17) indicating the presence or absence of proteins in all the screened species. The most represented core of proteins is located at the base of the clustered profile (Figure 17, lightest orange box). These proteins are RAD51, RAD50, MSH3, PCNA, XPB, TOP3A, MLH1, BLM, SMC1A, and MSH6, all involved mainly in repair. The next block includes presences in ancient single celled eukaryotes. Expected absences are those corresponding to endosymbionts (*Guillardia theta* and *Bigelowiella natans*, pink names Figure 17) where the genomic sequences correspond to the nucleomorph (the remains of the prokaryotic-based engulfment in the first event of endosymbiosis). In this block, absences of orthologs (Blue boxes, Figure 17) start to appear more frequently, especially in *C. elegans*, *D. melanogaster* and fungi. This indicates potential rewiring of unrelated proteins in these lineages may accomplish the functional requirements for a proper DNA response. In the next block losses are prevalent and significant in plants, fungi, *C. elegans*, and *D. melanogaster*. The next block includes proteins mostly lost in fungi and plants, where the last block points to Chordate proteins.
To delineate the pace of growth, we next plotted the aggregated frequency of the orthologs present at least in one representative species of each age group (Figure 22). Around 10% of the Hsa-118 proteins are traceable to the prokaryotic group including Archaea and Bacteria. At the eukaryotic split represented by the free-life planktonic organism *Emiliania huxleyi (Hacrobia*), there was a large expansion of genes where most of the DDR components were acquired (around 60-70%). From this point, the incorporation of novel components was less remarkable being completely established at the *Vertebrata* group. No further innovations are detected after that evolutionary point.

b) Ancestral reconstruction algorithms

We also calculated the emergence of the different components of the DDR using two alternative algorithms (Wagner and Dollo parsimony) and plotted the results, where the same trend is maintained (Figure 22) with some expected differences in the relative numbers at the ancient *Eukaryota* stages (see Discussion).

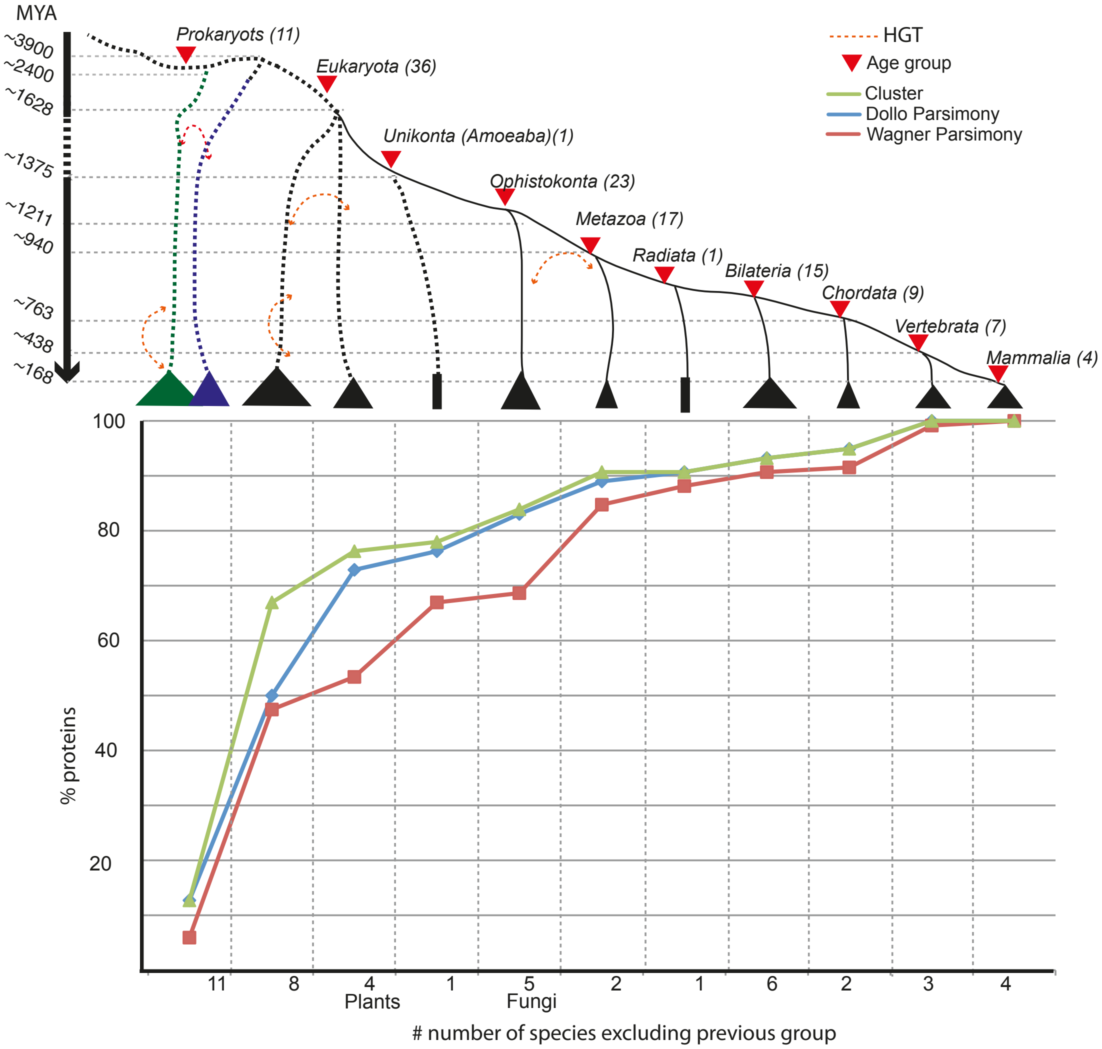
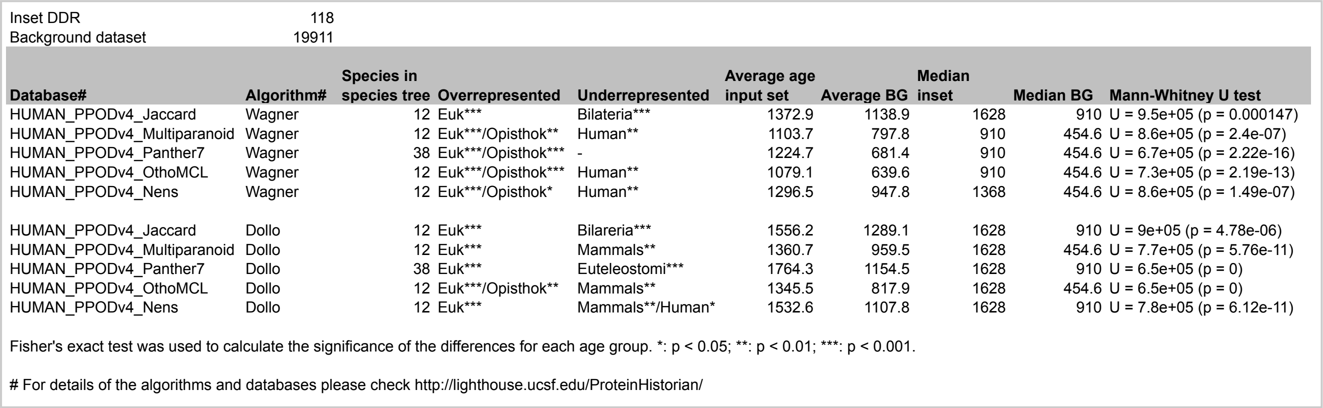


Figure 22. Relative presence of DDR proteins according to age groups. To represent the pace of growth according to the relative contribution of each age group on DDR components, we plotted the aggregated frequencies (normalized by group size) for each of three methods: hierarchical clustering, Dollo parsimony and Wagner parsimony. Red dotted arrows represent Horizontal Gene Transfer (HGT) events between phylogenetic groups.

4.6 GENE AGES (Figure 8, Panel D)

In agreement with our results, alternative gene age enrichments using different methods and databases, showed that the Hsa-118 set was significantly enriched in genes corresponding to the *Eukaryotic* age and the *Opisthokonta* split, while being significantly underrepresented in mammalian ages (Table 9).



Inset DDR	118
Background dataset	19911

Database#	Algorithm#	Species in species tree	Overrepresented	Underrepresented	Average age input set	Average BG	Median inset	Median BG	Mann-Whitney U test
HUMAN_PPODv4_Jaccard	Wagner	12	Euk***	Bilateria***	1372.9	1138.9	1628	910	U = 9.5e+05 (p = 0.000147)
HUMAN_PPODv4_Multiparanoid	Wagner	12	Euk***/Opisthok**	Human**	1103.7	797.8	910	454.6	U = 8.6e+05 (p = 2.4e-07)
HUMAN_PPODv4_Panther7	Wagner	38	Euk***/Opisthok***	-	1224.7	681.4	910	454.6	U = 6.7e+05 (p = 2.22e-16)
HUMAN_PPODv4_OthoMCL	Wagner	12	Euk***/Opisthok***	Human**	1079.1	639.6	910	454.6	U = 7.3e+05 (p = 2.19e-13)
HUMAN_PPODv4_Nens	Wagner	12	Euk***/Opisthok*	Human**	1296.5	947.8	1368	454.6	U = 8.6e+05 (p = 1.49e-07)
HUMAN_PPODv4_Jaccard	Dollo	12	Euk***	Bilareria***	1556.2	1289.1	1628	910	U = 9e+05 (p = 4.78e-06)
HUMAN_PPODv4_Multiparanoid	Dollo	12	Euk***	Mammals**	1360.7	959.5	1628	454.6	U = 7.7e+05 (p = 5.76e-11)
HUMAN_PPODv4_Panther7	Dollo	38	Euk***	Euteleostomi***	1764.3	1154.5	1628	910	U = 6.5e+05 (p = 0)
HUMAN_PPODv4_OthoMCL	Dollo	12	Euk***/Opisthok**	Mammals**	1345.5	817.9	1628	454.6	U = 6.5e+05 (p = 0)
HUMAN_PPODv4_Nens	Dollo	12	Euk***	Mammals**/Human*	1532.6	1107.8	1628	910	U = 7.8e+05 (p = 6.12e-11)

Fisher's exact test was used to calculate the significance of the differences for each age group. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

For details of the algorithms and databases please check http://lighthouse.ucsf.edu/ProteinHistorian/

Table 9. Ages enrichment analysis for the Hsa-118 set of proteins. Wagner and Dollo algorithms were used to calculate age enrichment in different databases. In most cases genes corresponding to the Eukaryotic age and the *Opisthokonta* split are enriched, while being significantly underrepresented in bilaterians, mammalian and human ages.

4.7 GENE-TREES AND SPECIES-TREES (Figure 8, Panel E)

Due to the polytomies at deep branches in our species phylogeny, we devised a phylogenetic pipeline that conducted phylogenetic gene trees to identify family-specific evolutionary trends. Not unexpectedly, some gene trees were statistically unsupported at deep branches, where discrepancies with the species-tree were largely found. The overall results of the phylogenetic pipeline indicate that the evolutionary history of most families may be more complex than expected, and also extensive HGT must have occurred, as it is very frequent to observe arthropod, fungal, and worm sequences grouping along ancient eukaryotes instead of their assumed more related lineages. This might indicate that certain genes in these organisms may be older due to HGT events.

We generated a total of 65 trees, of which 49 were single gene trees and 16 were multigene trees (i.e. that included homologous genes).
Eight trees were unreliable or insufficiently supported at deep branches due to the complexity of the MSA because of the high divergence of certain regions of specific proteins (10-RBX1, 14-USP11, 17-H2AX, 20-UBE2N/T, 23-RFA3, 25-TPYD1, 38-HUS1 and 52-MMS21). A deeper sampling of these trees would be required to obtain more reliable results, but his was not done due to computational constraints.
On the other hand, out of 90 genes, 24 were in agreement with the species-tree. From these, only in 4 cases the species where perfectly sorted (MK03, ERCC3, RNF8, XRCC1) according to the species-tree, and 20 followed the species-tree with minor variations likely due to the quality of specific sequences (ERCC8, EXO1, PIAS1/4, TDP2, PARP1, PLK1, etc).
In general, large misplacements were observed for arthropods and worms (in 36 cases, from which *C. elegans* is misplaced in 26 trees) and to a lesser extent (17 trees) fungi and plants (ERCC2 example, figure 23), indicating that most of the families have suffered complex evolutionary histories.

Regarding the 16 multigene trees, in 7 trees while certain members of the family followed the taxonomic tree, the other members did not (for instance Family 17, where

ERCC3 follows the species order while FANCM does not, Family 11 where EME1 follows while MUS81 and XPF do not; or Family 6 where CUL1 follows while CUL4 does not).
In 6 multigene trees none of the members followed the species tree (Families MSH2/3/6, and ATR/ATM/ PRKDC, SF17)), while in 2 multigene trees: 55-Family 1 (PIAS1/4) and 88-Family 2 (MDM2/4) both proteins follow the species tree.

In all cases, when different domain architectures were found, phylogenies were conducted with the common domains only.

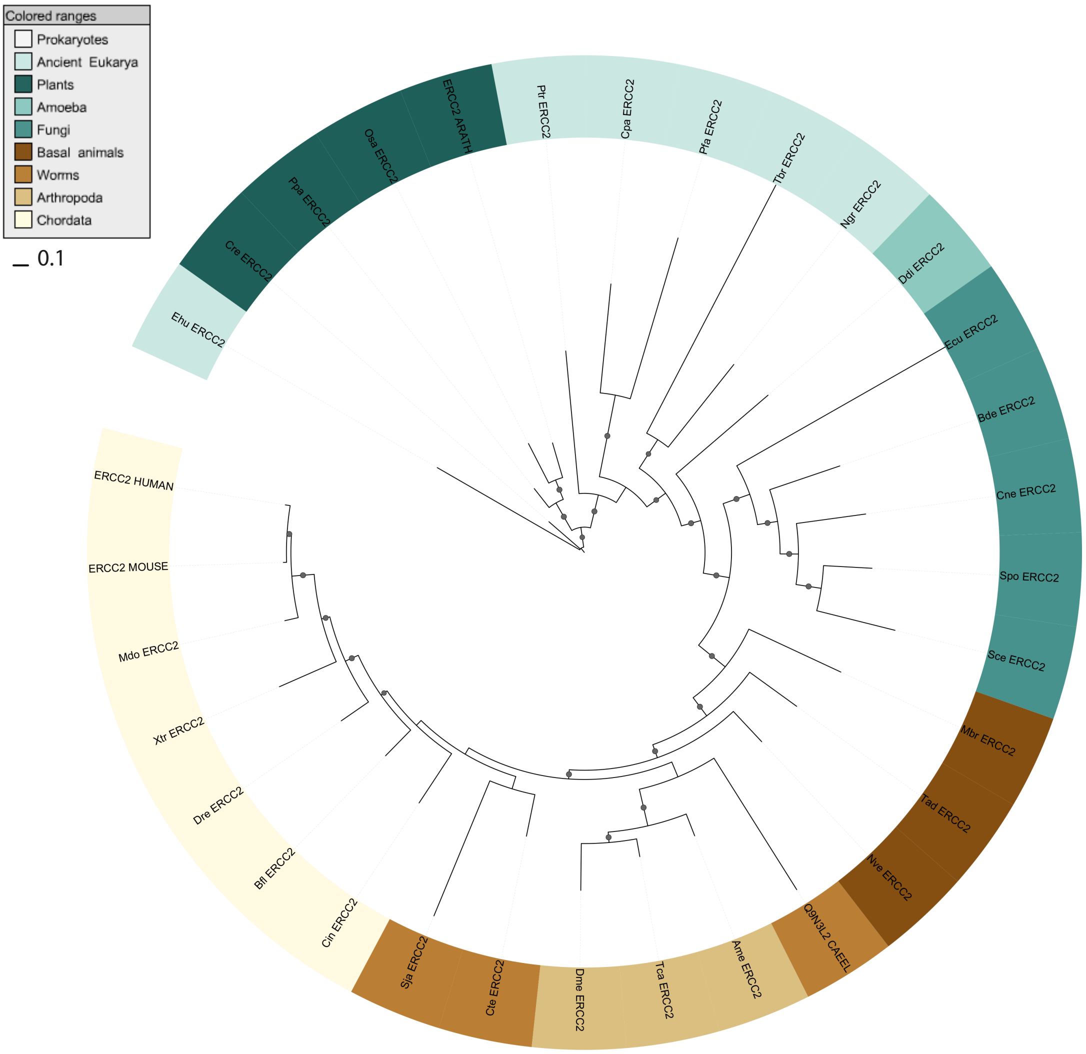


Figure 23. Phylogenetic tree of the DNA excision repair protein ERCC2. In this well supported tree at all levels, the sequences are sorted according to the species-tree, but for plants, which are grouped with the phytoplankton *E. huxleyi* (used for rooting the tree); and *S. japonicum* and *C.teleta,* which cluster closer to *chordata* than *arthropoda*. Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%.

One of the most conflicting phylogenetic placements corresponds to *C. elegans*, which does not group with worms in most trees (many with probability value > 80% or 0.08), like in the cases of BLM, ERCC1, MSH2, MUS81, PCNA, RAD50, RFA1, SKP1, or in Family13 consisting of MLH1 and PMS2 (see figure 24). According to this family tree, both MLH1 and PMS2 from *C. elegans* group near ancient eukaryotes, which might be the result of HGT from ancestral organisms. Other unexpected placements of orthologs are PMS2 from *C. reinhardtii*, likely due to its partial sequence, or *E. cuniculi* near the PMS2 root, and *C. intestinalis* close to basal eukaryotes. Besides, fungi are positioned closer to ancient eukaryotes than plants, and both the phytoplankton *E. huxleyi* and the diatom *P. tricornutum* cluster with plants. According to the domain architecture, the MLH1 orthologs detected in prokaryotes might actually be the ancestral gene that gave rise to the duplication of MLH1, and are most likely PMS2.

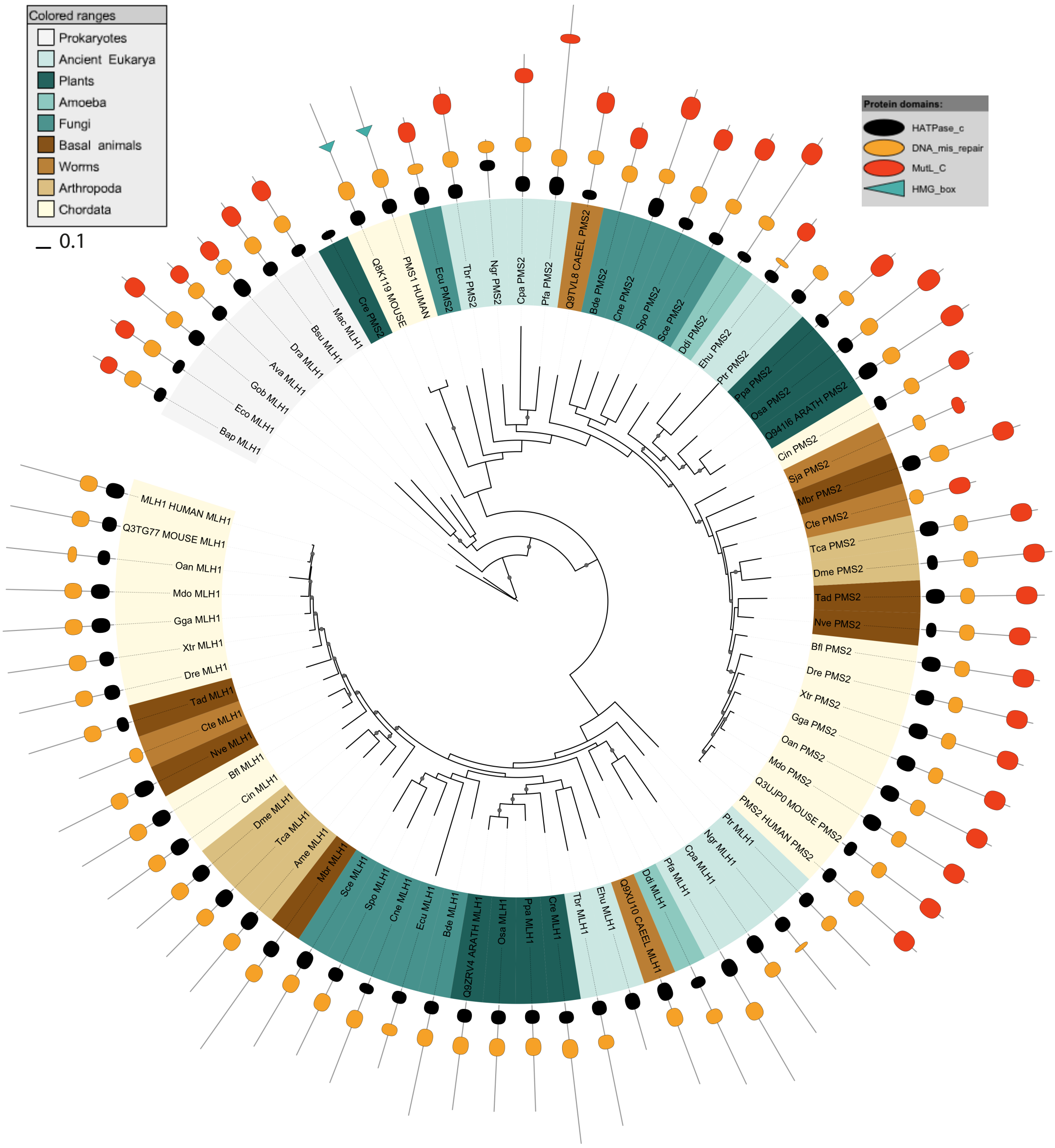


Figure 24. Phylogenetic tree of Family13, comprising MLH1 and PMS2. The sequences of PMS1, other member of this family (which includes additional domains), have also been included to clarify the correct position of in-paralogs. In this tree, *C. elegans* is located with ancient eukaryotes. Other unexpectedly positioned orthologs are PMS2 from *C. reinhardtii, E. cuniculi* and *C. intestinalis.* According to the domain architecture, the MLH1 orthologs detected in prokaryotes might actually be the ancestral gene that gave rise to the duplication of MLH1, and are most likely PMS2.

Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%. The differential domain architecture of the orthologs is represented. Sequences are depicted showing their relative length, being the N-terminal region the closest displayed to the centre of the tree.

In certain cases such as in the proteins MSH3 and MSH6 (figure 25), FANCM (figure 26), 14-3-3, or SMC1A, plants orthologs group closer to animals than fungi.

In the example of Family15, comprising MSH2, MSH3 and MSH6, besides the positioning of the fungi in an older clade than plants are, other sequences are located in a position that differs with the species phylogenetic tree. For instance, in MSH2, though the gene tree follows almost perfectly the species-tree, the *E. cuniculi* protein is at the base of the tree (maybe due to the sequence being incomplete or to a contamination), and *C. elegans* is also found by ancient eukaryotes.

In the MSH3 tree, the *E. huxleyi* protein clusters with the prokaryotic sequences (mean probability values > 80%), which could be due to HGT events or an artifact. Finally, in the case of MSH6, the basal organism *M. brevicollis* groups with fungi, which in the tree are closer to ancient eukaryotes than plants; *and N. vectensis* and *C. teleta* are found next to chordata, probably due to the fact that all these sequences present a PWWP domain (red rhombus in figure 25) in the N-terminal region of the protein.

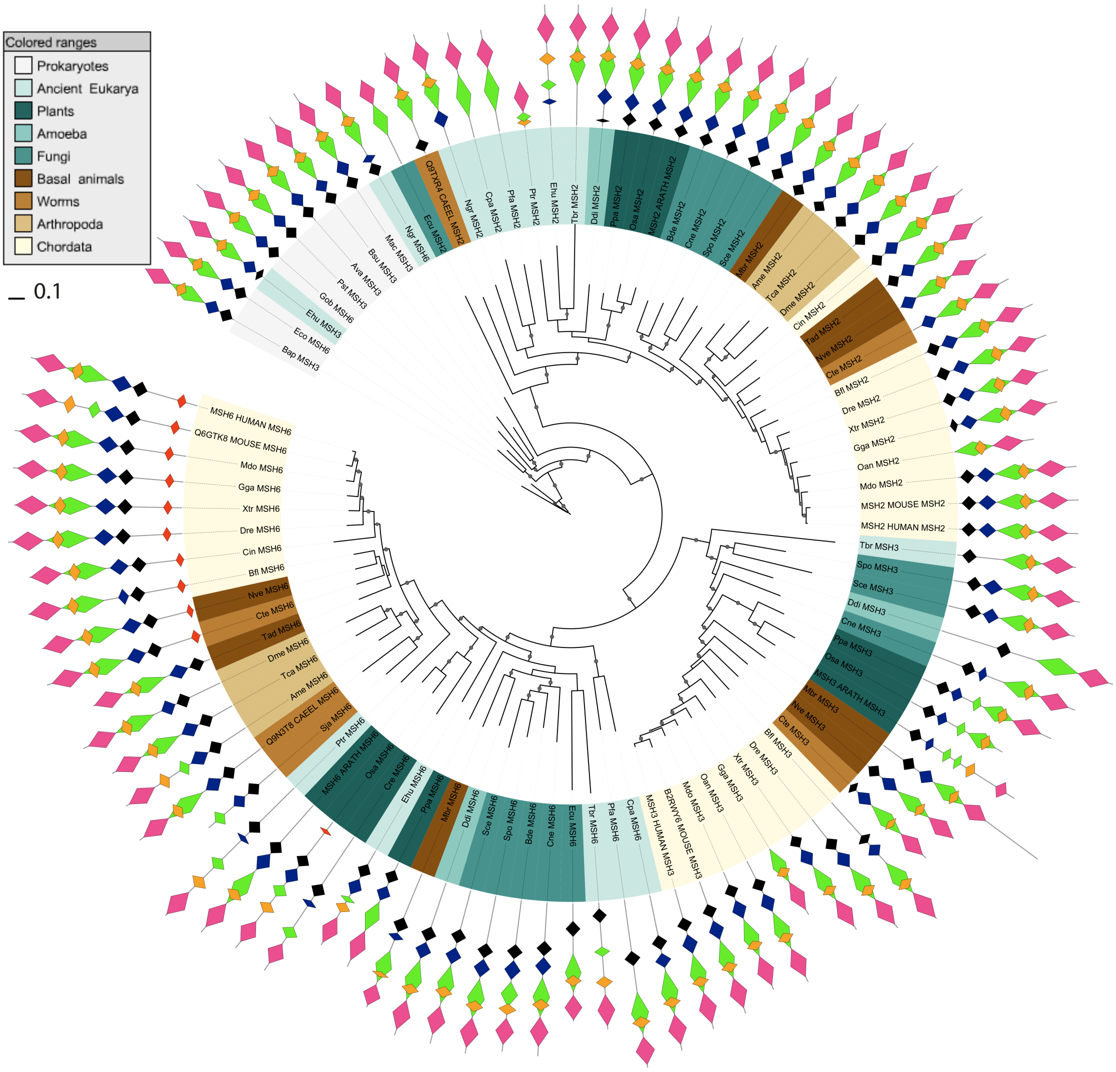


Figure 25. Phylogenetic tree of Family15, comprising MSH2, MSH3 and MSH6. Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%. The differential domain architecture of the orthologs is represented. Sequences are depicted showing their relative length, being the N-terminal region the closest displayed to the centre of the tree. The PWWP domain is detected in *chordata* plus *C. teleta* (polychaete) and *N. vectensis* (sea anemone). Also a tudor-like domain is detected in plants.

As explained previously, the accuracy of the existing methods to predict domain boundaries is not entirely satisfactory, as can be observed in figure 25, where the different domains (MutS I to V) are not always well defined: in most cases the MutS IV domain is identified within domain III, while in other occasions, domain III is detected as two repeats.

Besides the inversion fungi-plants, arthropods are also found closer to fungi than basal eukaryotes and annelidae in several trees. One of them is the case of Family17 (figure 26), comprised by ERCC3 and FANCM. The ERCC3 part of the tree is well supported and reflects almost perfectly the species-tree order. Interestingly, the sequence of the planctomycete *G. obscuriglobus* is found among the ancient eukaryotes (mean probability values > 80%), instead of being with the rest of bacterial sequences, which could indicate a eukaryotic inference (HGT event). On the other hand, the FANCM tree presents an inversion fungi-plants and other arthropods-basal eukaryote.

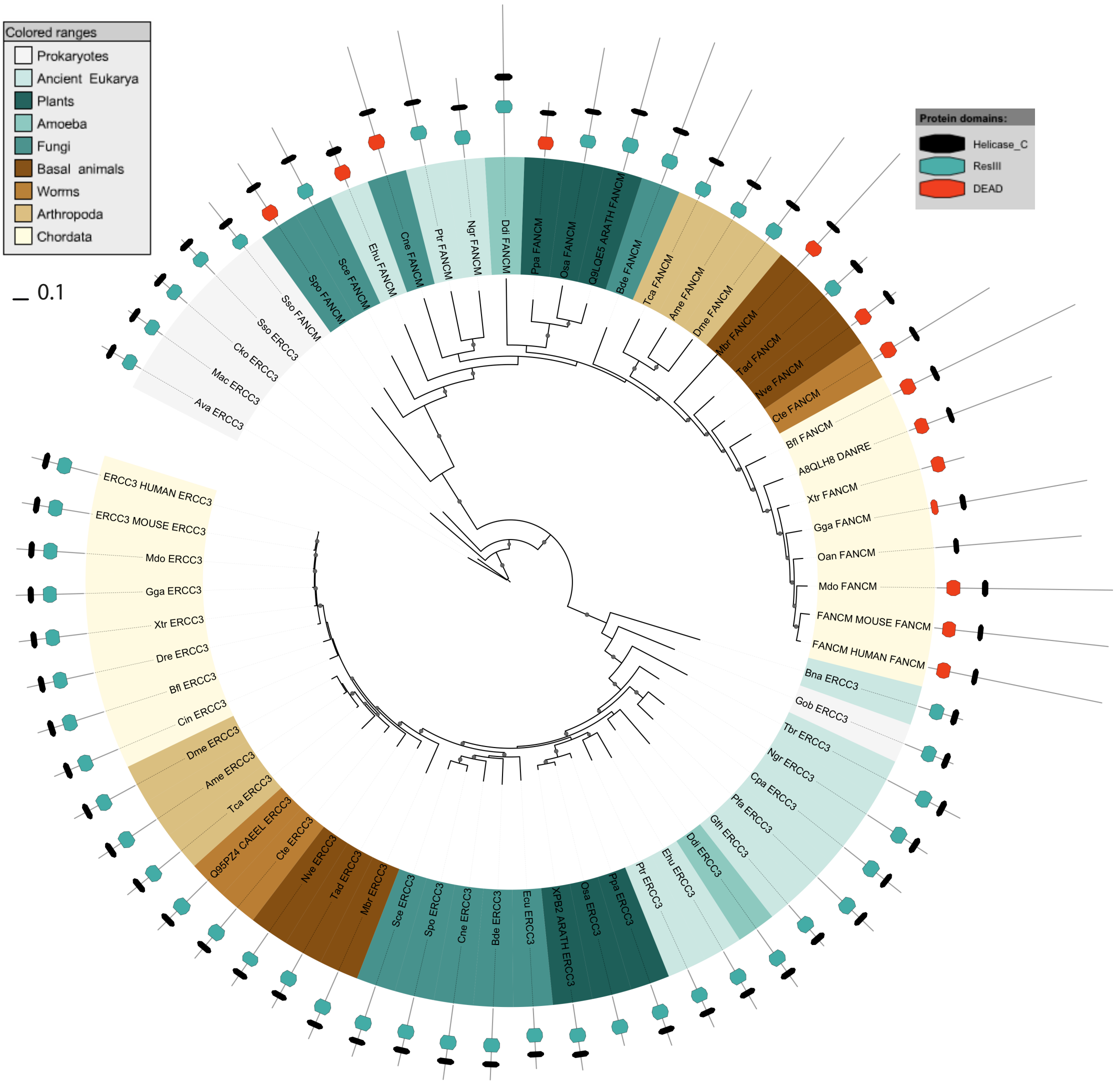


Figure 26. Phylogenetic tree of Family17, comprising ERCC3/XPB and FANCM. Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%. The *G. obscuriglobus* ERCC3 clusters among the ancient eukaryotes and the FANCM tree presents one inversion fungi-plants and other arthropods-basal eukaryotes.

Other cases where arthropods or worms are found closer to fungi than basal eukaryotes are RFA3, ERCC5, RAD1 or MUS81 (however, it seems that *C. teleta* (annelida) tends to escape this trend). In some other trees the same result is obtained but for *M. brevicollis*, which is found generally after fungi (TOP3A, XPF, TIM) or with ancient eukaryotes (MRE11). Also in the TIM tree *C. elegans* clusters by fungi, which happens with the RAD9 tree as well.

Within the 65 trees, at least three trees showed the same inconsistencies common to proteins being part of the same protein complex, like the case of XRCC5 and XRCC6, where *C. elegans* groups within arthropods instead of with other worms.
Another interesting example of trees, are the cases of RAD17 and TOPB1, both proteins that are part of a complex in the ATR pathway. These two replication stress proteins, which are grouped together in the protein cluster (figure 17, Results 4.3.4), show the same inconsistencies in the gene-tree (see figures STt24 and STt24 in Annex), maybe pointing towards a co-evolution process, where proteins are transferred in blocks.
Other example is the case of 06-Family13 (see figure 24), comprised by PMS2 and MLH1 –which dimerize to form MutL alpha in the MMR pathway–, where the sequences from *C. elegans* group close to ancient eukaryotes.

In 52 out of 65 trees the *Chordata* members followed the taxonomic-tree, with few exceptions probably due to artifacts given the incompleteness of certain sequences. One such example is the case of RAD51, whose tree is very problematic but well supported (mean probability values > 80%) (figure 27). In this tree the *O. anatinus* ortholog is remarkably out of place since it clusters with the *T. adhaerens* sequence, and both are located by ancient eukaryotes. Moreover, the *C. intestinalis* sequence is placed between Bacteria and Archaea. Nevertheless, this topology is well supported.
One of the most extreme cases is the gene-tree from XRCC5 where the genes of fungi, arthropods and worms are grouping with ancient eukaryotes (see figure STt29, Annex).

Regardless of the differences observed between some taxonomic and gene-trees (particularly in the cases of arthropods and worms), the reconstruction of phylogenetic trees helps detect wrongly assigned orthology. In addition to Family13 explained above (Figure 24), in Family8, the MYST1 and KAT5 orthologs seem to have been incorrectly identified, since according to the tree MYST1 from *N. gruberi, C. parvum*, *C. falciparium*, plants and *D. discoideum* could be KAT5 instead, having emerged in ancient eukaryotes, where the *P. tricornutum* ortholog has been correctly detected (Figure 28). Analogously, Family 9 with Artemis/Apollo, also shows some inconsistencies produced by automatic orthology assignations, whereby DCR1B of *P. patens*, fungi and *M. brevicollis* are DCR1A instead, another protein of the same family of Artemis/Apollo.

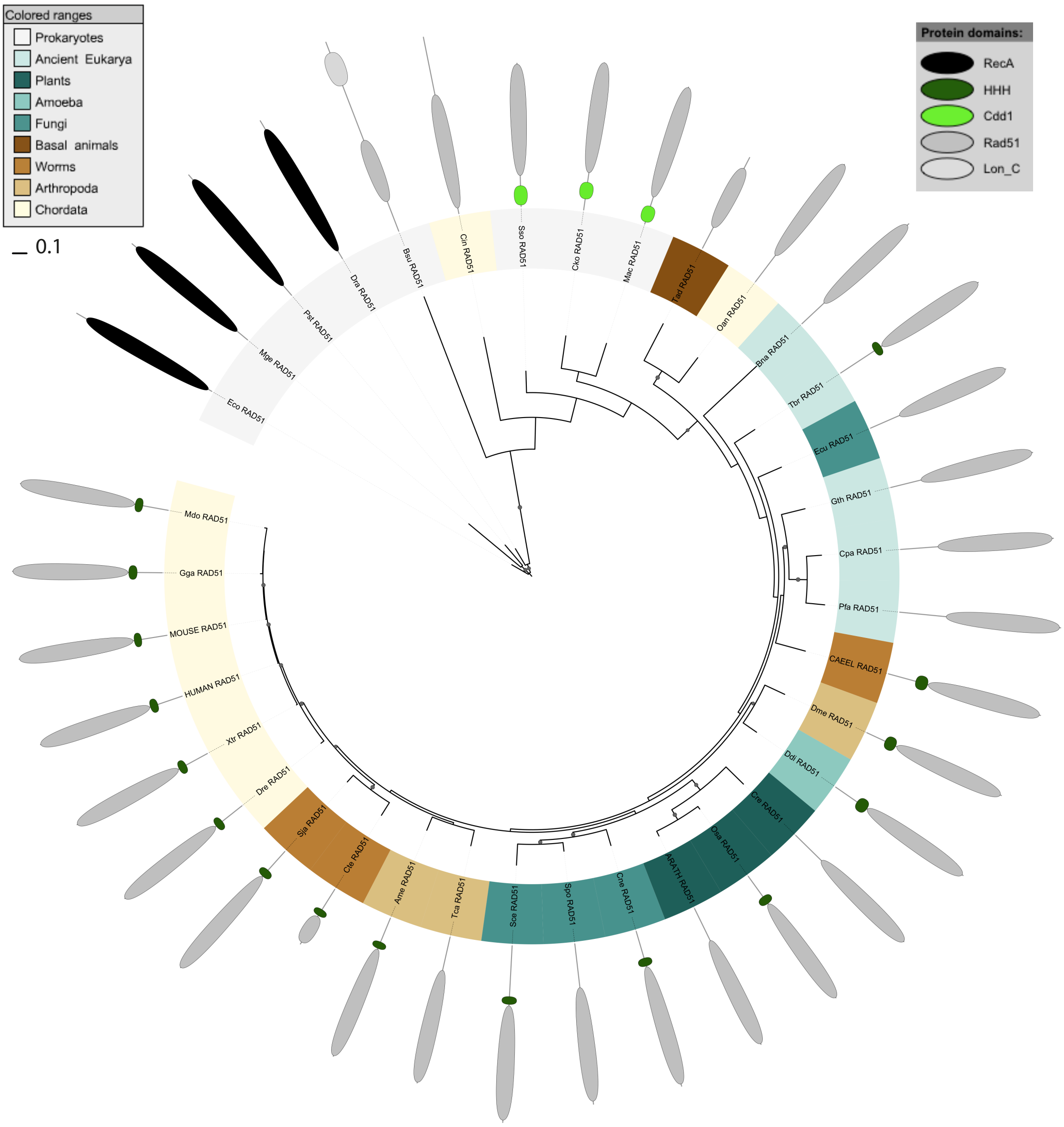


Figure 27. Phylogenetic tree of DNA repair protein RAD51. The *O. anatinus* ortholog strangely clusters with the *T. adhaerens* sequence, and both are located by ancient eukaryotes. Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%. The differential domain architecture of the orthologs is represented. Sequences are depicted showing their relative length, being the N-terminal region the closest displayed to the centre of the tree.

Regarding the age of the proteins and as expected, those more ancient (see figure SF1 in Annex, protein blocks 0, I and II), have more variations between the gene- and the species-tree, compared to the blocks comprised by more modern proteins (blocks III and IV). Considering the protein complexes, in some gene-trees we have seen the same variations in species order in proteins forming part of the same complex (such as the abovementioned case of RAD17 and TopBP1, or in XRCC5 and XRCC6 (Ku80 and Ku70 respectively), in which fungi are located prior to plants and where *C. elegans* clusters among arthropods, and all are before basal eukaryotes (see figures STt29 and STt30, Annex).

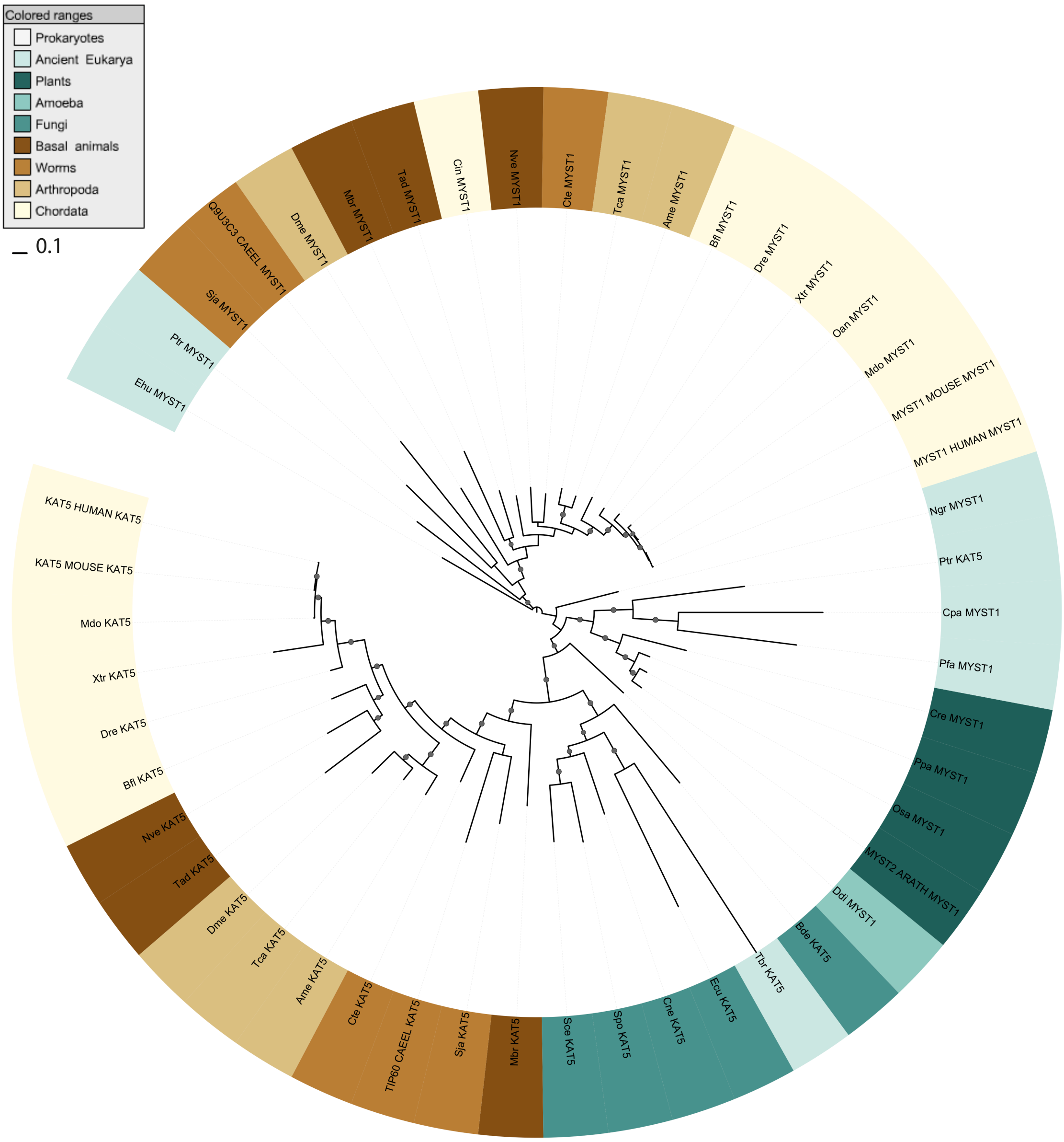


Figure 28. Phylogenetic tree showing Family8, including MYST1 and KAT5. The MYST1 orthologs in *N. gruberi, C. parvum, P. falciparum,* plants and *D. discoideum* (right half of the tree, in light green) were probably misidentified by InParanoid, and are KAT5 instead. Species are coloured according to their phylogenetic group. The dots in the tree branches mean clades with probability value > 80%.

A brief summary of results is included in Table 10, which shows examples of different cases found when comparing the species- and gene-trees built for the selected DDR proteins.

Comparison between taxonomic- and gene trees					
Gene tree follows taxonomy	*E. huxleyi* and plants at the base of the tree	Fungi near the base of the tree	Plants closer to basal eukaryotes than fungi	Worms closer to *metazoa* than *arthropoda*	Basal eukaryotes close to *metazoa*
03-ERCC3 12-XPF 29-TIM 45-MK03 47-XRCC1 51-ERCC8 82-RNF8	15-SKP1 19-XPC 22-ERCC1	08-SMC5 32-XRCC6 45-WEE1	02-MSH3 02-MSH6 23-RFA3	00-RAD51 01-RAD50 45-MAPK2 69-XPA	25-TDP1 28-RAD1 43-PARP2

Table 10. Examples of different cases found when comparing the taxonomic- and the gene-trees of DDR proteins.

The rest of the gene-trees generated in this study can be found in the Annex (figures SFt1 to 59).

4.8 EMERGENCE OF POTENTIAL FUNCTIONS OF DDR PROTEINS (Figure 8, Panels F / G) (Panel F) (Panel G)

a) GO assignment

The functional enrichment of the Hsa-118 based on GO analyses indicated strong associations with DNA repair and response, as all the proteins were enriched in terms related these processes (ST6, Annex**)**.

b) A broader classification

To have a more general view of the functions, we classified the Hsa-118 set in “Sensors”, “Mediators”, “Transducers” and “Effectors”, as described in the literature.
The most populated classes for *H. sapiens* are, the effectors (48 proteins) followed by mediators (with 40), sensors (with 32), and finally transducers (with 24) (ST7, Annex**,** figure 29). There is functional overlapping in the DDR set (figure 29) where the largest isthe one created by sensors (followed by mediators) with the rest of the classes.

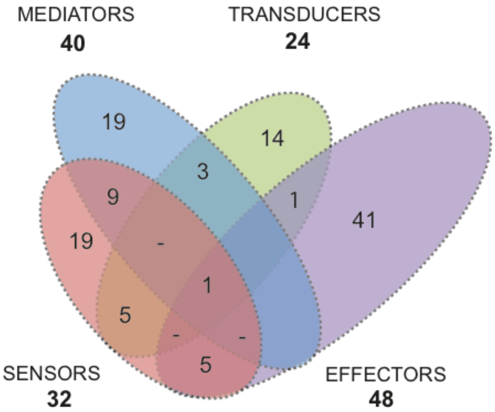


Figure 29. Venn diagram of the human 118 DDR proteins grouped in a four-tier functional classification: sensors, mediators, transducers and effectors.

When we plotted the incorporation of functions of the *Hsa*-118 along the evolutionary scale in our age groups, the ancestral core at the prokaryotic level was comprised of sensors and effectors (blue and purple lines, Figures 30a and 30b) with one mediator (PCNA).

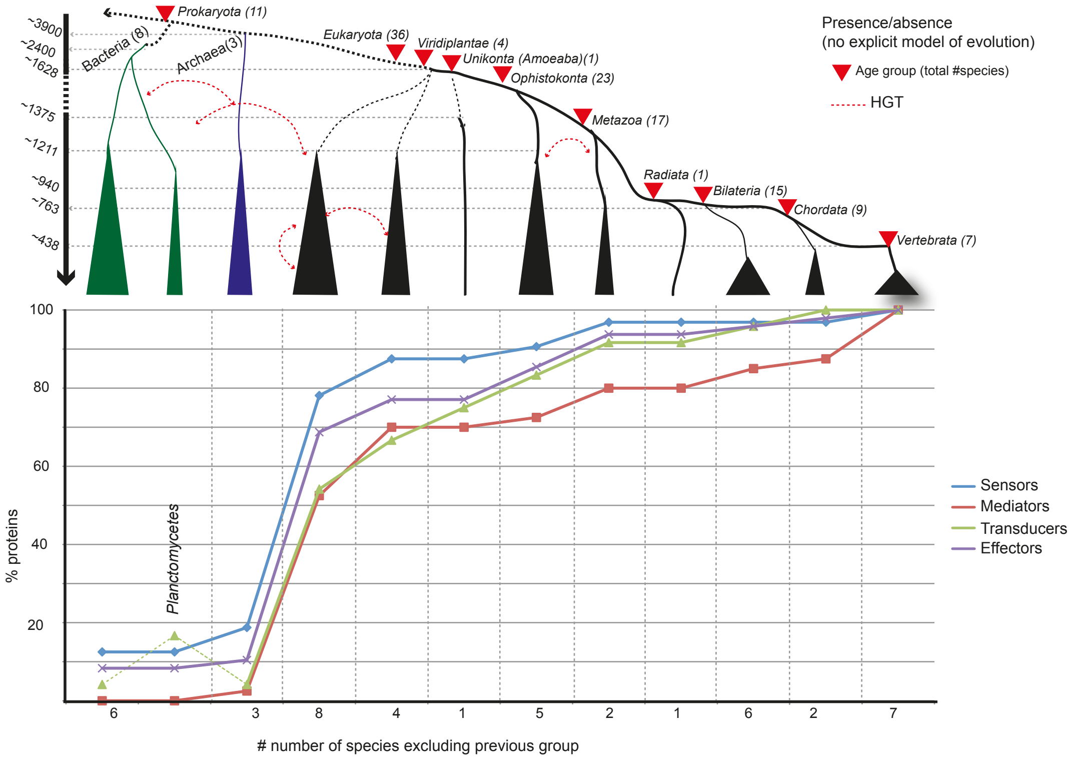


Figure 30a. Relative presence of DDR proteins according to a four-tier classification (sensors, mediators, transducers and effectors) considering the aggregated frequencies obtained by hierarchical clustering. Dotted arrows represent Horizontal Gene Transfer (HGT) events between phylogenetic groups.

Interestingly, two bacterial species from the *Planctomycetes* phylum contain homologs of transducers (green dashed line, Figure 30a), as well as *E. coli* containing an ortholog to MSH6 (Figure 17), though the regulatory function of this protein has likely specialized towards modern eukaryotes, due to the incorporation of a PWWP domain. At the eukaryotic age, most of the proteins are still effectors and sensors, although a large expansion of mediators and transducers are incorporated.

From this point on, sensors, effectors and, to a lesser extent, mediators, are incorporated at a steady pace, while transducers expand largely reaching convergence at the *Metazoa* group. Mediators are the last to reach this convergence, which takes place in *vertebrata*.

The different methods used for the emergence of functions of the *Hsa*-118 along the evolutionary scale in our age groups provide similar trends. The slight variations observed are due to the distinct evolutionary assumptions on which the algorithms are based.

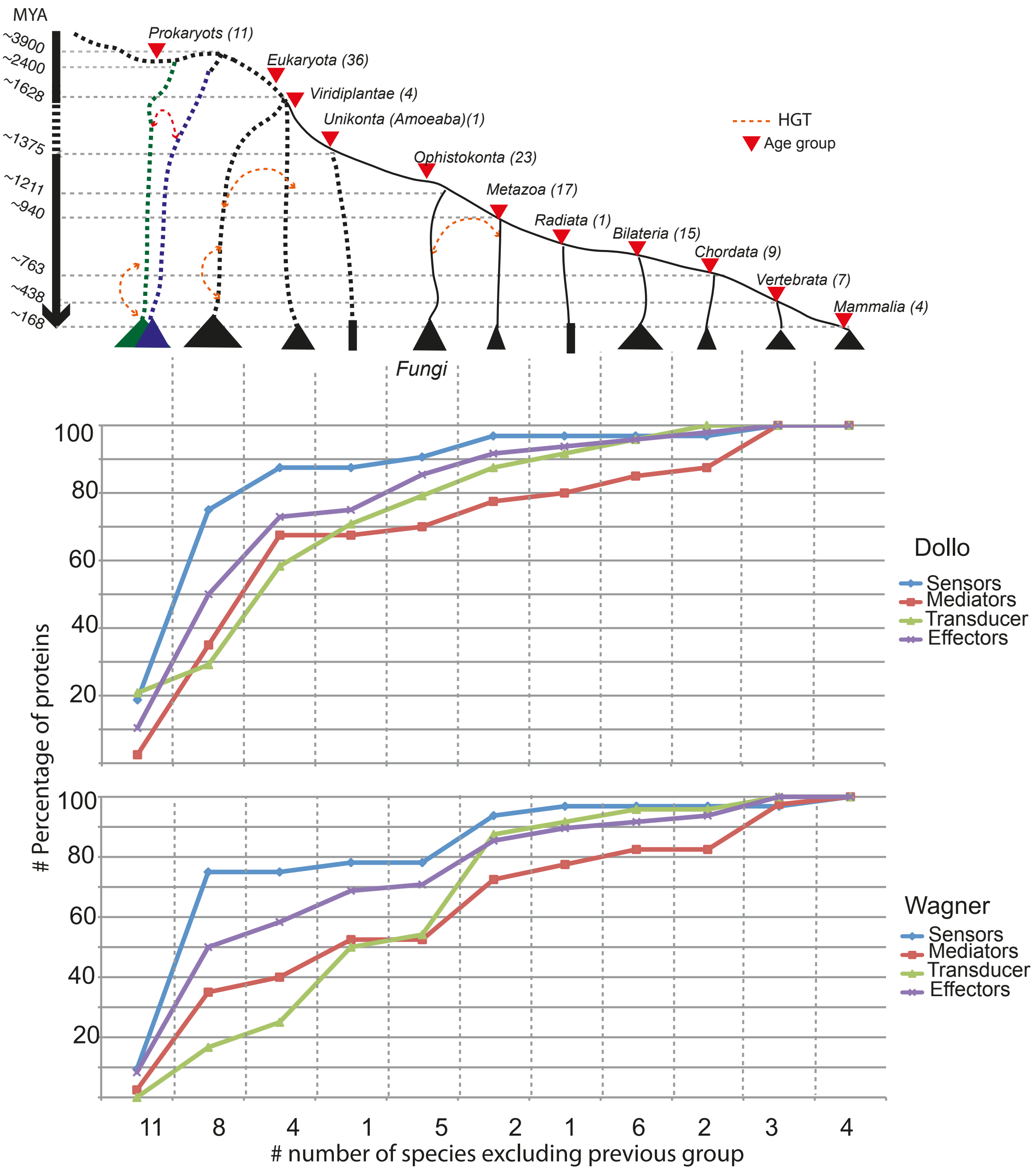


Figure 30b. Relative presence of DDR proteins according to a four-tier classification: sensors, mediators, transducers and effectors; considering the aggregated frequencies obtained by Dollo and Wagner parsimony. Red dotted arrows represent Horizontal Gene Transfer (HGT) events between phylogenetic groups.

Many genes were acquired in animals. By means of orthologs identified, novel incorporations such as the effectors MDM2 and hnRPK took place when animals emerged. MDM2 controls degradation of hnRPK, a p53 cofactor that plays key roles in coordinating transcriptional responses to DNA damage [117]. At the same age emerged MTA2, which forms a complex with the NURD protein in the repair of stalled forks (Figure 33), and the transducer UBR5, that interacts with TopBP1 [313].

In the *Bilateria,* the effector RBBP8 (CTIP), KAP1 (mediator and transducer) and the mediator NR4A2 were integrated in the network. RBBP8 is ubiquitinated in a BRCA1-dependent manner so CTIP, instead of being targeted for degradation, associates to chromatin and participates in the G2/M checkpoint control [314], while the NR4A nuclear orphan receptor has an essential role for DNA-PK-mediated phosphorylation in DBS repair [315].

Within bilaterians, similarly to fungi, extensive gene losses (blue boxes, Figure 17) have taken place as observed by the absence of orthologs in *C. elegans* and *D. melanogaster*, while they are present in ancestral relatives like annelids (segmented worms) or basal species as *N. vectensis* (*Cnidaria*) or even *Placozoa* (*T. adhaerens* -Tad- the most primitive animal).

The newest incorporations occurred in the chordates, where the urochordate *C. intestinalis*, incorporated two proteins, RNF168 (mediator and transducer) and SOX4 (an effector). The transcription factor SOX4 is required for the activation of p53 since it enhances its acetylation by interacting with and stabilizing p53, thus blocking its MDM2-mediated ubiquitination and degradation [116] (Figure 33). RNF168 is recruited by RNF8 to amplify the ubiquitination and recruit other proteins into the foci [99]. This way, the most recent evolutionary time incorporating functions is the vertebrate split, where sensors, mediators and effectors were added (as transducers got settled in the previous phylogenetic age (Figure 30a)). Vertebrate-specific genes are the effector MDM4, mediators Abraxas, MDC1 (also a sensor), RAP80, and PALB2, being central proteins to the different foci complexes occurring at damaged chromatin (Figure 32). On the other hand, TERF2 is involved in telomere maintenance [316] and interacts with Apollo and other proteins to protect telomeres from replicative damage [317].

Although most of the sensors emerged in ancient eukaryotes (Figure 30a, Table 6, ST7 in Annex), important proteins from all classes have been identified along eukaryotic evolution in different age groups. Thus, novel proteins in plants were the effectors EME1, DNA2L, PIAS1, the mediators BARD1, MRI40, TIPIN and XRCC4, and proteins showing overlapping functions like NBN (having assigned the four classes) ATRIP (sensor and mediator), ATM (sensor and transducer), and RMI1 (mediator and transducer). Other shared functions are represented by SLX4 (mediator and effector), a component of a complex involved in the resolution of Holliday junctions [318] in homologous recombination pathways, and by 53BP1 (sensor and mediator) a hallmark protein for foci formation [319] lately involved in end-resection [320] and which plays a fundamental role in DBS sensing and repair. HERC2 (mediator and transducer) is a crucial protein of the foci, where ATM phosphorylates this protein, thus stimulating its interaction with RNF8. HERC2 is also needed for RNF8 to promote UBC13-dependent poly-ubiquitylation of H2A-type histones **[98]** (Figure 32).

In summary, at the ancient eukaryotic stage, most of the proteins are orthologs of effectors and sensors, although a large expansion of mediators and transducers are also incorporated at younger ages. From this point on, sensors, effectors and mediators are getting incorporated at a steady pace, while transducers expand largely reaching convergence at the *Metazoa* group.

4.9 DDR AND POSTTRANSLATIONAL MODIFICATIONS (Figure 8, Panel G)

In addition to the fast Poly (ADP-Ribosyl)-ation of histone tails conducted by PARPs [321], various post-translational modifications such as sumoylation, phosphorylation, ubiquitination, acetylation or neddylation occur on DDR. Considering that conservation of a given protein performing a particular PTM does not necessarily indicate a conservation of the PTM, the "conservation" concept here should be taken as a proxy for potential existence. Keeping this in mind, we have analyzed the target-modifier pairs involved in PTM in the human set.

From the DDR human proteins, 53 are known targets of 24 modifiers (within set) (Figure 17: coloured triangles by the protein names, ST8, Annex) where some modifiers can be also targets (i.e.: UBE2T, CHK1/2, and PRKDC). If we plot the presence of modifiers by age and by the modification they exert it is noticeable that the potential to assemble the four PTMs were already on place very early on evolution. The incorporation of the remaining genes follows a step-wise manner, with the exception of phosphorylations (Figure 31).

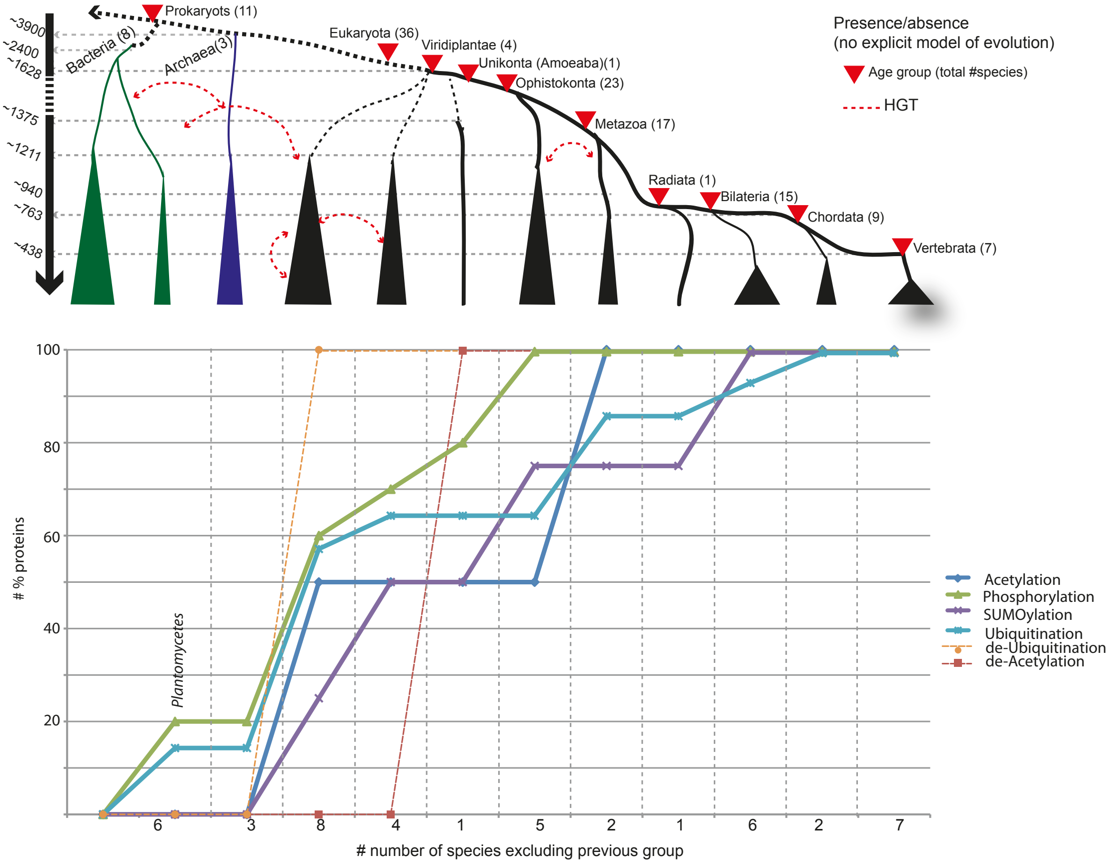


Figure 31. Relative presence of DDR proteins according to the PTMs they exert (acetylation, phosphorylation, SUMOylation, ubiquitination, deubiquitination and deacetylation) considering the aggregated frequencies obtained by hierarchical clustering. Red dotted arrows represent Horizontal Gene Transfer (HGT) events between phylogenetic groups.

a) Modifiers are traceable to early eukaryotes

The oldest group where orthologs of human proteins known to have modification activity have been identified (with ubiquitinase UBE2T) is *Bacteria*. In contrast, the newest ubiquitination potential activity probably acquired in evolution belongs to the *Chordata* taxonomic group (MDM2). Phosphorylations are by far the largest of the groups containing proteins, but since *Opisthokonta*, no potential phosphorylation capabilities have been identified. Within the human dataset (*Hsa*-118), 10 proteins are known to phosphorylate: the kinases ATM, ATR, CHK1, CHK2, MK03, PLK1, PRKDC, TAOK1, WEE1, and MK2. Interestingly, homologues for PLK1, TAOK1 have been found in bacterial *Planctomycetes*. Most of the kinases emerged at the ancient eukaryotic time, except ATM and CHK1/2 (but never after the metazoan split) indicating that phosphorylations were likely settled early on evolution, especially when no further proteins with phosphorylation capabilities were included after fungi in DDR (Figure 31). Ubiquitination is also old and it is widely represented in our dataset by BRCA1, CUL1, CUL4, HERC2, MDM2, RAD18, RBX1, RN168, RNF8, TRIPC, UBC13, UBE2T, KAP1 and UBR5 (ST8, Annex). Proteins with this function have been incorporated along the evolution of ancient eukaryotes till chordates. Homologs of UBE2T and UBE2N (also known as UBC13) have been identified in the *Planctomycetes* representatives, indicating that this function is very ancient. Deubiquitinating enzymes are represented only by BRCC36, which was present in early eukaryotes.

Although less studied, sumoylation has become an important process regulating networks. In DDR, orthologs with this potential function have also been incorporated until bilaterians. Proteins from Hsa-118 known to sumoylate substrates are MMS21, PIAS1, PIAS4, and KAP1. While MMS21 is detected in ancient eukaryotes, PIAS1 appeared on plants, and PIAS4 and KAP1 in Bilaterians. Thus, in general terms, this PTM has appeared late on evolution.

On the other hand, acetylations and deacetylations seem to be of old origin. The two acetyltransferases in this study, MYST1 (histone acetyltransferase which may be involved in transcriptional activation) and KAT5, both emerged in ancient eukaryotes. Regarding deacetylations, the deacetylase SIR1 appeared in amoebozoans.

b) Proteins representing target-modifier pairs for PTMs were present in ancient eukaryotes

In total, we compiled 99 target-modifier pairs (including auto-modifications) with experimental evidence in our dataset as registered by the literature (ST8, Annex).

Twenty-five pairs are “ancient pairs” as they appeared at this particular age, from which eleven pairs appeared simultaneously in the same species (being all of them phosphorylations by ATR and PRKDC). Within the same age group but in different species, in eight cases the target is older than the modifier (phosphorylations by PRKDC, ubiquitnation by RNF8, acetylation by KAT5, and sumoylation by MMS21) while in six cases the modifier is older than the target (phosphorylations by ATR, and ubiquitinations by CUL1/4 and RBX1).

At the time of the plants split, the phosphorylation repertoire produced by ATM was already on place, and later in *Amoebozoa*, phosphorylation regulating cell-cycle checkpoints was established (by the presence of the CDC25A-CHK2 pair). In fungal species additional phosphorylations were included, represented by CDC25C with CHK1/2 and MAPK2, and the youngest age where a potential pair was acquired is at *Bilateria*, with the acquisition of KAP1 (described to sumoylate and ubiquitinate).
The remaining pairs include cases where targets are older than modifiers and vice-versa spanning different ages, sometimes really distant, such as the cases of the ATR-PALB2 or ATR-Abraxas, being the kinase from ancient eukaryotes and the other two proteins from *vertebrata*, or the pair UBR5-TOPB1, where the E3 ubiquitin-protein ligase from *metazoa* ubiquitinates the DNA topoisomerase TOPB1, that was already present in ancient eukaryotes.

Overall, at the end of the ancient eukaryotes age, pairs of interactors that could potentially interact for phosphorylations, ubiquitinations, sumoylations, and acetylations were already on place.

Regarding the age of appearance of a given PTM, out of the 99 target-modifier pairs, in 32 cases the target is more ancient than the modifier, while in 33 pairs the modifier emerged before the target. In the remaining 34 pairs, both proteins appeared in the same age (see Table 11 below and ST9, Annex). If we analyze the pairs considering the species phylogeny, in 41 cases the target is present in a species more ancient than the modifier is, in 40 pairs the modifier is older than the target, and in 18 cases, both target and modifier emerged in the same species. These results show that for the DDR interacting pairs analyzed, there is no particular trend for the modifier to appear in evolution before the target, nor vice-versa. Interestingly the age group differences showed by the members of the pair are larger when modifiers are older than targets (Table 11).
Besides, for those interacting pairs emerging in the same species or in very distant ages, we have not detected any clear relationship with the results obtained in the gene-trees of these proteins.

	Target older than modifer		Modifier older than target		Modifier/Target from same age group			
Type of PTM	Number	Age group differences *	Number	Age group differences **	Same age	Age group	Same organism	Age group
Phosphorylation	26	18s/5m/3b	24	4s/6m/14b	9	7(2), 2(4)	13	9(2), 2(3), 1(4),1(5)
Ubiquitination	3	2s/2b	6	3s/2m/1b	4	(2)	0	
Sumoylation	2	1s/1m	2	2b	2	(2)	0	
Acetylation	0		1	1s	1	(2)	0	
deAcecetylation	1	1m	0		0		0	
TOTAL	32		33		16		13	

* In all cases the oldest age where a Target was incorporated was at Opisthokonta (Fungi and *M. brevicollis*)
** In all cases the oldest age where a Modifier was incorporated was at Opisthokonta (Fungi and *M. brevicollis*)

Table 11. Summary of PTMs. There are a total of 99 pairs of target-modifiers in the Hsa-118 set (5 pairs are automodifications so they have been excluded from the table). 'Age group differences' indicates jumps of distances among the proteins: s is small (one age group of difference), m is medium (two age groups of difference) and b is big (more than 2 age groups of difference). 'Age group' 2 is Ancient Eukaryotes, 3 is Plants, 4 is Unikonta (Amoeba), and 5 is Opisthokonta (Fungi and *M. brevicollis*). Regarding the pairs from the same age group, phosphorylations are the most frequent PTM and nearly all of them were already present in ancient eukaryotes, though the last incorporation was in *Opisthokonta*.

c) Conservation of modified residues

Regarding the modified residues, for some targets, there is precise information available in the human DDR proteins (ST8, Annex). In total we have compiled 114 residues (in 58 target proteins) where 85 are phosphorylations, 13 are ubiquitinations and 10 sumoylation sites (although for both modifications half of the residues are not precisely identified), 4 acetylations, and 2 deacetylations. The most modified protein in the dataset is BRCA1, which is phosphorylated, sumoylated and ubiquitinated. H2AX is also heavily and widely modified, although some specific residues that are ubiquitinated are not precisely identified.

When we checked the potential conservation of a particular residue in multiple alignments, the overall conservation of the sites was usually poor and in addition the neighbouring residues were also poorly conserved. In many cases, the regions were poorly aligned, especially in serine-rich regions, and frequently clusters of serines were found. Therefore, there are potentially many sites that could be phosphorylated.

4.10 CONSERVATION OF PROTEIN COMPLEXES IN DDR AND MAPPING TO THE HUMAN NETWORK (Figure 8, Panel G)

Next, we mapped the evolutionary conservation of known protein complexes into the human network (Figure 32-34). For illustrative purposes we have distinguished three sub-networks: the general repair (Figure 34), the replicative stress (Figure 33), and the DSB (Figure 32), although many components play roles in alternative pathways. For instance, BLM, H2AX, FACD2 participate in the three schemes. There is also an overlap among proteins involved in mismatch repair and replicative stress (Figure 33).

Overall, the most conserved pathways are those involved in DNA repair and replicative stress. For instance, the NER pathway in both flavours (global and transcription coupled, Figure 34), since all of its components were present at least in ancient eukaryotes, while some of them were present even in prokaryotes (XPB/ERCC3, XPD/ERCC2).

Noticeably, there are extensive losses of genes in fungi and invertebrates (fly and worm) within the whole network (Table 12) despite the high degree of conservation. For example, in the BER pathway, PARP1 and PARP2 are lost in several lineages, being the former missing in all fungi, while the latter is missing in all fungi except *B. dendrobatidis*, and also in nematodes and arthropods. XRCC1 is also missing in all fungi, as well as in *C. elegans*. This suggests that the BER pathway, as we understand it from the remaining species, must either be accomplished by a different set of genes in these particular organisms, or alternative pathways are fulfilling the role of these components.

In the case of NER, the protein CUL4 is absent from *S. cerevisiae* while present in *S. pombe*. It should be mentioned that both fungi are very divergent. XPA is missing in plants, and the regulation of XPC by SIR1 [322] should have appeared later in evolution, as SIR1 appeared in the *Amoebozoa*.

The NHEJ pathway senses damage by DBS (Figure 34). The core components are Ku70, Ku80, PRKDC, and LIG4, which are present in ancient eukaryotes. Later additions to the complex appeared in evolution, for instance, XRCC4 was incorporated in plants and XLF late in fungi. XLF is missing in *C. elegans*, and XRCC4 is missing in all fungi and in our nematoda/arthropoda representatives, while present in early bilaterians.

Regarding the regulation of these core components (Figure 32-34), MYST1, ATR, ATM, SIR1 and MTA2 appeared sequentially in evolution and at different phylogenetic ages, suggesting an increase of the complexity in the regulation of core components. RMI1/2 appeared in plants, while TOP3A and BLM are ancient from prokaryotes. This suggests that the assembly of complexes involving the RMI1 protein, like RMI1/TOP3A/BLM [323] or RMI1/RMI2/TOP3A [324] may have occurred at the plants stage. Thus, maybe in ancient eukaryotes the dissolution of Holliday junctions could be achieved without the RMI proteins or additional proteins could accomplish that function. Other proteins important for Holliday junction resolution are SLX1 and SLX4 [57]. While the former is ancient but missing in plants, the latter has been acquired in animals but has been lost in nematodes and arthropods. At telomeres, HR is performed by a complex that includes SMC4/5 and MMS21 [55], all of them form ancient eukaryote origin, while its regulation via TERF2 is recent in vertebrates [316].
Besides its known involvement in telomere maintenance and protection [325], novel roles have been assigned to the exonuclease Apollo, which interacts with MUS81, MRE11 and FACD2, all proteins of ancient origin. Therefore a potential regulation of the core proteins is conserved.

The DDR in replicative stress is also quite conserved, and it is noticeable that most of the core components assembling at the fork were already present in eukaryotic times. Additional components involved in dynamic properties controlling the cell cycle were incorporated at the *Opisthokonta* split (Figure 33), so these functions regulating cell cycle dynamic events are old but not ancient.

The damage response to DSB is initially mediated by the action of Ku70/80 and PRKDC (Figure 34). In the event this fails, an alternative way mediated by ATM takes over (Figure 32) exerting a complete variety of actions at the breaks. This constitutes the less conserved and more recent incorporation in the network, and is by far the less conserved module, where many of the proteins involved in protein complexes are of metazoan origin (i.e.: HERC2, 53BP1), chordate origin (RNF168) and vertebrate origin (MDC1, F175A/Abraxas, RAP80). Moreover, a substantial part of the less recent components have been lost (Figure 32) in invertebrate and fungal species indicating that the particular function accomplished by the BRCA1-A complex should be conducted by another set of proteins in these organisms. Most probably, the proteins in this alternative set are quite different in domain architecture to the ones detected in the other species, since in the case of BRCC45 (one the most ancient proteins in this complex), the BRE domain contained has not been detected in any of the five fungi in our study; and regarding BRCC36, the JAB1/Mov34/MPN/PAD-1 ubiquitin protease domain, though identified in all the fungi, it is found in DDR-unrelated proteins, such as in some 26S proteasome regulatory subunits in yeast.

DDR subpathways	Known Complexes	Proteins Acting on the complex	General framework	Orthologs lost in fungi	Orthologs lost in invertebrates	Orthologs lost in fungi and invertebrates	Orthologs lost in plants
DBS breaks	H2AX-MRN-ATM-53BP1-PTIP-RAD18	KAT5	break sites	Nbs1	53BP1		RAD18, KAT5
	H2AX-MDC1-RNF8-RAD18-UBC13-HERC2		break sites		RNF8		
	UBC13-RNF168-HERC2	PIAS4; TRIP12	break sites				
	RAP80-Abraxas-BRCC45-BRCC36-BRCA1-MERIT40-BARD1	UBE2T, RBBP8; PIAS1	break sites	BARD1	RBBP8	BRCC45,BRCC36, BRCA1,MERIT40, UBE2T	
	PALB2-BRCA2					BRCA2	
	MDMX-MDM2	ATM	cell cycle arrest				
	MDMX-MDM2; CUL1-FBX031	KAP1	cell cycle arrest		FBX031,KAP1		
Replicative stress	SMARCAL1-RPA-RAD17-911-ATR-ATRIP-TOPB1-TIPIN-TIMELESS-CLASPIN	UBR5	break sites	SMARCAL1			
	SMARCAL1-RPA-RAD17-911-ATR-ATRIP-TOPB1-TIPIN-TIMELESS-CLASPIN	MTA2-NR4A2			MTA2 (lost in *arthropoda*)		
	BTRCP-SKP1-CUL1-RBX1		cell cycle progression				
	BTRCP-SKP1-CUL1-RBX1		cell cyclle delay				
							PLK1
	DDB1-CUL4-CDT2-CDT1-PCNA		block to re-replication	CDT1			
MMR	MSH3-MSH2-MSH6	EXO1, (SLK1-SLX4)			MSH3, SLX4		SLX1
	MLH1-PMS2	EXO1					
BER	PARP1-PARP2-XRCC1	NR4A2		PARP1, XRCC1		PARP2	
NHEJ	Ku70-Ku80-PRKDC-XRCC4-KLF-LIG4	Artemis, SIR1, ATM, MYST1		MYST1		Artemis	
NER-global	RAD23B-XPC	SIR1					
NER-transcription	CSA-DDB1-CSB-CUL4-RPA						
NER-common	XPB-XPA-XPD-XPG-RPA-XPF-ERCC1						XPA
HR	RMI1-TOP3A-BLM; MUS81-EME1	EXO1, DNA2L, Apollo, FANCD2, MRE11, FANCM		EME1, FANCD2			
telomeres	SMC5-SMC6-MMS21	TERF2					
Holliday junction	SLX1-SLX4				SLX4		SLX1
Adducts removal	TDP1-TDP2			TDP2			

Table 12. Summary of protein complexes and losses in lineages. Complexes are depicted in Figures 32-34.

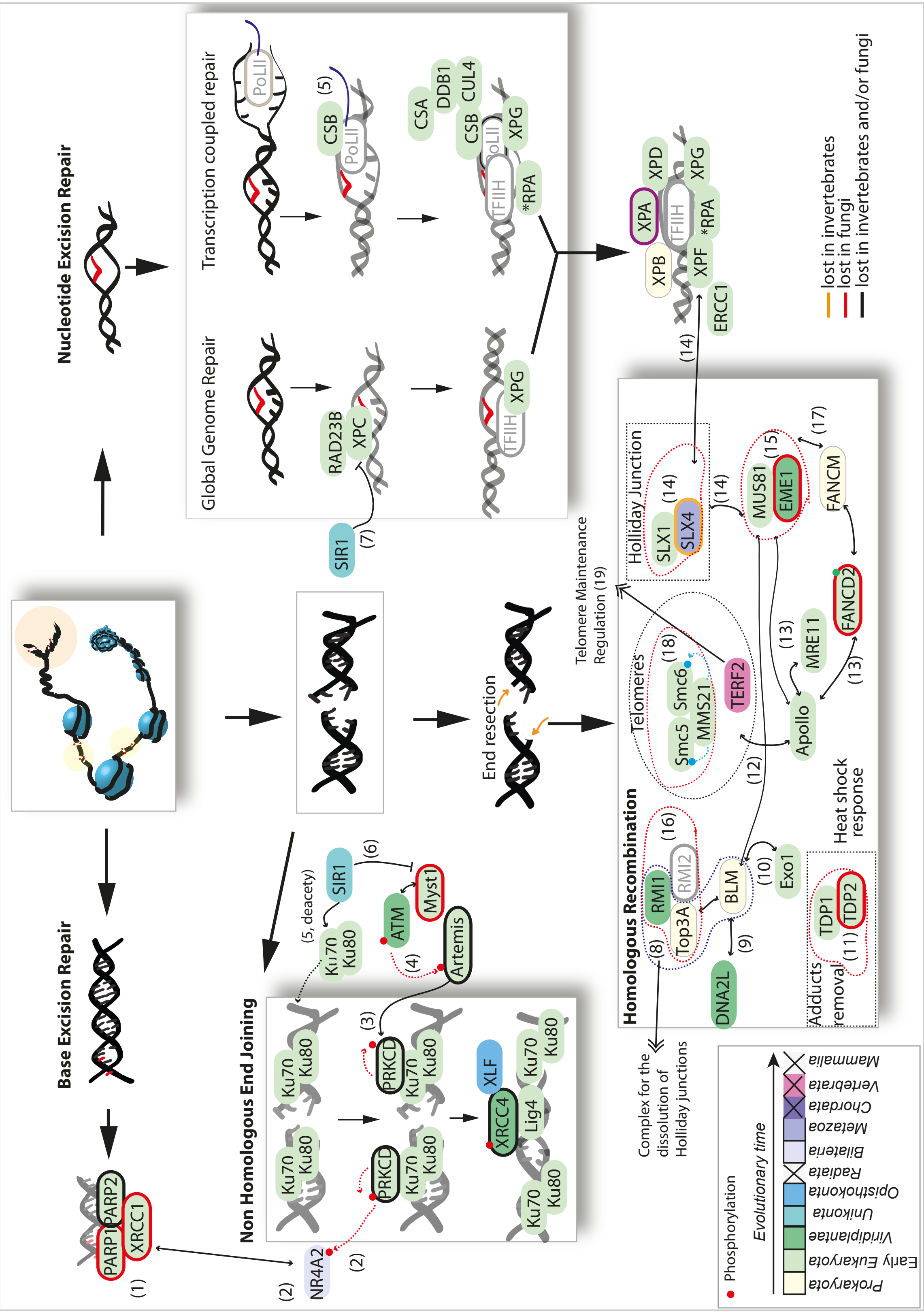


Figure 32. DDR by ages: General repair pathways: BER, NHEJ and HR. Colours refer to the evolutionary point where the different proteins in the human pathway emerged in evolution according to our results. Red dots represent phosphorylations. Proteins with bold margins represent losses or absences in specific taxa. *RPA: complex formed by three proteins RPA1, 2 and 3.

Bibliographic references: (1)[326]; (2)[315] (3)[327]; (4)[328]; (5)[329]; (6)[330]; (7)[322]; (8)[323]; (9)[61]; (10)[331]; (11)[332]; (12)[317]; (13)[325]; (14)[57]; (15)[58]; (16)[324]; (17)[333]; (18)[55]; (19)[316]

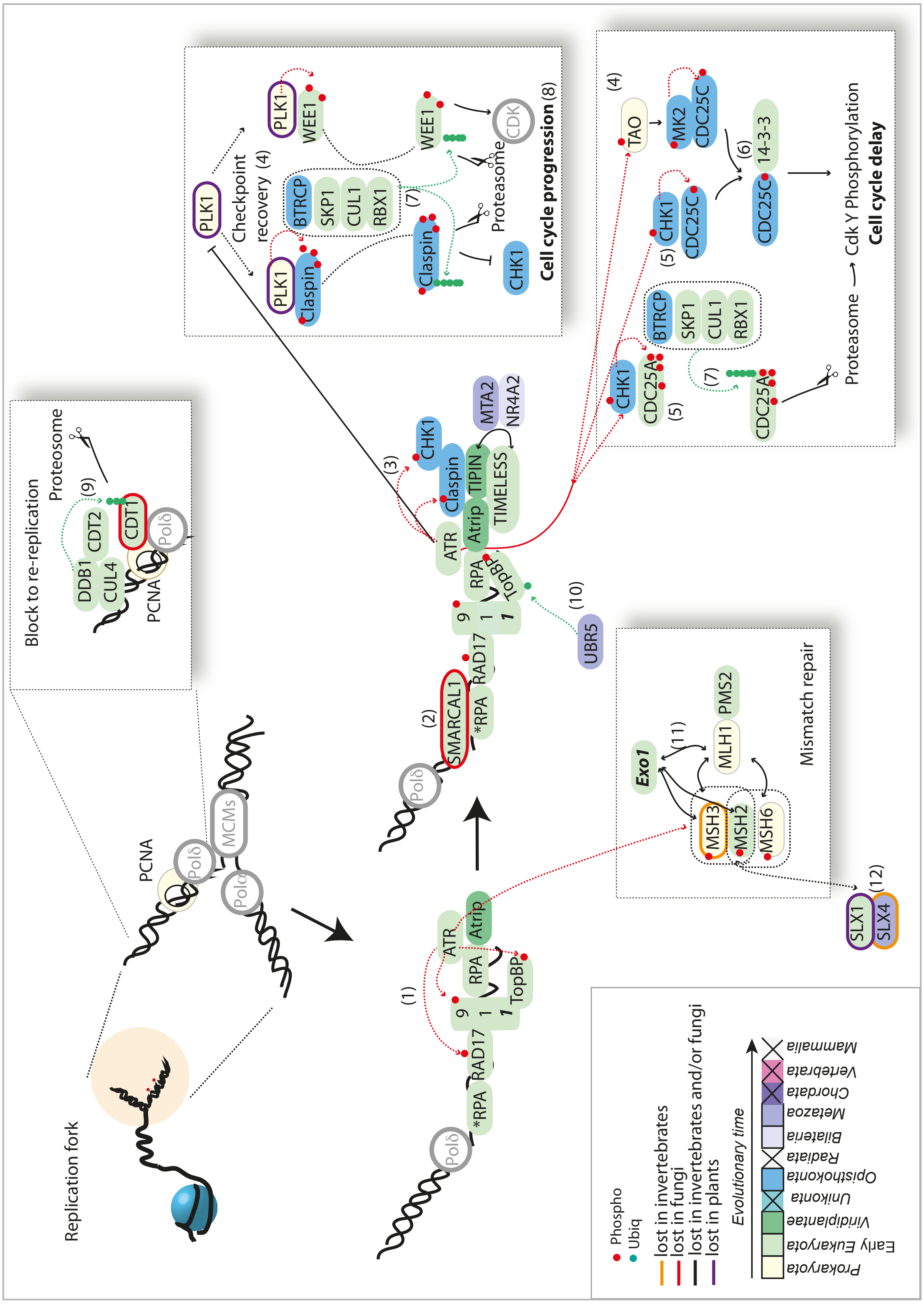


Figure 33. DDR by ages: damage at replication forks. Colours refer to the evolutionary point where the different proteins in the human pathway emerged in evolution according to our results. Red dots represent phosphorylations and green ones are ubiquitinations. Proteins with bold margins represent losses or absences in specific taxa. Bibliographic references: (1)[120]; (2)[121] (3)[7]; (4)[122] (5)[7]; (6)[123]; (7)[124]; (8)[126]; (9)[129]; (10)[313] (11)[37]; (12)[334].

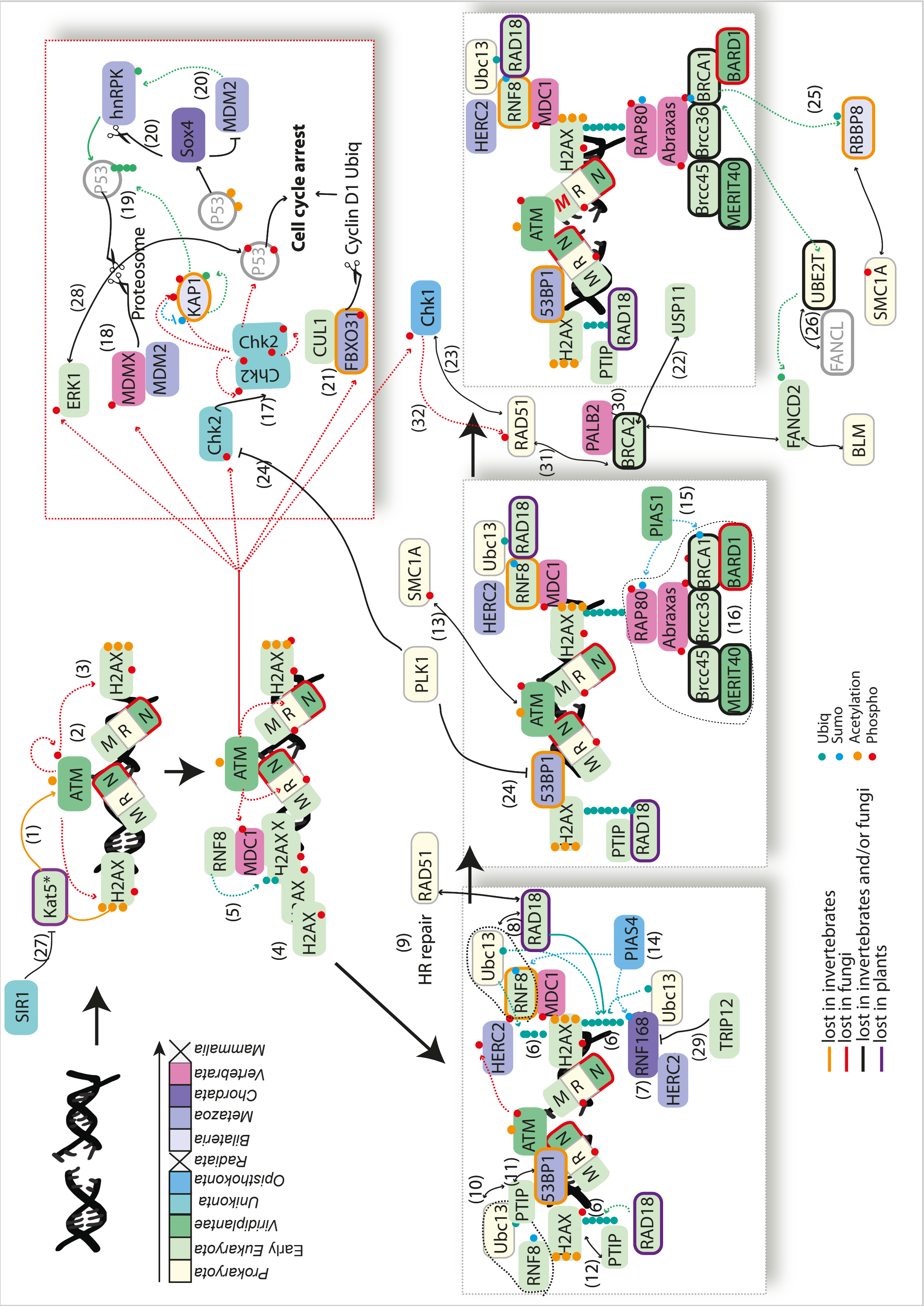


Figure 34. DDR by ages: Double Strand Breaks. Colours refer to the evolutionary point where the different proteins in the human pathway emerged in evolution according to our results. Coloured dots represent various PTMS: phosphorylations, ubiquitinations, sumoylations and acetylations. Proteins with bold margins represent losses or absences in specific taxa.

Bibliographic references: (1)[91]; (2) [90]; (3)[93]; (4) [96]; (5) [97]; 6)[98]; (7) [99]; (8) [100]; (9)[100]; (10) [101]; (11)[102] (12) [103]; (13)[104,105]; (14)[112]; (16) [107-109]; (15)[113] ; (17) [114]; (18)[115]; (19)[116]; (20)[117]; (21)[119]; (22)[110]; (23)[111]; (24)[128]; (25)[314]; (26)[335]; (27)[330]; (28)[336]; (29)[337]; (30)[59]; (31) [338]; (32)[339].

5 DISCUSSION

The emergence and conservation of the DNA Damage Response (DDR) network is still an open issue [48] receiving wide attention.
In terms of evolutionary inference no systematic analyses of the network as a whole have been so far conducted, with the exception of particular parts of the pathway (i.e. chromatin modifiers [263]). Reasons for this rely on the difficulty to establish a consensus set of DDR components, and more importantly, due to the fact that the DDR network involves the concerted action of different sub-networks in a dynamic context [5] driven by post-translational modifications (reviewed in [83]), which adds another layer of complexity to identify crucial role members. Therefore, while consensus genes involved in well-defined sub-networks such as repair are relatively straightforward, a consensus set of the remaining DDR proteins has not been defined yet, probably due to the fact that some components exert different functions in alternative sub-networks.

Due to large diversity in the Biology of the different species, it is expected to find gross variations in the assembly of this concerted summa of pathways. Nevertheless, given that a proper DDR is crucial for cell viability, it is likely that the core network components should have appeared early in evolution and may have suffered a tight evolutionary control regarding expansions or losses.

However, network analysis along the evolutionary timeline is often difficult due to the underlying limitations of ancestral reconstruction in highly divergent organisms, where some knowledge regarding general evolutionary rules is required. In this regard, recent work conducting large-scale and organism-wide analyses has provided a useful framework to understand particular evolutionary features along the evolutionary time using the concept of gene ages [227]. This work provides evidence that the distribution of evolutionary rates of protein coding genes is universal and uses a model whereby the distributions of loss rates are the same for genes gained and lost over a long time interval [227], making amenable to screen the DDR pathway along a wide evolutionary scale regardless the large evolutionary distance.

Thus, in the context of fixed "age" groups, we can draft the evolution of the whole network, to further analyze the acquisition of certain genes in particular organisms.

The use of sequence homology to extend the DDR pathways is informative

When studying networks, the usual procedure is to focus in a particular model organism. This approach presents several issues for evolutionary inference. For instance, the use of single species does not provide a comprehensive view of the evolutionary process because species-based bias is likely to have an effect. So, to transcend the restrictions of using a single species, it is desirable to analyze a given network in different organisms covering as much as possible the amplitude of the tree of life. In this spirit, we have compiled "seed" datasets of DDR components of four model organisms from different taxa, including genes that are well described in literature (Table ST1 Annex). Literature-mining approaches, despite being trustable, are sometimes limited to conduct large-scale analyses, and as expected, the overlap among our collected DDR literature-generated datasets from different organisms is low (Figure 12a).

To circumvent this, we have used one of the most useful approaches to extend pathways in organisms with available genomic sequences -but unknown functional data-, which is the use of homology-based annotation transfer [164,340,341]. This helps in discriminating between universal and non-universal components of a system [342].

It should be noted that even when orthology does not necessarily correspond to equal function, it is widely accepted that orthologous proteins likely retain function along evolution as compared to paralogs [343], and this is the assumption we are using here. This limitation have likely a minor impact as most of the Hsa-118 proteins lack paralogs; therefore, neo-functionalizations or sub-functionalizations due to species-specific gene duplications is expected to be low in the DDR framework of protein families as opposed to other families that suffered extensive gene duplication events, like the RAS superfamily of proteins [157].
On the other hand, both LSEs and gene losses, pinpoint to aspects of the network that are likely to be critical in organism-specific functions **[344,345]**.

As we focus on evolution towards *H. sapiens*, we have neglected particular LSEs that may indicate particularities in other important lineages, especially *Viridiplantae*. This point, although deserving much attention is not the primary focus of this study. Nonetheless, our preliminary findings identified some proteins (in principle not widely acknowledged in recent literature to be DDR involved in human, but described as DDR-related for other species) in model species that showed significant homology to those from other species having annotations related to DDR (Table ST3 Annex), increasing the final overlap among the four species (from 4 to 13 proteins, Figure Figure 12b), while as expected, lineage-specific proteins decreased in all the groups. These findings indicate the presence of proteins in all four organisms that have potentially more importance than expected in DDR processes. This awaits experimental confirmation.

Inferring orthology is still a challenge

The most accurate manner to assess orthology is by careful phylogenetic [193,194] inference. This usually involves the reconstruction of a phylogenetic tree using a given model to describe the evolutionary relationships among the sequences and species involved. However, the large demands of time and computing power needed to generate reliable trees have limited their use to single gene families or datasets of moderate size. Moreover, phylogenetic trees are difficult to automate for genome scale data, and the topology of the tree is strongly dependent on the tree building method chosen. Besides, in some occasions pair-wise comparison approaches have outperformed more complex algorithms that use sophisticated tree reconstruction and reconciliation approaches [198].
Because of all these drawbacks and considering the relative size of our data set, we decided to use a pair-wise automatic clustering method for orthology detection and further phylogeny for selected proteins.
In this regard, this study has some limitations derived from the quality of the available data, since most of the chosen organisms have only draft assemblies (see table ST2,

Annex), and while some of them are in the first versions, the proteomes are still incomplete.

This fact has various implications: first, in absence of well-annotated sequences, is it very difficult to identify the truly characterized proteins, so a conservative orthology detection strategy seems to be adequate to minimize obtaining false positives, with the caveat that most likely we will be missing true proteins. On the other hand, this strategy also is affected by missing data, which may result in false negatives.

Secondly, some orthologs may not be detected since in most cases data are protein predictions. And finally, we should consider that as any automatic method, the computational pipeline devised here might yield false positives. Despite these caveats, we decided to include all these species to have the widest range available set of deep branching organisms.

As mentioned in Results 4.3 and as shown in Table 3, several proteins detected as orthologs were excluded from the study since they misaligned in the MSA of orthologs and lacked the characteristic domains of a given DDR protein. The majority of these discarded sequences shared sequence similarity with the human protein only in certain regions where promiscuous domains were usually found. Though many of these discarded sequences are likely to be protein fragments (due to the use of draft genomes in this work), few of them might be evolutionarily related to the human protein and although sharing a common origin, might have diverged enormously and now perform different functions.

Our data indicate that our strategy has worked reasonably well as we have detected more than 50% of the Hsa-118 orthologs in 52% of the organisms (24 species out of the 46 used in this study), while the number of organisms in which more than 75% of the Hsa-118 orthologs were detected are 9 (representing 20% of the total 46 species): *N. vectensis* (a basal eukaryote, where 90 DDR proteins were detected using human as seed)*, C. teleta, B. floridae, D. rerio, X. tropicalis, G. gallus, O. anatinus, M. domestica* and *M. musculus.*

Age groups do not necessarily reflect the DDR proteins evolutionary history

Genome evolution involves extensive loss and gain of genes, and therefore the propensity of gene loss values of individual genes differ widely [346,347]. These events have a strong influence in the evolutionary history of different families, and we can find different properties for diverse genes within the same species. Within the *Hsa*-118 genes, we have identified 16 protein families containing homologous proteins (see Results 4.5.1) that could be affected by different evolution rates and could have very different and complex evolutionary histories. Assuming the universality of the model proposed by Wolf *et al.* [227], we have approached the emergence of the network in a broader –cruder– sense in a gene content framework spanning several species that can be formalized as a phylogenetic profile (Figure 17). This points to the appearance of DDR components in evolution and when these components acquired the potential to get assembled, without making any assumptions about the underlying evolutionary process. For this particular purpose, phylogenetic profiles (widely used in alternative contexts like the study of protein correlated evolution [148], physical interactions of proteins [348] and protein annotation [349]) are amenable tools to analyze the evolution of gene content to delineate protein families evolution [301]. Besides, phylogenetic profiling is a well established method for predicting domain associations and functional relations and physical interactions between proteins [348]. Moreover, this method has been used to annotate genes and infer gene and protein networks [349].

Subsequently, DDR orthologs were assigned to particular gene age groups (Figure 10) that have been also established previously [227], where each gene age group indicates the maximum phylogenetic depth where a particular gene can be found using sequence similarity as a proxy of homology (see methods). Although the concept of age group has been widely explored [227,302], a general concern is that there is not a single optimal method to define the age of a particular gene [227,302]. In fact, different estimation strategies may produce different ages for the same proteins due to the complex evolutionary histories of proteins and to constraints of the methods [302]**.**

Our analyses of gene content indicates a general good agreement using the three methods, where around 10% of the proteins of the *Hsa*-118 is of ancient origin traceable to prokaryotic organisms and present in the three supra-*Phyla* (Figure 22). From this point on, the largest expansion of DDR is also of ancient origin and likely happened at the time of the *Eukaryotic* split (about 1628 MYA), where the DDR network grew to about 50-70% of its current components. So, beyond the *Metazoa* group, both hierarchical clustering and Dollo algorithm provide very similar patterns (Figure 22) while the Wagner's method estimates a smaller number of proteins by age. Our results are in agreement with previous work where it's been reported that Dollo parsimony produces overall older protein age estimates that Wagner parsimony [302] (Figure 22). As this method assumes that each protein family was only gained once, false positives and horizontal gene transfers (HGT) can inflate protein ages. Wagner however produces younger ages on average [302]. This is an effect of how each method accounts for gene losses and gains as in Cluster and Dollo a family's origin is the most recent common ancestor of all species in which it is observed, regardless gene losses, whereas Wagner's parsimony allows multiple gain and loss events in an ancestral family reconstruction. Therefore, gene losses greatly influence the outcomes of the method.

To complicate things further, there is unavailability of functional data for most of the selected species that are in draft state. Besides, The use of organisms with particular life-styles such as parasites and pathogens, as well as the nucleomorphs of *G. theta* and *B. natans*, has expectedly produced large absences in the taxonomic distribution of some proteins. Thus, there is a combined effect of genome incompleteness (overestimation of gene loss), the difficulty to identify the nature of a gain (overestimation inflated by HGT) coupled to the inaccuracy in protein annotations for most of the draft genomes, and the difficulty to identify highly divergent proteins using conservative assumptions as the ones used in this work, which could explain the differences in the methods obtained in the content of proteins from early eukaryotes (Figure 22). These effects decrease after the *Metazoa* split (~940 MYA)) where the trend follows a steady pace for all the three methods showing minimal differences that can be disregarded.

Taxonomy- and gene-trees in DDR are difficult to reconcile

In addition to gene gains and losses, gene transfers have played an important role throughout evolution. This has been particularly true for the prokaryotes [350], although recent work revisits the impact of HGTs in eukaryotes and how particular life-styles can enhance them [351]. Given the fact that most of the DDR genes are of ancient or very old origin (Figure 22), they may likely have been gained by this mechanism, and in fact this may reflect the large discrepancies observed between the species and genes trees. Other possible explanation is the incorrect placement of given species in the taxonomic tree, and then, certain studies (for instance the one by Wolf *et al.* [227]) based on this taxonomical distribution should be revised to account for the evolutionary model of particular gene families.

The "gene ages" approach is particularly relevant to identify the discrepancies observed between the species and gene trees, alleviating the uncertainty on the evolutionary model that shaped some of the DDR protein families (i.e. gains and loses). This is especially important when deep phylogenies are particularly prone to poor resolution due to large sequence divergence where it is frequent to find cases where a set of homologous genes or proteins may be quite useful in resolving species or genus-level relationships, but it might be quite poor at resolving phylum-level relationships due to poor conservation or short sequence length. In this regard, considering that the species tree used here shows an extensive degree of polytomies at deep nodes **[268]** (Figure 10) we have not attempted to reconcile the estimated gene trees and rather have used the concept of gene ages to infer the timeline of evolutionary emergence of DDR proteins.

Nonetheless, we have conducted pylogenetic analyses in selected proteins, and the overall results of the phylogenetic analysis indicate that the evolutionary history of most families may be more complex than expected (Table 10), as it is very frequent to observe arthropod, fungal, and worm sequences grouping close to ancient eukaryotes instead of their assumed more related lineages, which indicates that maybe certain genes in these organisms are older due to HGT events and have been later acquired by these organisms.

Despite the discrepancies observed (especially regarding arthropods and worms), the reconstruction of a phylogenetic tree describing the evolutionary relationships among the sequences and species involved is very useful since it allows a better assessment of gene orthology than Blast-based methods. Also, they help detect wrongly assigned orthology. In addition to Family13 explained in results (Figure 24), in Family8, the MYST1 and KAT5 orthologs detected by the computational pipeline seem to be wrongly identified, since according to the tree, which is generally well supported, MYST1 from *N. gruberi, C. parvum*, *C. falciparium*, plants and *D. discoideum* could be KAT5 instead, having its emergence in ancient eukaryotes, where the *P. tricornutum* ortholog has been correctly detected (Figure 28). Other example of errors in the orthology detection can be seen in the case of Family9, comprised by DCR1A, DCR1B and DCR1C (see figure STt38 in Annex).

In those cases where information about in-paralogs was available, these were included in the initial trees to help clarify the position of paralogs and orthologs, which is especially useful when running phylogenies of multigene families. Examples of these trees are Family13, comprising MLH1 and PMS2 (Figure 24); Family6, comprising Cullin-1 and 4 (STt19); or Family 9 (STt38).

A widely used repository of evolutionary relationships is ENSEMBL COMPARA. Although COMPARA considered PTIP (PAXI1) and MDC1 as being part of the same family, we finally did not regard Family7 as such because these sequences only share the promiscuous BRCT domain and beyond this, there is no detectable sequence similarity between these proteins. As seen in the domain analyses (Figures 18 and 19), the BRCT domain is shared by many protein families, so it is not a good candidate to infer homologous relationships.

Regarding gene-trees order, interacting or functionally related proteins have been frequently shown to have similar phylogenetic trees [352] because of co-adaptation processes (compensatory changes between the two proteins) or due to similar evolutionary pressure on the sequences [353], and such seems to be the case of RAD17 and TOPB1. The two proteins are part of a complex involved in ATR-dependent checkpoint activation at stalled replication forks.
However, beyond the aforesaid example and others mentioned in Results 4.7, we have not detected a clear trend in the origin and taxonomical distribution of the modifiers-target pairs analyzed in this work (see results 4.9), likely due to the fact that those modifiers interact with many other proteins, the same as the targets, and also probably due to those proteins also having a role in other pathways.

Protein domain content and functional analogy: common misinterpretations

Another important contributors to genome evolution are the protein domains (structurally defined modules within the protein, arranged in a particular order) that usually point to precise functions. These modules are believed to constitute major evolutionary units, as phylogenetic analysis based on protein domain content has shown to be comparable to standard phylogenetic methods based on molecular markers (such as rRNA [354]). Moreover, domain shuffling is a great source of functional variability and has been extensive in the evolution of protein families [355], where different arrangements of the same protein domains can achieve alternative functions [356]. Aravind *et al.* [342] showed an increase in the complexity of domain architectures in proteins involved in chromatin structure, suggesting that this will have an effect in the number of interactions between proteins, as combining multiple domains in a polypeptide allows for more combinatorial interactions.
Also regarding protein modules, presence of common domains in evolutionary unrelated proteins likely points to functional analogy and is frequently misinterpreted as orthology.
An example is illustrated by the proteins 53BP1 (human) and Crb2 (fission yeast) generally considered as orthologs because they exert the same function [240]. However, besides the BRCT domain (spanning a short stretch of length in both sequences), these proteins are evolutionary unrelated (therefore cannot be referred as orthologs). Moreover, no identifiable orthologs for 53BP1 are found for fungal species in our analyses (Figure 17). Another example of proteins evolutionary unrelated but

considered as orthologs because they exert the same function is illustrated by XRCC4 (modern eukaryotes) and LIF1 (*S. cerevisiae*) [357], however these proteins are not similar at all at the sequence level. Interestingly, XRCC4 has been completely lost in plants, fungi, fly and worms, although it is present in basal animals (*T. adhaerens*) (Figure 17). Similarly, RAD50 is often related to the SMC family [30], although in *H. sapiens* RAD50 lacks the SMC domain that is partially present in other species (*O. sativa, S. pombe*, etc) (Table ST4 Annex).

It is widely accepted that the acquisition of additional domains may confer novel capabilities to a protein [358] and that the balance of losses/gains of domains affects the functional repertoire of species [359]. Different domain content in orthologs has functional effects. For instance, the XRCC1 protein is well conserved in eukaryotes (although completely lost in fungi and *C. elegans*). However, in plants, the XRCC1 orthologs lack the N-terminal domain (required for DNA polymerase β binding) and the C-terminal BRCTII domain (necessary for DNA ligase III interaction) (Figure 19). Interestingly, they do retain the BRCTI domain that mediates interaction with PARP1 and PARP2 **[360].** We could then speculate that as plants do not contain DNA polymerase β and DNA ligase III genes **[258]** the BER pathway must be remarkably different in this lineage, probably a simplified version.

Our results indicate that in the DDR framework is quite usual to find the same domain scattered in evolutionary unrelated proteins, examples are BRCT repeats (present in MDC1, PARP1, TOPBP1), the FHA domain (CHK2, RNF8, MDC1 and NBN), and the PWWP domain (MSH6, ATM - only *in A. thaliana*) (Figure 19, Table ST4 Annex). These could have implications when inferring functions.

It is possible however that the same domain appears in evolutionary unrelated proteins belonging to the same protein complex, that brings us back to the aforementioned example of LIF1 and XRCC4. LIF1 contains a “XLF” domain, which is present in a different protein of the same complex, the NHEJ1 protein. However, LIF1 does not have the common domain shared by all XRCC4 orthologs.

Another interesting example is the case of DNA-PK that contains four domains (NUC194, FAT, PI3_PI4_kinase, and FATC). Although an ortholog was detected in a non-model plant (Figure 17, Ppa - *P. patens*), no equivalent proteins were detected in in the remaining species of the lineage, therefore a deep analysis should be done in this genome to discard any contamination or artifacts. Similarly, in *Arthropoda* an ortholog was detected in *T. castaneum*, but seems to be missing in the remaining species, including model species. The fact that no ortholog has been detected in either *C. elegans* or fungi has questioned the dispensability of this protein in the NHEJ pathway [361]. However, it is likely that unrelated proteins containing the same domains in different arrangements would perform the same function when a canonical ortholog is missing in one of the species, such as the aforementioned case of XLF.

Therefore, in absence of clear orthologous relationships, a sensible way to detect suitable candidates to fulfill a functional role is to screen for the presence of common domains in different proteins belonging to the same complex.

When obvious domains are not detected, a possible way of inferring function is by searching for distant homology based on conserved regions. In this regard, we did not identify Pfam domains in few sequences within our Hsa-118 set of DDR orthologs, but these had conserved regions with hits to domains known to be involved in DNA repair, replication fork arrest or checkpoint processes (Table 6). Interestingly, certain hits revealed similarities between the structure of the 26S proteasome and the BRCA1-A complex. A further characterization of these domains will elucidate their potential role in DDR events.

Regarding the enrichment analysis of domains involved in DDR, the results obtained suggest that most of these domains bind DNA, are from ancient origin, and are shared among the analyzed species. Interestingly, the number of statistically significant enriched domains shared between *H. sapiens* and *A. thaliana* was higher than the ones shared by *H. sapiens* and *S. cerevisiae* (Figure 21), even though yeast are more closely related to human than plants.

The majority of DDR components of the network are traceable to ancient Eukaryotes

Overall, our results indicate that the core components of the network are of ancient origin traceable to prokaryotes, that big expansion of DDR components is also ancient and traceable to *Eukaryotic* origin (1628 MYA) as indicated by the presence of DDR components in *Emiliania huxleyi* (*Haptophyceae*) the oldest eukaryote with a planktonic free life-style. This is in agreement with age enrichment analyses for the human DDR set of genes [302], were the *Hsa*-118 set was significantly enriched at the *Eukaryotic* age regardless of the algorithm, age group classification or the database screened. In contrast, the *Hsa*-118 set is significantly underrepresented in mammalian genes and do not follow the general trend of the rest of the protein families for human proteins (Table 9).

Therefore, most of the DDR components were available at the *Metazoa* split (~940 MYA) with little incorporation afterwards. The youngest age where a DDR-related protein was incoporated corresponds to *Vertebrata* (~438 MYA) group.

We next interrogated if the incorporation of functions followed the path of the general components. Due to the unfeasibility of experimentally characterizing all the proteins, computational methods are suitable alternatives although not exempt of problems [184]. Then we extended known functions from well-described species to orthologous proteins from the remaining genomes. However, the definition of **"function" should be taken cautiously here**. If we understand the DDR network as a dynamic framework where many components and pathways are largely regulated [5,74] by a concerted action of many different proteins, from enzymes to regulators, then there is a wide plethora of functions making functional assignments very difficult to define. In cases like this, standard GO analyses, useful in different contexts, provide very little information (ST6, Annex) (enrichments for DNA repair, DNA metabolic process, and response to DNA damage stimulus, which are expected).

Subsequently, for consistency purposes, we have used a broader classification widely used in the field of DDR where proteins are assigned to four main supra-functional classes: "*Sensors*" (S), "*Mediators*" (M), "*Transducers*" (T) and "*Effectors*" (E) (ST7, Annex) **[5].** In this schema, sensors and effectors would represent the extremes of a directed pathway, and the addition of mediators and transducers would incorporate functions to increase the complexity of the network and would allow its cross-talk with other pathways.

Most of the ancestral components represent the extremes of a directed pathway

In our dataset, the most populated class is *Effectors* followed by *Sensors*, *Mediators* and finally *Transducers* (Figure 29**,** ST7, Annex), and before the eukaryotic split, most of the identified orthologs correspond to *Sensors* and *Effectors* represented by about 10-20% of the proteins, so there is potential for the ancestral pathway to be enriched in the two extreme parts of the network (Figure 30a and b). Interestingly enough, the members of the bacterial phylum *Planctomycetes* are the only prokaryotic members sharing homologous sequences of *Transducers* typical of the eukaryotic kingdom (Figure 30a**,** dashed green line). More precisely, homologous sequences to the ubiquitinases UBC13 (UBE2N) and UBE2T, and the kinase domain of both PLK1 and TAO kinases. These results could provide some lights to recent work published for these particular organisms that revisit the evolutionary origins of this particular phylum in the context of eukaryotic evolution [362-364], that is a subject of intense scientific debate [365].

Still in ancient groups (at the *Eukarya* time of split), orthologs of the intermediate classes were incorporated, although most of the functions were *Sensors* and *Effectors*. At the *Metazoa* split all the classes represented by orthologs where steadily incorporated.

The incorporation of components with PTMs potential to the DDR network is traceable to prokaryotes and early eukaryotes

Post-translational modifications [5] are essential regulators of the dynamics of the DDR. Regarding the emergence of these potential functions, proteins with phosphorylation capability were settled early on evolution, as well as orthologs of proteins showing ubiquitination activities (Figure 31, ST9, Annex) and this activity has been incorporated since ancient eukaryotes towards *Chordata*. Accordingly, orthologs for deubiquitinating enzymes (BRCC36) are also present in early eukaryotes, and as pointed out previously, interestingly enough it has been lost in fungi and invertebrates (Figure 17**,** Figures 32-34, Annex). Although less studied, sumoylation is also present in ancient eukaryotes (MMS21) and has also been incorporated several times in evolution (PIAS1 in plants and posterior duplications in bilaterians, PIAS4). Finally, acetylations and deacetylations are ancient as well. MYST1 and KAT5, the two acetyltransferases in this study, appeared in ancient eukaryotes, while the deacetylase SIR1 appeared in amoebozoans. While estimating the evolutionary timeline for individual components is feasible, a complete different matter is to infer whether these components acquired capabilities to exert precise functions. Although orthology [366] (a phylogenetic term) is widely used as a surrogate for functional equivalence, in the case of post-translational modifications this correspondence has to be taken more cautiously. Besides, the lack of comprehensive data for many organisms makes this assumption general [367-370] where the sequence conservation of a given site correlates to functional conservation. In our hands, this approach, at least for the DDR genes was unfeasible given the high disorder present in the particular regions where modifications take place (using as reference the human sequences (ST8, Annex)) being poorly conserved in the identified orthologs. This lack of conservation does not preclude the function, as additional regions of the proteins could be also modified. An interesting feature of the human DDR modifications is that various modifiers can act upon the same DDR proteins in different contexts, in agreement with previous work addressing the co-evolution of various PTMs in several eukaryotic species (from fungi to human) where it is described that proteins that are regulated by one kind of PTM are likely to be regulated by a different one, but not necessarily at the same time and not essentially effecting the same behaviours of the protein [370].

In our analyses of 99 target-modifier pairs (Table 11), the numbers of cases where the target is more ancient than the modifier and *viceversa*, are nearly identical. Intuitively, it should be expected otherwise, where modifiers appear later to modify existing targets. Moreover, in cases of interacting pairs, where the modifiers are older than their targets, the age distances within the members of the pair are way larger (Table 11, 14 cases where there are more than three ages of distance) than distances observed when targets are older than modifiers (only 3 cases).

This could be due to the fact that these modifiers have alternative targets (not considered here) of older origin or because originally these modifiers exerted different functions more related to metabolism (most of the PTMs in the DDR pairs are phosphorylations, which are also the most common proteome-wide PTMs identified experimentally [371] and are highly linked to metabolism), and along evolution their function varied and diversified through the modification of their sequences and the addition of domains [372]. These modifications might have led to a function more related to regulatory and signaling processes, and to the capability to interact with new targets. It would be interesting to make a global genome analysis of targets-modifiers to check whether the differences observed in their ages of origin are DDR specific or is something common in the PTMs.

As discussed above, regarding the appearance of novel components along evolutionary lineages, our data suggest that general post-translational regulatory mechanisms such as phosphorylation and ubiquitination may have been incorporated in evolution before sumoylation. Thus, proteins involved in phosphorylation (ATM, ATR, check point kinases, etc.) and those proteins involved in ubiquitination (UBE2N, CUL1, TRIPC, etc) are mainly in blocks II and III of the cluster (Figure 17). In the case of sumoylation (PIAS1 and PIAS4), up to date the core machinery does not contain any proteins with such functionality.

The distribution of these PTM proteins along the pathways suggests that metabolic reactions are previous to the regulation activities, which would be expected since the former are simpler in terms of components and can be modeled easily.

Reconstructing the pace of DDR pathways from cellular organisms to modern eukaryotes, where gene losses shaped the network

How do the aforementioned findings fit within the different sub-networks (pathways)? Our careful literature-based compilation of the human network (Figures 32-34) serves as a comprehensive framework to map evolutionary relationships. When evolutionary conservation is mapped into the human DDR network, it is noticeable that there are proteins of ancient origin traceable to *Prokarya* present in all the sub-networks: general repair (mismatch, HR, NER, yellow Figure 34), replication stress (kinase domain of PLK1 and TAO, yellow Figure 33) and DBS (ubiquitinases, SMC1A, and kinase domain of PLK1, yellow Figures 32). The general repair and replicative stress pathways are the oldest ones (yellow and light green colours in Figures 32-34), although in the case of replicative stress essential components important for regulatory processes have been added to the pathway at different evolutionary splits (in plants, fungi, bilaterians and metazoans (Figure 33). In the case of general repair, the most conserved sub-networks are BER, NER, and HR (in order of conservation, Figure 34). The less conserved pathway is the DSB. Even if the core sensing components in this particular module are quite conserved, it is noticeable that most components involved in regulation are young proteins (i.e.: the cell cycle arrest components are mostly of animal origin), while at the foci RAP80, MDC1, and Abraxas are exclusively of vertebrates (Figure 32).

There are also striking losses of key parts of the network, especially in plants, fungi and invertebrates (fly and worms, blue boxes Figure 17) where fundamental proteins in complexes have been lost in several species. Examples of this in particular pathways like BER, HR or NHEJ (Figure 34) are the PARP proteins, PRKDC, or Artemis. PLK1 is missing in plants and MSH3 in invertebrates (Figure 33). As we focus on *H. sapiens* genes, we are surely neglecting specific lineage expansions that could accommodate the same functional roles. In this regard, it is known that in certain species the loss of a single important gene can dramatically affect entire sub-modules of regulatory networks (i.e., yeast) [373]. In metazoans, the availability of genomic sequences of basal animals (cnidarians and placozoans included in this study) shows that nematodes have lost several modules of regulatory networks [374]. An illustrative example is the RNAi system; while it is highly developed in plants (and in some fungi and animals), it has been lost in other lineages (entirely lost in *S. cerevisiae*, but present in *S. pombe)* [375]. In animals, this system has suffered multiple partial losses; although nematodes have it complete, some insects and vertebrates lack different parts of the network (i.e.: the siRNA replicating encompassing the RNA-dependent RNA polymerases, that is however present in the basal *Branchiostoma* [375]).

Analogously, in the case of the DSB complexes at the *foci*, *S. pombe* contains more orthologs of DDR than *S. cerevisiae* (Figure 17), and there are losses of DDR complexes in both fungi and invertebrates, being the most striking case the absence of almost half of the components of complexes forming at DSB (Figure 32), where 5 proteins of the same complex (MERIT40, BARD1, BRCC36, BRCC45, and BRCA1) are missing in invertebrates and/or fungi (Table 12). This suggests the existence of at least two independent losses of this sub-network, one in the line leading to fungi, and one to invertebrates. As these proteins participate in extensive post-translational modifications, it is possible that these functions have an indirect back-up mechanism from alternative functionally comparable systems (like in the case of the RNAi, where chromatin-level gene silencing [376] and post-transcriptional protein degradation systems [377] perform functions that are related, like modulating the levels of gene products).

Another plausible explanation to these apparent losses could be the existence of LSE in protein families [344], where estimations indicate that 20% (yeast) to 80% (plants and vertebrates) are comprised of families of lineage-specifically expanded families. Examples are proteins exerting their functions at the termini of signaling cascades (i.e.: E3s and MAP kinases) [342], the Ub/ubiquitin-like proteins conjugation network [358] and phosphorylation networks on yeast [378] although most of these processes are development related.

Therefore, this suggests again that functions are not necessarily linked to orthology, as functional analogs (genes with different evolutionary origin but performing the same function) can compensate this. An example is the NHEJ protein XRCC4, without identifiable orthologs in yeast, but where functional analogues have been identified: LIF1 in *S. cerevisiae* and Nej1 in *S. pombe*. In this regard, the use of traditional model systems in comparative genomic studies to target function may not be desirable.

D. discoideum is an intriguing species. Although its phylogenetic placement is still unresolved, recent proteome-based phylogeny suggests that *amoebozoa* diverged from the animal–fungal lineage after the plant–animal split [379]. The very important kinase CHK2 and the deacetylase SIR1 have orthologs in this species. The kinase CHK2 is central to the induction of cell cycle arrest and apoptosis by DNA damage [114] and interestingly, the amoebozoan lacks CHK1, whose function may be performed instead by the expanded RAD53 family in this organism [380].

Another interesting feature detected in this species is the fact that out of the 65 DDR orthologs detected in this species, 27 proteins are markedly longer than their human counterparts, suggesting this organism might have longer genes than the majority of the other species.

It is important to notice that the absence of orthologs in certain organisms does not mean that these do not have the corresponding DNA damage repair systems; for example, NHEJ has been reported in certain bacteria [233], though we have not detected orthologs of proteins in this pathway in the chosen organisms from that kingdom. Even so, there is emerging evidence of functional crosstalk between bacterial NHEJ proteins and components of other DNA-repair pathways [381].

The shape of the human DDR network as inferred from evolution

So far there is a lack of a formal representation of the human DDR network. In the pathways repositories, only partial networks can be found, especially the repair fraction (Reactome, Kegg, etc). However, as we have explained in this work, there are intricate relationships among the different sub-networks. One of the most known canonical DDR pathway representation is the one described by Harper and Elledge [74]. As there is certain overlapping of components involved in related pathways, it is important to incorporate these elements into a more comprehensive way. For instance, it is desirable to illustrate how components of the sensing part are also participating in related pathways like cell cycle checkpoints and DNA repair.

We have manually reconstructed the DDR network to account for the data available in the literature (Figure 32-34). A description by pathway follows:

Mismatch repair

Our results agree with previous studies showing the MutS and MutLα components of the MMR system are widely conserved among the species studied **[382]**. The MSH3 and MLH1 sequences appear to have varied less along evolution since orthologs for these proteins have been identified in the three kingdoms of life, while MSH6 is detected in Bacteria and Eukaryotes and orthologs of MSH2 and PMS2 have been detected only in Eukaryotes. It is noticeable that, though highly conserved, several species in our set lack some of the MMR proteins, like the *apixomplexa* and *arthropoda* representatives or *Caenorhabditis elegans* where the MSH3 protein is missing **[47,383]**.

MSH2, MSH3 and MSH6 mismatch repair proteins are all highly similar to the bacterial MutS protein, and their domain composition seem to be well conserved along evolution. However, the MSH6 orthologs in higher eukaryotes possess an additional N-terminal region comprising a PCNA binding motif, a large region of unknown function with a globular PWWP domain and a nonspecific DNA binding fragment. This PWWP domain binds double-stranded DNA, without any preference for mismatches or nicks, whereas its apparent affinity for single-stranded DNA is about 20 times lower **[384]**.

Interestingly, while the fungi and *arthropoda* representatives and the *C. elegans* MSH6 orthologs do not exhibit an N-terminal PWWP domain, *A. thaliana* and *O. sativa* MSH6 present an N-terminal Tudor domain which probably share functional properties with the PWWP domain of human MSH6.

Base excision repair

Eukaryotes have several functional analogs of bacterial BER enzymes, and the mechanism of BER is similar to that of prokaryotes. However, eukaryotes have additional specific BER enzymes. To date, poly(ADP-ribose) polymerases (PARP) and XRCC1 have been identified as eukaryotic-specific enzymes **[245]**.
PARP1 and PARP2 are activated by SSBs and catalyze poly(ADP-ribose) (PAR) synthesis at DNA breaks triggering local chromatin relaxation and recruitment to the damaged site of repair factors with strong affinity for PAR, such as XRCC1, where this protein operates as a scaffold that interacts with and stimulates the enzymatic activity of other components of the BER machinery **[326,385]**.

Both PARP1 and PARP2 are involved in DNA damage sensing and signaling when single strand break repair or BER pathways operate. PARP1 uses NAD to add branched ADP-ribose chains to proteins and functions as a DNA nick-sensor in DNA repair and as a negative regulator of the activity of DNA polymerase β in LP-BER **[386]** while a functional role of PARP2 has been found in the maintenance of telomere integrity [387]. Recently, both PARP1 and PARP2 were found to intervene in DSB repair since it binds to stalled replication forks that contain small gaps, where they mediate Mre11-dependent replication restart [388].

The members of the PARP superfamily present a modular architecture characterized by a conserved core responsible for the catalytic activity to which a variety of targeting and regulatory modules have been added. Although PARP1 and PARP2 C-terminal catalytic domains have the strongest resemblance among all the other family members, their N-terminal parts differ completely. The PARP1 DNA binding domain contains two zinc fingers while PARP2 presents a nuclear location signal and a functional DNA binding domain that targets DNA gaps but not nicks like PARP1. Also, PARP1 has a central BRCT motif, which is a protein–protein interacting interface found predominantly in proteins involved in the maintenance of genomic integrity and cell cycle checkpoint functions responsive to DNA damage [389]. Due to these differences it is argued that PARP2 is involved in a later step of the BER process than PARP1 and that they may have distinct DNA targets [390].

According to our results and to previous studies [391] both PARP1 and PARP2 are highly conserved from simple eukaryotes to human, but are absent in yeast and some fungi, as well as in our *apicomplexa* representatives. PARP2 seems to be less conserved than PARP1 since it was not detected in *Gallus gallus* (probably because this organism was in a draft genome state) and our selected *arthropoda*.

PARP domain composition appears to be well conserved along evolution, though in the case of PARP2 there are two N-terminal SAP domains in the plants orthologs not present in the other groups analyzed. Experiments have shown that in *Arabidopsis thaliana* both PARP1 and PARP2 genes are induced by DNA strand breaks and ionizing radiation. However, expression of the AtPARP2 gene was found to be induced by different types of environmental stress, which suggests an additional role for AtPARP2 that would be independent of DNA damage [360].

XRCC1 does no have any known enzymatic activity, but it can physically interact with other proteins involved in the SSB repair and BER pathways. XRCC1 interacts with DNA polymerase β through its N-terminal domain, and the central section of the protein is involved in the interaction with other proteins involved in the repair of SSBs. Also, XRCC1 contains two BRCT domains, BRCTI and BRCTII, located centrally and at the C-terminal part of the protein, respectively. The BRCTI domain is responsible for the physical interaction with PARP1 and PARP2 and is indispensable for their recruitment at SSB sites, while the BRCTII domain specifically interacts with DNA ligase III [41].

The XRCC1 protein seems to be well conserved in eukaryotes, with the exception of fungi and simpler organisms, though there are important differences between the plant orthologs and those found in other species. Plants XRCC1 orthologs lack the N-terminal domain required for DNA polymerase β binding and the C-terminal BRCTII domain necessary for DNA ligase III interaction, but retain the BRCTI domain that mediates interaction with PARP1 and PARP2. [360]. As in plants there are no DNA

polymerase β and DNA ligase III genes **[258]**, the plant SP-BER pathway must differ notably from that in other eukaryotes.
O. sativa XRCC1 protein binds ssDNA as well as dsDNA and also interacts with PCNA forming a complex, suggesting a different contribution of XRCC1 to DNA repair pathways compared to the mammalian BER system **[392].**

In fungi, this repair mechanism may be different to that in higher eukaryotes since no orthologs of PARP1/2 or XRCC1 have been detected in our fungal representatives (but for PARP2 in *B. dendrobatidis*). Previous works have described a role of several DNA N-glycosylase/AP lyases in in BER in *S. cerevisiae* [393], and of other genes such as the OGG1 in *C. albicans* [394].

Nucleotide excision repair

Though there is a conserved NER mechanism in all domains of life, there are notable differences between the eukaryotic and prokaryotic systems. In human, more than 20 proteins are known to be involved in NER **[395],** while bacterial cells require only three proteins (UvrA, B and C) to accomplish a similar effect in the much simpler prokaryotic NER **[396].** In spite of the clear functional parallels between the bacterial and eukaryal NER pathways, an independent evolution of these two systems has been postulated due to the lack of sequence homology between the bacterial Uvr proteins and the eukaryotic XP proteins.
In archaea, the scenario is intriguing since the NER system varies depending on the species. Most possess orthologs of the eukaryotic nucleases XPB and XPD, like *S. solfataricus* and Candidatus *Korarchaeum cryptofilum*; in addition, homologs of XPF and XPG have been detected in other archaeal species. Nevertheless, there are a few species of archaea with a NER machinery similar to the UvrABC system from bacteria, which could be explained by horizontal gene transfer events. Besides, there are also archaea with a mixture of eukaryotic and bacterial NER orthologs **[397]**.
The presence of detectable orthologs of eukaryotic NER proteins in the archaea has led to the hypothesis that the archaeal NER machinery is a simpler version of the eukaryal one, however some studies suggest this idea should be accepted with caution **[398].**

NER proteins are in general terms well conserved among eukaryotes, though there are some significant differences regarding particular phylogenetic groups. The fact that orthologs of certain NER proteins have not been detected in some lower phyla of the Eukarya suggests that this repair pathway may have acquired proteins and gained complexity relatively late in evolution.

Orthologs of CSA (ERCC8) have not been detected in *T. brucei*, *N. gruberi*, some fungi, *C. elegans* and in our *apicomplexa* and *arthropoda* representatives, which has made this protein to cluster in a different group than the other proteins in this pathway. Also, no orthologs of XPA have been detected in plants, thus these organisms may have other divergent unknown proteins playing this function.

In *P. falciparum*, besides XPA, the global repair XPC protein is also missing, suggesting that this organism may have a different mechanism for DNA damage detection in GG-NER (global genome repair).

In the other proteins analyzed in the NER pathway: ERCC1, XPA, XPB (ERCC3), XPC, XPD (ERCC2), XPF, XPG (ERCC5), RAD23B, CSA (ERCC8) and CSB (ERCC6), the domain composition of the orthologous sequences detected is practically identical in all the species analyzed.

Non-homologous end-joining

Originally, NHEJ was thought to be limited to eukaryotes since *E. coli*, the most and best studied prokaryote, is unable to ligate DNA ends. However, bioinformatics analyses lead to the discovery of a distantly diverged Ku-like gene and an ATP-dependent ligase (LigD) gene in the same operon in various prokaryotic genomes [399,400]**.** Later, a pathway similar to NHEJ was shown to function in some bacterial species (mainly those that form endospores**)** [401], indicating that this repair mechanism has been conserved in the course of evolution.

Apart from the scarce homology between the prokaryotic and eukaryotic Ku proteins, and the fact that the bacterial Ku homologue forms a stable homodimer similar in structure to the ring-shaped eukaryotic Ku heterodimer [401], the other essential agent of the bacterial NHEJ repair pathway, LigD, presents clear differences between prokaryotes and eukaryotes; unlike the eukaryotic DNA Ligase IV, the bacterial LigD is a single polypeptide that contains three domains: polymerase, phosphoesterase and ligase [381]**.** Besides, bacterial genomes do not encode an obvious DNA-PKcs homologue, thus, though the end-joining machinery in bacteria seems to be a direct ancestor of the NHEJ pathway in higher organisms, there are elements of end-joining in eukaryotic cells which either have arisen independently or have developed later in evolution [233,402]**.**

In the case of archaea, even though LigD 3’ phosphoesterase DNA repair homologs have been identified in some species (among them Candidatus *K. cryptofilum* and *M. acetivorans*) and were found to catalyze the same reactions of ribonucleoside resection and 3’-phosphate removal as the bacterial phosphoesterase domains [403], the DNA repair pathway in which they are involved (if any) must differ from that of bacteria since so far no homolog of Ku has been detected in the archaeal kingdom.

As mentioned above, though Ku70 and Ku80 homologs have been described in certain bacteria. In contrast, with our pipeline we have not detected any ortholog of the human proteins in our prokaryotic species, though a Ku domain is detected in the *B. subtilis* ykoV protein.

All the NHEJ orthologous proteins analyzed have a conserved domain composition and the slight differences found are probably caused by misannotated sequences and due to the proteins being predictions.

Given their critical role, one would expect NHEJ proteins to be evolutionarily conserved with relatively few sequence changes. However, while crucial domains are conserved, in certain proteins and species the sequence variations seem to be high enough not to allow us to detect more orthology relationships. For example, though we have detected few orthologs of the XRCC4 protein (which has made it cluster in a more modern group of proteins and thus appears as the latest addition to the NHEJ pathway), homologs of this protein seem to be present in most species among the Eukarya. In this regard, we have found the XRCC4 Pfam domain in proteins of *A. mellifera, D. melanogaster, S. japonicum, N. vectensis, C. neoformans, B. dendrobatidis, P. patens, C. reinhardtii, N. gruberi, C. parvum, P. tricornutum* and *E. huxleyi*.
Another example regarding XRCC4 is the case of *S. cerevisiae*, where Lif1 (Ligase-interacting factor 1) was identified as the homolog of human Xrcc4. Though the yeast protein has a XLF domain instead of the XRCC4 domain present in the XRCC4 orthologs, and despite the apparent low level of sequence identity, it highly conserves the primary binding site to DNA ligase IV **[357]**.

Curiously, a XLF domain is present in the higher eukaryotes NHEJ1 protein, and also in the *S. cerevisiae* homolog of this protein **[404]**, showing there is a limited repertoire of conserved domains involved in DNA repair. In this regard, we have identified NHEJ1 orthologs only from *Trichoplax adhaerens* to higher eukaryotes, though we have detected XFL domains in some lower eukaryotes as well as in bacteria and archaea.

Another protein showing differences among species is DNA-PK. Interestingly, an ortholog having the same domain composition as higher eukaryotes (NUC194-FAT-PI3_PI4_kinase-FATC) was detected in the moss *P. patens*, but no equivalent proteins were detected in the other plants representatives, and the same occurs in the case of *arthropoda*, where an ortholog of DNA-PK is detected in *T. castaneum* but seems to be missing in *D. melanogaster and A. mellifera.* Also, DNA-PKcs orthologs have not been found in yeast or *C. elegans* suggesting that the function of DNA-PKcs is not evolutionarily conserved and might be dispensable for NHEJ **[361]**. Some authors have suggested that the Mre11-Rad50-XRS1 (homolog of human NBS1) complex might act as a nuclease and play an equivalent role to Artemis-DNA-PK in organisms lacking these proteins, like yeast, and some plants and invertebrates **[405]**. This would be a system compensating for other mechanisms.

Homologous recombination

Few HR proteins are clearly conserved at the amino acid sequence level between prokaryotes and eukaryotes. Some examples are the recombinase, named RecA in prokaryotes and Rad51 in eukaryotes, and the RecQ helicase in prokaryotes and its eukaryotic counterpart BLM (Bloom syndrome protein).
While other eukaryotic proteins involved in this pathway have no clear homologs in bacteria and archaea, a number of them do have similar biochemical activities. In this regard, the MRN complex, consisting of the Mre11, Rad50 and NBS1 proteins, is in part equivalent to the RecBCD complex in prokaryotes, which, apart from acting in the repair of DSBs, is involved in bacterial conjugation and transduction, and thus in the horizontal transfer of genes **[406]**.

Regarding Rad51, homology searches have shown that most eubacteria possess only one recA gene, while many archaeal species contain two recA/Rad51 homologs (radA and radB) (though it seems none of the three archaeal species analyzed in our set of organisms have a radB gene), and eukaryotes have multiple members (Rad51, Rad51B, Rad51C, Rad51D, DMC1, XRCC2, XRCC3, and recA) **[407]**

Nevertheless, the domain composition among orthologs of these two widely conserved HR proteins (Rad51 and BLM) is dissimilar (see Results 4.4.2, Figure 19, Table ST4 Annex). The results obtained point towards an hypothetical emergence of the eukaryotic Rad51 caused by a combined evolution of bacterial and archaeal sequences, while the particular case of BLM represents fairly well the acquisition of novel functions due to diverse protein domain architectures reflecting substantial differences at the species level.

Regarding HJs resolution, the bibliography [57] describes an ortholog of human SLX4 in fungi. However, or pipeline has not detected any ortholog of this protein in fungi since the sequence in common between the human SLX4 and the proposed fungi ortholog is quite short, and there are no shared domains; thus, the fungi protein is probably a functional analog and was wrongly named as ortholog in the bibliography.

The ATM/ATR pathway

Regarding the MRN complex, homologs of Mre11 and Rad50 have been found in most organisms studied to date, and form as a stable complex, even in prokaryotes [408]**.** In eukaryotes, the Mre11/Rad50 (MR) complex also contains Nbs1 in plants and most modern eukaryotes, while in yeast, functional homologs of Nbs1 were identified in *S. pombe* [409] and in *S. cerevisiae*, where the protein was named Xrs2, to form MRX **[410]**.

As aforementioned, the IRIF forming proteins such as HERC2, 53BP1, MDC1 and the BRCA1-A complex RAP80 and Abraxas, or others involved in regulatory processes such as MDM2/4, Sox4, FBXO31, etc. were acquired recently in evolution and no orthologs of these proteins have been detected in ancient eukaryotes.

Interestingly, lower eukaryotes such as yeast do not contain obvious MDC1 orthologs. Instead, other conserved IRIF-forming factors such as the PAXI1/PTIP ortholog in *S. pombe* or the fungi orthologs of RAD18 are known to aggregate in regions with phosphorylated H2A **[411],** which is exclusively present at sites of DNA damage **[412].**

In regard to the structural components of the replication fork, no 9-1-1 complex (Rad9, Rad1 and Hus1) orthologs have been detected for the *apicomplexa* representatives (*C. parvum* and *P. falciparum*) and *T. brucei*. Besides, the RFA2 subunit of the RPA complex has not been detected by our pipeline in *P. falciparum*, where this protein is encoded by an unusual transcript that lacks the RAD52 interacting domain [413].

The presence of homologs of TIPIN, TIM and CLASPIN with similar functions in yeast indicates that the overall process of fork stabilization is conserved between complex and simple eukaryotes [48], though we have not detected these proteins in some of the fungal proteomes included in this study.

Concerning the regulatory activities, sumoylation was incorporated to the pathway by PIAS1, which seems to have emerged in plants and poses an interesting case of domain variation among orthologs from different phylogenetic groups. The PIAS family of proteins has a conserved zf-MIZ domain found in many ubiquitin E3 ligases and able to interact with the E2 enzyme, and a SAP (from SAF/ACINUS/PIAS) domain, which associates to DNA sequences of matrix attachment regions. Besides these two domains, the PIAS1 orthologs detected in plants have a PHD domain, which binds bromodomain proteins and thus contributes to the SUMO ligase function of the PIAS1 in plants [414]**.** Curiously, the same PIAS plant domain architecture is found in a putative SUMO ligase (GI: 296005550) in *P. falciparum*, which is the only case detected in an organism that is not a plant.

Control of DNA integrity: cell-cycle checkpoints

Fungi and animals have evolved cell cycle checkpoints to maintain genome integrity. Plants should also have surveillance mechanisms to enable them to arrest their cell cycle on DNA aggression or stress conditions, especially when, because of their immobile life style, these organisms are constantly exposed to putative DNA-damaging conditions (for example UV-B light).

The pathways through which ATM and ATR signal DNA damage to the cell-cycle machinery must be different to that in fungi and animals since orthologs of p53 and CDC25 are absent in plants according to our results and to previous works [415]. Also, plant homologues of CHK1/CHK2 have not been identified so far. However, in plants the replication checkpoint functions through phosphorylation of B-type CDKs, and it has been proposed that the CDC25-controlled onset of mitosis might have been evolutionarily replaced by a B-type CDK-dominated pathway, eventually resulting in the loss of the CDC25 gene [415]

On the other hand, an ortholog of the CDC25-counteracting WEE1 kinase has been detected in plants, and was shown to have an important role in arresting the cell cycle under DNA-damaging conditions [416]. Interestingly, PLK1 seems to be absent in plants, even though this kinase phosphorylates WEE1 in higher eukaryotes. Maybe no homolog of PLK1 exists in plants, though insignificant Pfam-A matches to a POLO_box domain have been obtained in a hypothetical protein in each of our four plants representatives (*A. thaliana, O. sativa, P. patens* and *C. reinhardtii*).

Examples of selected proteins: domain architecture and function

Here follow some selected examples of proteins where the variations in their domain composition lead to differences in protein function:

- When the addition of a domain involves a gain of function:

As explained in this work, the helicase BLM presents variations in its domain architecture in different organisms, for instance in *D. radiodurans,* where the C-terminal region of the sequence has three HRDC repeats that increase the effectiveness of the helicase activity [417], instead of the single copy found in the other orthologs.

- When the loss of a domain involves a differentiation of the function:

In modern eukaryotes the MMR MSH6 protein has an additional N-terminal region consisting of a PCNA binding motif, a globular PWWP domain and a nonspecific DNA binding fragment. This extra region most likely confers MSH6 a regulatory function in the NHEJ repair pathway [38]. However, this domain seems to have been lost in fungi, arthropoda and *C. elegans*, suggesting that these organisms may have other regulatory elements playing this role or other mechanisms of cross-talk between the MMR and NHEJ pathways.

- When domains are present in different proteins of the same complex in one organism, while being in the same protein in another species:

LigD, a fundamental component of the bacterial NHEJ repair pathway, presents clear dissimilarities with its eukaryotic counterparts; while the eukaryotic DNA Ligase IV has only ligation activity, the bacterial LigD is a single polypeptide that contains three domains: polymerase, phosphoesterase and ligase [381], which could play the role of other proteins in the NHEJ pathway in eukaryotes.

Summary of findings

As mentioned in the Introduction (section 5.1), evolution of organisms and DNA repair are highly interconnected because of the influence that repair mechanisms and pathways have in evolutionary patterns and mutation rates. Thus, there is a complex interplay between the need for evolvability and the need for fidelity of transmission of genetic information to the offspring.

Throughout evolution, those organisms with high levels of genetic variation had better chances to survive sudden environmental changes, but as organisms evolved more complex genomes, genomic instability became mostly detrimental and systems safeguarding the integrity of DNA increased.

The added complexity of the DDR in eukaryotes may reflect the requirement of a variety of mechanisms and pathways specialized for different conditions, such as the ability to signal and repair damage in DNA densely packaged into chromatin, the necessity to perform repair in distinct specialized cell types, the need for tight control to avoid ectopic recombination of repetitive DNA sequences, or to perfectly tune multifunctional components that are also used in other facets of genome and cell metabolism.

We have provided an extensive repository of human DDR proteins, carefully mined from literature and we have analyzed its evolutionary properties.

We have found that most of these genes are traceable to early eukaryots and some of the core components were already in prokaryotes.

Most of the ancestral components of the network are sensors and effectors, representing the two extremes of the pathway, which seems to have increased its complexity in between them including recruiting and signaling activities.

Our comparisons of gene trees versus species trees indicate that the evolutionary processes that have shaped the DDR network are way more complex than originally expected, likely involving massive HGT events.

Interestingly, most of the components have been lost at least more than twice during evolution, especially in fungi, nematodes and arthropods, where fundamental proteins in complexes have been lost in several species.

In certain cases of losses, we have found that functional analogs could compensate the absences, or alternative systems might intervene.

Protein domain shuffling has incorporated novel functions in various organisms, with impact in certain lineages, especially in plants.

A notable aspect of the distribution of repair systems in different life forms is that, although certain domains, such as helicases and nucleases, are largely conserved in all organisms, the number of orthologous proteins shared by bacteria, archaea and eukaryotes is very small [400].

According to our results, different proteins and modules have been added to the DDR system (Figures 32-34), which may have increased its complexity in terms of fine-tuning and cross-talk to other pathways, as seems to have occurred in other systems, like the chromatin modification machinery [263]. The addition of proteins could have, for instance, increased the efficiency of the process or system. In this regard, NBN may have joined the MRN complex to facilitate the linking of ATM in sites of DSBs, and ATRIP could have evolved to link the ATR kinase to SSB sites. Another example is the RNF168 protein, which interacts with ubiquitylated H2A, assembles at DSBs in an RNF8-dependent manner and, by targeting H2AX, amplifies local concentration of lysine 63-linked ubiquitin conjugates to the threshold required for retention of 53BP1 and BRCA1 [99].

6 CONCLUSIONS

- The literature-based overlap of DDR sets is unexpectedly low, indicating large differences in the trends and research conducted in DDR by different scientific communities.

- Most of the DDR genes are of old origin, being some of the core components traceable to prokaryotes. The DDR pathways seem to have grown around these ancient modules.

- All the sub-pathways contain at least one member traceable to ancient prokaryotes and DNA repair is the most ancient module with remarkable variations to accommodate different life-styles.

- The ancestral network is mainly metabolic. The majority of the initial components are sensors and effectors, representing the two extremes of the pathway. Further additions of components, in particular regulatory elements, have increased the complexity of the network.

- The evolutionary history of DDR protein families is more complex than expected. A plausible explanation is the massive HGT events coupled to gene loses. Therefore, the gene age framework should be revisited.

- Most of the DDR components have been lost at least more than twice during evolution, especially in fungi, nematodes and arthropods, where essential proteins in complexes have been lost in several species.
  In specific cases of losses, functional analogs could compensate the absences, or alternative systems might intervene.

- Lineage-specific and domain rearrangement events may have included novel functions in various organisms, principally in plants.

- Human DDR proteins are enriched in specific domains such as the BRCT repeats, Rad51, Helicase_C or AAA domains. Besides, certain domains are specifically found in determined functional tiers, such as MutS, MutL, Ku and PARP related domains in Sensors; Histone, UIM and Tower in Mediators; POLO_box and UQ_con in Transducers, and finally, the XRCC4 and SWIB domains in Effectors.

- The enlargement of the network has occurred through the addition of new components that have evolved to interact and work together with the ancient ones, which may have increased the complexity of the DDR network in terms of fine-tuning and cross-talk to other pathways.

7. FUTURE PERSPECTIVES

- The DDR components should be studied in the frameworks of additional systems, like i.e.: replisome or proteosome.

- These analyses should be extended to alternative model species to identify precise lineage specific expansions.

- The genomic regulation of these components could shed some light into the evolutionary process of the pathway.

- A detailed and comprehensive analysIs of uncharacterized regions of DDR proteins could provide useful hints regarding potential functions.

- The identification of sequence/structure specific signatures in relevant domains and/or uncharacterized regions could help to spot sub and neo-functionalization processes.

- A detailed analysIs of post-translational modifications in the DDR system compared to genome-wide could help to evaluate the evolutionary trends observed for DDR modifiers in this study.

BIBLIOGRAPHY

1 Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb Symp Quant Biol* **65**, 127-133, (2000).

2 Friedberg, E. C. *DNA repair and mutagenesis*. 2nd edn, (ASM Press, 2006).

3 Friedberg, E. C. *Correcting the blueprint of life : an historical account of the discovery of DNA repair mechanisms*. (Cold Spring Harbor Laboratory Press, 1997).

4 Hoeijmakers, J. H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-1485, (2009).

5 Polo, S. E. & Jackson, S. P. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev* **25**, 409-433, (2011).

6 Branzei, D. & Foiani, M. Regulation of DNA repair throughout the cell cycle. *Nat Rev Mol Cell Biol* **9**, 297-308, (2008).

7 Reinhardt, H. C. & Yaffe, M. B. Kinases that control the cell cycle in response to DNA damage: Chk1, Chk2, and MK2. *Curr Opin Cell Biol* **21**, 245-255, (2009).

8 Huen, M. S. & Chen, J. Assembly of checkpoint and repair machineries at DNA damage sites. *Trends Biochem Sci* **35**, 101-108, (2010).

9 Wang, P. *et al.* microRNA-21 negatively regulates Cdc25A and cell cycle progression in colon cancer cells. *Cancer Res* **69**, 8157-8165, (2009).

10 Cannell, I. G. & Bushell, M. Regulation of Myc by miR-34c: A mechanism to prevent genomic instability? *Cell Cycle* **9**, 2726-2730, (2010).

11 Zhang, X. *et al.* Oncogenic Wip1 phosphatase is inhibited by miR-16 in the DNA damage signaling pathway. *Cancer Res* **70**, 7176-7186, (2010).

12 Boucas, J. *et al.* Posttranscriptional regulation of gene expression-adding another layer of complexity to the DNA damage response. *Front Genet* **3**, 159, (2012).

13 Jansson, M. D. & Lund, A. H. MicroRNA and cancer. *Mol Oncol* **6**, 590-610, (2012).

14 Liu, Y. & Lu, X. Non-coding RNAs in DNA damage response. *Am J Cancer Res* **2**, 658-675, (2012).

15 Ljungman, M. The DNA damage response--repair or despair? *Environ Mol Mutagen* **51**, 879-889, (2010).

16 Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715, (1993).

17 Sander, M. *et al.* Proceedings of a workshop on DNA adducts: biological significance and applications to risk assessment Washington, DC, April 13-14, 2004. *Toxicol Appl Pharmacol* **208**, 1-20, (2005).

18 De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169-185, (2004).

19 Sancar, A., Lindsey-Boltz, L. A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* **73**, 39-85, (2004).

20 Freitas, A. A. & de Magalhaes, J. P. A review and appraisal of the DNA damage theory of ageing. *Mutat Res* **728**, 12-22, (2011).

21 Kow, Y. W. & Dare, A. Detection of abasic sites and oxidative DNA base damage using an ELISA-like assay. *Methods* **22**, 164-169, (2000).

22 Boiteux, S. & Guillet, M. Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisiae. *DNA Repair (Amst)* **3**, 1-12, (2004).

23 Friedberg, E. C. A brief history of the DNA repair field. *Cell Res* **18**, 3-7, (2008).

24 Schlissel, M., Constantinescu, A., Morrow, T., Baxter, M. & Peng, A. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. *Genes Dev* **7**, 2520-2532, (1993).

25 Lord, C. J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287-294, (2012).

26 Lisby, M., Barlow, J. H., Burgess, R. C. & Rothstein, R. Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. *Cell* **118**, 699-713, (2004).

27 Bartek, J. & Lukas, J. Mammalian G1- and S-phase checkpoints in response to DNA damage. *Curr Opin Cell Biol* **13**, 738-747, (2001).

28 Bernstein, C., Bernstein, H., Payne, C. M. & Garewal, H. DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis. *Mutat Res* **511**, 145-178, (2002).

29 Hoeijmakers, J. H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-374, (2001).

30 Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204, (2010).

31 Giglia-Mari, G., Zotter, A. & Vermeulen, W. DNA Damage Response. *Cold Spring Harb Perspect Biol*, (2010).

32 Weber, S. Light-driven enzymatic catalysis of DNA repair: a review of recent biophysical studies on photolyase. *Biochim Biophys Acta* **1707**, 1-23, (2005).

33 Iyer, R. R., Pluciennik, A., Burdett, V. & Modrich, P. L. DNA mismatch repair: functions and mechanisms. *Chem Rev* **106**, 302-323, (2006).

34 Jiricny, J. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* **7**, 335-346, (2006).

35 Hong, Z. *et al.* Recruitment of mismatch repair proteins to the site of DNA damage in human cells. *J Cell Sci* **121**, 3146-3154, (2008).

36 Kadyrov, F. A., Dzantiev, L., Constantin, N. & Modrich, P. Endonucleolytic function of MutLalpha in human mismatch repair. *Cell* **126**, 297-308, (2006).

37 Li, G. M. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**, 85-98, (2008).

38 Shahi, A. *et al.* Mismatch-repair protein MSH6 is associated with Ku70 and regulates DNA double-strand break repair. *Nucleic Acids Res* **39**, 2130-2143, (2011).

39 Gredilla, R. DNA damage and base excision repair in mitochondria and their role in aging. *J Aging Res* **2011**, 257093, (2010).

40 Sattler, U., Frit, P., Salles, B. & Calsou, P. Long-patch DNA repair synthesis during base excision repair in mammalian cells. *EMBO Rep* **4**, 363-367, (2003).

41 Fortini, P. & Dogliotti, E. Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways. *DNA Repair (Amst)* **6**, 398-409, (2007).

42 Fousteri, M. & Mullenders, L. H. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res* **18**, 73-84, (2008).

43 Tapias, A. *et al.* Ordered conformational changes in damaged DNA induced by nucleotide excision repair factors. *J Biol Chem* **279**, 19074-19083, (2004).

44 Leibeling, D., Laspe, P. & Emmert, S. Nucleotide excision repair and cancer. *J Mol Histol* **37**, 225-238, (2006).

45 Handa, N., Morimatsu, K., Lovett, S. T. & Kowalczykowski, S. C. Reconstitution of initial steps of dsDNA break repair by the RecF pathway of E. coli. *Genes Dev* **23**, 1234-1245, (2009).

46 Weiner, A., Zauberman, N. & Minsky, A. Recombinational DNA repair in a cellular context: a search for the homology search. *Nat Rev Microbiol* **7**, 748-755, (2009).

47 Lopez-Camarillo, C. *et al.* DNA repair mechanisms in eukaryotes: Special focus in Entamoeba histolytica and related protozoan parasites. *Infect Genet Evol* **9**, 1051-1056, (2009).

48 Errico, A. & Costanzo, V. Differences in the DNA replication of unicellular eukaryotes and metazoans: known unknowns. *EMBO Rep* **11**, 270-278, (2010).

49 Wang, H., Perrault, A. R., Takeda, Y., Qin, W. & Iliakis, G. Biochemical evidence for Ku-independent backup pathways of NHEJ. *Nucleic Acids Res* **31**, 5377-5388, (2003).

50 Walker, J. R., Corpina, R. A. & Goldberg, J. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* **412**, 607-614, (2001).

51 Lees-Miller, S. P. & Meek, K. Repair of DNA double strand breaks by non-homologous end joining. *Biochimie* **85**, 1161-1173, (2003).

52 Kurosawa, A. & Adachi, N. Functions and regulation of Artemis: a goddess in the maintenance of genome integrity. *J Radiat Res* **51**, 503-509, (2010).

53 Hentges, P. *et al.* Evolutionary and functional conservation of the DNA non-homologous end-joining protein, XLF/Cernunnos. *J Biol Chem* **281**, 37517-37526, (2006).

54 Cappelli, E., Townsend, S., Griffin, C. & Thacker, J. Homologous recombination proteins are associated with centrosomes and are required for mitotic stability. *Exp Cell Res* **317**, 1203-1213, (2011).

55 Potts, P. R. & Yu, H. The SMC5/6 complex maintains telomere length in ALT cancer cells through SUMOylation of telomere-binding proteins. *Nat Struct Mol Biol* **14**, 581-590, (2007).

56 San Filippo, J., Sung, P. & Klein, H. Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* **77**, 229-257, (2008).

57 Fekairi, S. *et al.* Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases. *Cell* **138**, 78-89, (2009).

58 Ogrunc, M. & Sancar, A. Identification and characterization of human MUS81-MMS4 structure-specific endonuclease. *J Biol Chem* **278**, 21715-21720, (2003).

59 Kennedy, R. D. & D'Andrea, A. D. The Fanconi Anemia/BRCA pathway: new faces in the crowd. *Genes Dev* **19**, 2925-2940, (2005).

60 Williams, R. S., Williams, J. S. & Tainer, J. A. Mre11-Rad50-Nbs1 is a keystone complex connecting DNA repair machinery, double-strand break signaling, and the chromatin template. *Biochem Cell Biol* **85**, 509-520, (2007).

61 Stracker, T. H. & Petrini, J. H. The MRE11 complex: starting from the ends. *Nat Rev Mol Cell Biol* **12**, 90-103, (2011).

62 Mimitou, E. P. & Symington, L. S. Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. *Nature* **455**, 770-774, (2008).

63 Nimonkar, A. V., Ozsoy, A. Z., Genschel, J., Modrich, P. & Kowalczykowski, S. C. Human exonuclease 1 and BLM helicase interact to resect DNA and initiate DNA repair. *Proc Natl Acad Sci U S A* **105**, 16906-16911, (2008).

64 Tomimatsu, N. *et al.* Exo1 plays a major role in DNA end resection in humans and influences double-strand break repair and damage signaling decisions. *DNA Repair (Amst)* **11**, 441-448, (2012).

65 Wyman, C., Ristic, D. & Kanaar, R. Homologous recombination-mediated double-strand break repair. *DNA Repair (Amst)* **3**, 827-833, (2004).

66 Friedberg, E. C., Lehmann, A. R. & Fuchs, R. P. Trading places: how do DNA polymerases switch during translesion DNA synthesis? *Mol Cell* **18**, 499-505, (2005).

67 Bassermann, F. & Pagano, M. Dissecting the role of ubiquitylation in the DNA damage response checkpoint in G2. *Cell Death Differ* **17**, 78-85, (2010).

68 Langerak, P. & Russell, P. Regulatory networks integrating cell cycle control with DNA damage checkpoints and double-strand break repair. *Philos Trans R Soc Lond B Biol Sci* **366**, 3562-3571, (2011).

69 Shrivastav, M., De Haro, L. P. & Nickoloff, J. A. Regulation of DNA double-strand break repair pathway choice. *Cell Res* **18**, 134-147, (2008).

70 Miller, K. M. *et al.* Human HDAC1 and HDAC2 function in the DNA-damage response to promote DNA nonhomologous end-joining. *Nat Struct Mol Biol* **17**, 1144-1151, (2010).

71 Shibata, A. *et al.* Factors determining DNA double-strand break repair pathway choice in G2 phase. *EMBO J* **30**, 1079-1092, (2011).

72 Haber, J. E. Partners and pathwaysrepairing a double-strand break. *Trends Genet* **16**, 259-264, (2000).

73 Helleday, T. Homologous recombination in cancer development, treatment and development of drug resistance. *Carcinogenesis* **31**, 955-960, (2010).

74 Harper, J. W. & Elledge, S. J. The DNA damage response: ten years after. *Mol Cell* **28**, 739-745, (2007).

75 de Souza-Pinto, N. C., Wilson, D. M., 3rd, Stevnsner, T. V. & Bohr, V. A. Mitochondrial DNA, base excision repair and neurodegeneration. *DNA Repair (Amst)* **7**, 1098-1109, (2008).

76 Palm, W. & de Lange, T. How shelterin protects mammalian telomeres. *Annu Rev Genet* **42**, 301-334, (2008).

77 Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071-1078, (2009).

78 Lee, J. H. & Paull, T. T. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* **308**, 551-554, (2005).

79 Wang, M. *et al.* PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Res* **34**, 6170-6182, (2006).

80 Durocher, D. & Jackson, S. P. DNA-PK, ATM and ATR as sensors of DNA damage: variations on a theme? *Curr Opin Cell Biol* **13**, 225-231, (2001).

81 Huen, M. S. & Chen, J. The DNA damage response pathways: at the crossroad of protein modifications. *Cell Res* **18**, 8-16, (2008).

82 Costes, S. V. *et al.* Image-based modeling reveals dynamic redistribution of DNA damage into nuclear sub-domains. *PLoS Comput Biol* **3**, e155, (2007).

83 Lukas, J., Lukas, C. & Bartek, J. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. *Nat Cell Biol* **13**, 1161-1169, (2011).

84 Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316**, 1160-1166, (2007).

85 Paulsen, R. D. *et al.* A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Mol Cell* **35**, 228-239, (2009).

86 Goodarzi, A. A. *et al.* ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin. *Mol Cell* **31**, 167-177, (2008).

87 Noon, A. T. *et al.* 53BP1-dependent robust localized KAP-1 phosphorylation is essential for heterochromatic DNA double-strand break repair. *Nat Cell Biol* **12**, 177-184, (2010).

88 Jakob, B. *et al.* DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin. *Nucleic Acids Res* **39**, 6489-6499, (2011).

89 Shiotani, B. & Zou, L. Single-stranded DNA orchestrates an ATM-to-ATR switch at DNA breaks. *Mol Cell* **33**, 547-558, (2009).

90 Lee, J. H., Goodarzi, A. A., Jeggo, P. A. & Paull, T. T. 53BP1 promotes ATM activity through direct interactions with the MRN complex. *EMBO J* **29**, 574-585, (2010).

91 Sun, Y., Jiang, X., Chen, S., Fernandes, N. & Price, B. D. A role for the Tip60 histone acetyltransferase in the acetylation and activation of ATM. *Proc Natl Acad Sci U S A* **102**, 13182-13187, (2005).

92 Murr, R. *et al.* Histone acetylation by Trrap-Tip60 modulates loading of repair proteins and repair of DNA double-strand breaks. *Nat Cell Biol* **8**, 91-99, (2006).

93 Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* **273**, 5858-5868, (1998).

94 Cook, P. J. *et al.* Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions. *Nature* **458**, 591-596, (2009).

95 Mok, M. T. & Henderson, B. R. Three-dimensional imaging reveals the spatial separation of gammaH2AX-MDC1-53BP1 and RNF8-RNF168-BRCA1-A complexes at ionizing radiation-induced foci. *Radiother Oncol* **103**, 415-420, (2012).

96 Lou, Z. *et al.* MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. *Mol Cell* **21**, 187-200, (2006).

97 Panier, S. & Durocher, D. Regulatory ubiquitylation in response to DNA double-strand breaks. *DNA Repair (Amst)* **8**, 436-443, (2009).

98 Bekker-Jensen, S. *et al.* HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. *Nat Cell Biol* **12**, 80-86; sup pp 81-12, (2010).

99 Doil, C. *et al.* RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins. *Cell* **136**, 435-446, (2009).

100 Huang, J. *et al.* RAD18 transmits DNA damage signalling to elicit homologous recombination repair. *Nat Cell Biol* **11**, 592-603, (2009).

101 Jowsey, P. A., Doherty, A. J. & Rouse, J. Human PTIP facilitates ATM-mediated activation of p53 and promotes cellular resistance to ionizing radiation. *J Biol Chem* **279**, 55562-55569, (2004).

102 Gong, Z., Cho, Y. W., Kim, J. E., Ge, K. & Chen, J. Accumulation of Pax2 transactivation domain interaction protein (PTIP) at sites of DNA breaks via RNF8-dependent pathway is required for cell survival after DNA damage. *J Biol Chem* **284**, 7284-7293, (2009).

103 Yan, W. *et al.* Structural basis of gammaH2AX recognition by human PTIP BRCT5-BRCT6 domains in the DNA damage response pathway. *FEBS Lett* **585**, 3874-3879, (2011).

104 Yazdi, P. T. *et al.* SMC1 is a downstream effector in the ATM/NBS1 branch of the human S-phase checkpoint. *Genes Dev* **16**, 571-582, (2002).

105 Wu, J., Prindle, M. J., Dressler, G. R. & Yu, X. PTIP regulates 53BP1 and SMC1 at the DNA damage sites. *J Biol Chem* **284**, 18078-18084, (2009).

106 Bauerschmidt, C. *et al.* Cohesin phosphorylation and mobility of SMC1 at ionizing radiation-induced DNA double-strand breaks in human cells. *Exp Cell Res* **317**, 330-337, (2011).

107 Sobhian, B. *et al.* RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science* **316**, 1198-1202, (2007).

108 Shao, G. *et al.* MERIT40 controls BRCA1-Rap80 complex integrity and recruitment to DNA double-strand breaks. *Genes Dev* **23**, 740-754, (2009).

109 Feng, L., Huang, J. & Chen, J. MERIT40 facilitates BRCA1 localization and DNA damage repair. *Genes Dev* **23**, 719-728, (2009).

110 Schoenfeld, A. R., Apgar, S., Dolios, G., Wang, R. & Aaronson, S. A. BRCA2 is ubiquitinated in vivo and interacts with USP11, a deubiquitinating enzyme that exhibits prosurvival function in the cellular response to DNA damage. *Mol Cell Biol* **24**, 7444-7455, (2004).

111 Wiltshire, T. D. *et al.* Sensitivity to poly(ADP-ribose) polymerase (PARP) inhibition identifies ubiquitin-specific peptidase 11 (USP11) as a regulator of DNA double-strand break repair. *J Biol Chem* **285**, 14565-14571, (2010).

112 Galanty, Y. *et al.* Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. *Nature* **462**, 935-939, (2009).

113 Morris, J. R. *et al.* The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. *Nature* **462**, 886-890, (2009).

114 Ahn, J., Urist, M. & Prives, C. The Chk2 protein kinase. *DNA Repair (Amst)* **3**, 1039-1047, (2004).

115 Chen, L., Gilkes, D. M., Pan, Y., Lane, W. S. & Chen, J. ATM and Chk2-dependent phosphorylation of MDMX contribute to p53 activation after DNA damage. *EMBO J* **24**, 3411-3422, (2005).

116 Pan, X. *et al.* Induction of SOX4 by DNA damage is critical for p53 stabilization and function. *Proc Natl Acad Sci U S A* **106**, 3788-3793, (2009).

117 Moumen, A., Masterson, P., O'Connor, M. J. & Jackson, S. P. hnRNP K: an HDM2 target and transcriptional coactivator of p53 in response to DNA damage. *Cell* **123**, 1065-1078, (2005).

118 Shiloh, Y. FBXO31: a new player in the ever-expanding DNA damage response orchestra. *Sci Signal* **2**, pe73, (2009).

119 Santra, M. K., Wajapeyee, N. & Green, M. R. F-box protein FBXO31 mediates cyclin D1 degradation to induce G1 arrest after DNA damage. *Nature* **459**, 722-725, (2009).

120 Cimprich, K. A. & Cortez, D. ATR: an essential regulator of genome integrity. *Nat Rev Mol Cell Biol* **9**, 616-627, (2008).

121 Bansbach, C. E., Betous, R., Lovejoy, C. A., Glick, G. G. & Cortez, D. The annealing helicase SMARCAL1 maintains genome integrity at stalled replication forks. *Genes Dev* **23**, 2405-2414, (2009).

122 Allen, C., Ashley, A. K., Hromas, R. & Nickoloff, J. A. More forks on the road to replication stress recovery. *J Mol Cell Biol* **3**, 4-12, (2011).

123 Gardino, A. K. & Yaffe, M. B. 14-3-3 proteins as signaling integration points for cell cycle control and apoptosis. *Semin Cell Dev Biol* **22**, 688-695, (2011).

124 Guardavaccaro, D. & Pagano, M. Stabilizers and destabilizers controlling cell cycle oscillators. *Mol Cell* **22**, 1-4, (2006).

125 Mailand, N. *et al.* Regulation of G(2)/M events by Cdc25A through phosphorylation-dependent modulation of its stability. *EMBO J* **21**, 5911-5920, (2002).

126 Gewurz, B. E. & Harper, J. W. DNA-damage control: Claspin destruction turns off the checkpoint. *Curr Biol* **16**, R932-934, (2006).

127 Toyoshima-Morimoto, F., Taniguchi, E. & Nishida, E. Plk1 promotes nuclear translocation of human Cdc25C during prophase. *EMBO Rep* **3**, 341-348, (2002).

128 van Vugt, M. A. *et al.* A mitotic phosphorylation feedback network connects Cdk1, Plk1, 53BP1, and Chk2 to inactivate the G(2)/M DNA damage checkpoint. *PLoS Biol* **8**, e1000287, (2010).

129 Arias, E. E. & Walter, J. C. Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev* **21**, 497-518, (2007).

130 Medema, R. H. & Macurek, L. Checkpoint recovery in cells: how a molecular understanding can help in the fight against cancer. *F1000 Biol Rep* **3**, 10, (2011).

131 Bartkova, J. *et al.* DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* **434**, 864-870, (2005).

132 Nuciforo, P. G., Luise, C., Capra, M., Pelosi, G. & d'Adda di Fagagna, F. Complex engagement of DNA damage response pathways in human cancer and in lung tumor progression. *Carcinogenesis* **28**, 2082-2088, (2007).

133 Poehlmann, A. & Roessner, A. Importance of DNA damage checkpoints in the pathogenesis of human cancers. *Pathol Res Pract* **206**, 591-601, (2010).

134 Canman, C. E. & Lim, D. S. The role of ATM in DNA damage responses and cancer. *Oncogene* **17**, 3301-3308, (1998).

135 O'Driscoll, M., Ruiz-Perez, V. L., Woods, C. G., Jeggo, P. A. & Goodship, J. A. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nat Genet* **33**, 497-501, (2003).

136 Digweed, M. & Sperling, K. Nijmegen breakage syndrome: clinical manifestation of defective response to DNA double-strand breaks. *DNA Repair (Amst)* **3**, 1207-1217, (2004).

137 Spivak, G. The many faces of Cockayne syndrome. *Proc Natl Acad Sci U S A* **101**, 15273-15274, (2004).

138 Maslov, A. Y. & Vijg, J. Genome instability, cancer and aging. *Biochim Biophys Acta* **1790**, 963-969, (2009).

139 Branzei, D. & Foiani, M. Maintaining genome stability at the replication fork. *Nat Rev Mol Cell Biol* **11**, 208-219, (2010).

140 Wang, J., Gong, Z. & Chen, J. MDC1 collaborates with TopBP1 in DNA replication checkpoint control. *J Cell Biol* **193**, 267-273, (2011).

141 Stewart, G. S. *et al.* The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. *Cell* **136**, 420-434, (2009).

142 Devgan, S. S. *et al.* Homozygous deficiency of ubiquitin-ligase ring-finger protein RNF168 mimics the radiosensitivity syndrome of ataxia-telangiectasia. *Cell Death Differ* **18**, 1500-1506, (2011).

143 Mah, L. J., El-Osta, A. & Karagiannis, T. C. GammaH2AX as a molecular marker of aging and disease. *Epigenetics* **5**, 129-136, (2010).

144 Murga, M. *et al.* Exploiting oncogene-induced replicative stress for the selective killing of Myc-driven tumors. *Nat Struct Mol Biol* **18**, 1331-1335, (2011).

145 Melo, S. A. & Kalluri, R. Molecular pathways: microRNAs as cancer therapeutics. *Clin Cancer Res* **18**, 4234-4239, (2012).

146 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512, (1995).

147 Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**, D571-579, (2012).

148 Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288, (1999).

149 Frohlich, H., Fellmann, M., Sultmann, H., Poustka, A. & Beissbarth, T. Predicting pathway membership via domain signatures. *Bioinformatics* **24**, 2137-2142, (2008).

150 Garcia-Jimenez, B., Juan, D., Ezkurdia, I., Andres-Leon, E. & Valencia, A. Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS One* **5**, e9969, (2010).

151 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129, (1999).

152 Aravind, L. & Subramanian, G. Origin of multicellular eukaryotes - insights from proteome comparisons. *Curr Opin Genet Dev* **9**, 688-694, (1999).

153 Light, S. & Kraulis, P. Network analysis of metabolic enzyme evolution in Escherichia coli. *BMC Bioinformatics* **5**, 15, (2004).

154 Mazurie, A., Bonchev, D., Schwikowski, B. & Buck, G. A. Evolution of metabolic network organization. *BMC Syst Biol* **4**, 59, (2010).

155 Zmasek, C. M., Zhang, Q., Ye, Y. & Godzik, A. Surprising complexity of the ancestral apoptosis network. *Genome Biol* **8**, R226, (2007).

156 Diez, D., Sanchez-Jimenez, F. & Ranea, J. A. Evolutionary expansion of the Ras switch regulatory module in eukaryotes. *Nucleic Acids Res* **39**, 5526-5537, (2011).

157 Rojas, A. M., Fuentes, G., Rausell, A. & Valencia, A. Evolution: The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *J Cell Biol* **196**, 189-201, (2012).

158 Alvarez-Ponce, D., Aguade, M. & Rozas, J. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Res* **19**, 234-242, (2009).

159 Kultz, D. Molecular and evolutionary basis of the cellular stress response. *Annu Rev Physiol* **67**, 225-257, (2005).

160 Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**, 823-826, (1986).

161 Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94, (1999).

162 Thornton, J. W. & DeSalle, R. Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet* **1**, 41-73, (2000).

163 Janga, S. C., Diaz-Mejia, J. J. & Moreno-Hagelsieb, G. Network-based function prediction and interactomics: the case for metabolic enzymes. *Metab Eng* **13**, 1-10, (2011).

164 Loewenstein, Y. *et al.* Protein function annotation by homology-based inference. *Genome Biol* **10**, 207, (2009).

165 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, (1990).

166 Reid, A. J., Yeats, C. & Orengo, C. A. Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics* **23**, 2353-2360, (2007).

167 Kaessmann, H., Zollner, S., Nekrutenko, A. & Li, W. H. Signatures of domain shuffling in the human genome. *Genome Res* **12**, 1642-1650, (2002).

168 Basu, M. K., Carmel, L., Rogozin, I. B. & Koonin, E. V. Evolution of protein domain promiscuity in eukaryotes. *Genome Res* **18**, 449-461, (2008).

169 Buljan, M. & Bateman, A. The evolution of protein domain families. *Biochem Soc Trans* **37**, 751-755, (2009).

170 Sigrist, C. J. *et al.* PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* **38**, D161-166, (2010).

171 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**, D225-229, (2011).

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

173 Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-215, (2009).

174 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **38**, D211-222, (2010).

175 Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40**, D302-305, (2012).

176 Taylor, W. R. & Orengo, C. A. Protein structure alignment. *J Mol Biol* **208**, 1-22, (1989).

177 Knudsen, M. & Wiuf, C. The CATH database. *Hum Genomics* **4**, 207-212, (2010).

178 Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-229, (2004).

179 Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-248, (2005).

180 Berman, H. M. *et al.* The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**, 899-907, (2002).

181 Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).

182 Doyle, J. J. & Gaut, B. S. Evolution of genes and taxa: a primer. *Plant Mol Biol* **42**, 1-23, (2000).

183 Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**, e16, (2007).

184 Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **5**, e1000605, (2009).

185 Wong, W. C., Maurer-Stroh, S. & Eisenhaber, F. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol* **6**, e1000867, (2010).

186 Tress, M. *et al.* Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins* **69 Suppl 8**, 137-151, (2007).

187 Lopez, G., Rojas, A., Tress, M. & Valencia, A. Assessment of predictions submitted for the CASP7 function prediction category. *Proteins* **69 Suppl 8**, 165-174, (2007).

188 Marabotti, A. & Facchiano, A. When it comes to homology, bad habits die hard. *Trends Biochem Sci* **34**, 98-99, (2009).

189 Sonnhammer, E. L. & Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**, 619-620, (2002).

190 Ouzounis, C. Orthology: another terminology muddle. *Trends Genet* **15**, 445, (1999).

191 Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**, 544-549, (2004).

192 Diaz, R. *et al.* argC Orthologs from Rhizobiales show diverse profiles of transcriptional efficiency and functionality in Sinorhizobium meliloti. *J Bacteriol* **193**, 460-472, (2011).

193 Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-338, (2005).

194 Gabaldon, T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* **9**, 235, (2008).

195 Chen, L., DeVries, A. L. & Cheng, C. H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc Natl Acad Sci U S A* **94**, 3817-3822, (1997).

196 Hulsen, T., Huynen, M. A., de Vlieg, J. & Groenen, P. M. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**, R31, (2006).

197 Altenhoff, A. M. & Dessimoz, C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**, e1000262, (2009).

198 Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. Computational methods for Gene Orthology inference. *Brief Bioinform* **12**, 379-391, (2011).

199 Storm, C. E. & Sonnhammer, E. L. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* **13**, 2353-2362, (2003).

200 Zmasek, C. M. & Eddy, S. R. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3**, 14, (2002).

201 Jothi, R., Zotenko, E., Tasneem, A. & Przytycka, T. M. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* **22**, 779-788, (2006).

202 Pryszcz, L. P., Huerta-Cepas, J. & Gabaldon, T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res* **39**, e32, (2011).

203 Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22-28, (2001).

204 Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**, D196-203, (2010).

205 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, (2003).

206 Chen, T. W., Wu, T. H., Ng, W. V. & Lin, W. C. DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection. *BMC Bioinformatics* **11 Suppl 7**, S6, (2010).

207 Doolittle, R. F. The roots of bioinformatics in protein evolution. *PLoS Comput Biol* **6**, e1000875, (2010).

208 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518, (2005).

209 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, (2000).

210 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, (2007).

211 Kemena, C. & Notredame, C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**, 2455-2465, (2009).

212 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574, (2003).

213 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739, (2011).

214 Hall, B. G. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol Biol Evol* **22**, 792-802, (2005).

215 Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105, (2005).

216 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755, (2001).

217 Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407-415, (2004).

218 Soltis, D. E. & Soltis, P. S. The role of phylogenetics in comparative genetics. *Plant Physiol* **132**, 1790-1800, (2003).

219 Hillis, D. & Bull, J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**, 182-192, (1993).

220 Sarode, V. R., Savitri, K., Banerjee, C. K., Narasimharao, K. L. & Khajuria, A. Primary extrarenal Wilms' tumour: identification of a putative precursor lesion. *Histopathology* **21**, 76-78, (1992).

221 Woese, C. R. Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* **97**, 8392-8396, (2000).

222 Rojas, A. & Doolittle, R. F. The occurrence of type S1A serine proteases in sponge and jellyfish. *J Mol Evol* **55**, 790-794, (2006).

223 Smith, S. E. *et al.* Comparative genomic and phylogenetic approaches to characterize the role of genetic recombination in mycobacterial evolution. *PLoS One* **7**, e50070, (2012).

224 Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679-687, (2005).

225 Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891, (2000).

226 Wagner, G. P., Fried, C., Prohaska, S. J. & Stadler, P. F. Divergence of conserved non-coding sequences: rate estimates and relative rate tests. *Mol Biol Evol* **21**, 2116-2121, (2004).

227 Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A* **106**, 7273-7280, (2009).

228 Wolf, Y. I., Carmel, L. & Koonin, E. V. Unifying measures of gene function and evolution. *Proc Biol Sci* **273**, 1507-1515, (2006).

229 Koonin, E. V. & Wolf, Y. I. Evolutionary systems biology: links between gene evolution and function. *Curr Opin Biotechnol* **17**, 481-487, (2006).

230 Eisen, J. A. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* **26**, 4291-4300, (1998).

231 Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat Res* **435**, 171-213, (1999).

232 Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**, 482-496, (2002).

233 Hiom, K. DNA repair: bacteria join in. *Curr Biol* **13**, R28-30, (2003).

234 Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* **11**, 356-372, (2001).

235 Levesque, M., Shasha, D., Kim, W., Surette, M. G. & Benfey, P. N. Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Curr Biol* **13**, 129-133, (2003).

236 Doerks, T., van Noort, V., Minguez, P. & Bork, P. Annotation of the M. tuberculosis hypothetical orfeome: adding functional information to more than half of the uncharacterized proteins. *PLoS One* **7**, e34302, (2012).

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

238 Modrich, P. & Lahue, R. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu Rev Biochem* **65**, 101-133, (1996).

239 Kim, M. Y., Zhang, T. & Kraus, W. L. Poly(ADP-ribosyl)ation by PARP-1: 'PAR-laying' NAD+ into a nuclear signal. *Genes Dev* **19**, 1951-1967, (2005).

240 Zgheib, O., Pataky, K., Brugger, J. & Halazonetis, T. D. An oligomerized 53BP1 tudor domain suffices for recognition of DNA double-strand breaks. *Mol Cell Biol* **29**, 1050-1058, (2009).

241 Mesquita, R. D., Woods, N. T., Seabra-Junior, E. S. & Monteiro, A. N. Tandem BRCT Domains: DNA's Praetorian Guard. *Genes Cancer* **1**, 1140-1146, (2010).

242 Nishino, T. & Morikawa, K. Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. *Oncogene* **21**, 9022-9032, (2002).

243 Sykorova, E. & Fajkus, J. Structure-function relationships in telomerase genes. *Biol Cell* **101**, 375-392, 371 p following 392, (2009).

244 Kirschner, M. & Gerhart, J. Evolvability. *Proc Natl Acad Sci U S A* **95**, 8420-8427, (1998).

245 Morita, R. *et al.* Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. *J Nucleic Acids* **2010**, 179594, (2010).

246 Stracker, T. H., Usui, T. & Petrini, J. H. Taking the time to make important decisions: the checkpoint effector kinases Chk1 and Chk2 and the DNA damage response. *DNA Repair (Amst)* **8**, 1047-1054, (2009).

247 Huen, M. S., Sy, S. M. & Chen, J. BRCA1 and its toolbox for the maintenance of genome integrity. *Nat Rev Mol Cell Biol* **11**, 138-148, (2010).

248 Lancaster, J. M. *et al.* BRCA2 mutations in primary breast and ovarian cancers. *Nat Genet* **13**, 238-240, (1996).

249 Vorechovsky, I., Luo, L., Ortmann, E., Steinmann, D. & Dork, T. Missense mutations at ATM gene and cancer risk. *Lancet* **353**, 1276, (1999).

250 Schaffner, C., Idler, I., Stilgenbauer, S., Dohner, H. & Lichter, P. Mantle cell lymphoma is characterized by inactivation of the ATM gene. *Proc Natl Acad Sci U S A* **97**, 2773-2778, (2000).

251 Tanaka, A. *et al.* Germline mutation in ATR in autosomal- dominant oropharyngeal cancer syndrome. *Am J Hum Genet* **90**, 511-517, (2012).

252 Cohn, M. A. & D'Andrea, A. D. Chromatin recruitment of DNA repair proteins: lessons from the fanconi anemia and double-strand break repair pathways. *Mol Cell* **32**, 306-312, (2008).

253 Fernandez-Capetillo, O. Intrauterine programming of ageing. *EMBO Rep* **11**, 32-36, (2010).

254 Villarroel, M. C. *et al.* Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer. *Mol Cancer Ther* **10**, 3-8, (2011).

255 Wahl, G. M. & Carr, A. M. The evolution of diverse biological responses to DNA damage: insights from yeast and p53. *Nat Cell Biol* **3**, E277-286, (2001).

256 Wakabayashi, M., Ishii, C., Inoue, H. & Tanaka, S. Genetic analysis of CHK1 and CHK2 homologues revealed a unique cross talk between ATM and ATR pathways in Neurospora crassa. *DNA Repair (Amst)* **7**, 1951-1961, (2008).

257 Kazama, Y. *et al.* The Neurospora crassa UVS-3 epistasis group encodes homologues of the ATR/ATRIP checkpoint control system. *DNA Repair (Amst)* **7**, 213-229, (2008).

258 Kimura, S. & Sakaguchi, K. DNA repair in plants. *Chem Rev* **106**, 753-766, (2006).

259 Lane, D. P. *et al.* Mdm2 and p53 are highly conserved from placozoans to man. *Cell Cycle* **9**, 540-547, (2010).

260 Castro, M. A., Dalmolin, R. J., Moreira, J. C., Mombach, J. C. & de Almeida, R. M. Evolutionary origins of human apoptosis and genome-stability gene networks. *Nucleic Acids Res* **36**, 6269-6283, (2008).

261 Woods, N. T. *et al.* Charting the Landscape of Tandem BRCT Domain-Mediated Protein Interactions. *Sci Signal* **5**, rs6, (2012).

262 Lisby, M. & Rothstein, R. Localization of checkpoint and repair proteins in eukaryotes. *Biochimie* **87**, 579-589, (2005).

263 On, T. *et al.* The evolutionary landscape of the chromatin modification machinery reveals lineage specific gains, expansions, and losses. *Proteins* **78**, 2075-2089, (2010).

264 Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385-1389, (2010).

265 Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972, (2006).

266 Roger, A. J. & Simpson, A. G. Evolution: revisiting the root of the eukaryote tree. *Curr Biol* **19**, R165-167, (2009).

267 Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287, (2006).

268 Bhattacharya D, Yoon HS, Hedges SB & JD., H. *The Timetree of Life*. 116-120 (Oxford University Press, 2009).

269 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, (2004).

270 Lee, K. C., Webb, R. I. & Fuerst, J. A. The cell cycle of the planctomycete Gemmata obscuriglobus with respect to cell compartmentalization. *BMC Cell Biol* **10**, 4, (2009).

271 Clum, A. *et al.* Complete genome sequence of Pirellula staleyi type strain (ATCC 27377). *Stand Genomic Sci* **1**, 308-316, (2009).

272 Santarella-Mellwig, R. *et al.* The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. *PLoS Biol* **8**, e1000281, (2010).

273 Tanaka, M. *et al.* Analysis of Deinococcus radiodurans's transcriptional response to ionizing radiation and desiccation reveals novel proteins that contribute to extreme radioresistance. *Genetics* **168**, 21-33, (2004).

274 Slade, D. & Radman, M. Oxidative stress resistance in Deinococcus radiodurans. *Microbiol Mol Biol Rev* **75**, 133-191, (2011).

275 Makarova, K. S. *et al.* Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev* **65**, 44-79, (2001).

276 Narula, J., Devi, S. N., Fujita, M. & Igoshin, O. A. Ultrasensitivity of the Bacillus subtilis sporulation decision. *Proc Natl Acad Sci U S A* **109**, E3513-3522, (2012).

277 Douglas, A. E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. *Annu Rev Entomol* **43**, 17-37, (1998).

278 van Ham, R. C. *et al.* Reductive genome evolution in Buchnera aphidicola. *Proc Natl Acad Sci U S A* **100**, 581-586, (2003).

279 Douglas, S. E., Murphy, C. A., Spencer, D. F. & Gray, M. W. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* **350**, 148-151, (1991).

280 Douglas, S. *et al.* The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091-1096, (2001).

281 Gilson, P. R. *et al.* Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A* **103**, 9566-9571, (2006).

282 Curtis, B. A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59-65, (2012).

283 Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403, (1995).

284 Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res* **24**, 4420-4449, (1996).

285 Symonds, E. P., Trott, D. J., Bird, P. S. & Mills, P. Growth characteristics and enzyme activity in Batrachochytrium dendrobatidis isolates. *Mycopathologia* **166**, 143-147, (2008).

286 Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414**, 450-453, (2001).

287 Alvarez, M., Burn, T., Luo, Y., Pirofski, L. A. & Casadevall, A. The outcome of Cryptococcus neoformans intracellular pathogenesis in human monocytes. *BMC Microbiol* **9**, 51, (2009).

288 Loftus, B. J. *et al.* The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307**, 1321-1324, (2005).

289 Abrahamsen, M. S. *et al.* Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science* **304**, 441-445, (2004).

290 Perlmann, P. & Troye-Blomberg, M. Malaria blood-stage infection and its control by the immune system. *Folia Biol (Praha)* **46**, 210-218, (2000).

291 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511, (2002).

292 Ogbadoyi, E., Ersfeld, K., Robinson, D., Sherwin, T. & Gull, K. Architecture of the Trypanosoma brucei nucleus during interphase and mitosis. *Chromosoma* **108**, 501-513, (2000).

293 The Schistosoma japonicum genome reveals features of host-parasite interplay. *Nature* **460**, 345-351, (2009).

294 Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052, (2001).

295 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, (2011).
296 The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640, (2004).
297 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868, (1998).
298 Andreopoulos, B., An, A., Wang, X. & Schroeder, M. A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform* **10**, 297-314, (2009).
299 Eisen, M. B. & Brown, P. O. DNA arrays for analysis of gene expression. *Methods Enzymol* **303**, 179-205, (1999).
300 Saldanha, A. J. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248, (2004).
301 Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910-1912, (2010).
302 Capra, J. A., Williams, A. G. & Pollard, K. S. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput Biol* **8**, e1002567, (2012).
303 Heinicke, S. *et al.* The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One* **2**, e766, (2007).
304 The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* **5**, e1000431, (2009).
305 Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15, (2006).
306 Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-2141, (2003).
307 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128, (2007).
308 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, (2009).
309 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, (2000).
310 Rouse, J. & Jackson, S. P. LCD1: an essential gene involved in checkpoint control and regulation of the MEC1 signalling pathway in Saccharomyces cerevisiae. *EMBO J* **19**, 5801-5812, (2000).
311 Tran, H. J., Allen, M. D., Lowe, J. & Bycroft, M. Structure of the Jab1/MPN domain and its implications for proteasome function. *Biochemistry* **42**, 11460-11465, (2003).
312 Yang, H. *et al.* BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure. *Science* **297**, 1837-1848, (2002).
313 Honda, Y. *et al.* Cooperation of HECT-domain ubiquitin ligase hHYD and DNA topoisomerase II-binding protein for DNA damage response. *J Biol Chem* **277**, 3599-3605, (2002).

314 Yu, X., Fu, S., Lai, M., Baer, R. & Chen, J. BRCA1 ubiquitinates its phosphorylation-dependent binding partner CtIP. *Genes Dev* **20**, 1721-1726, (2006).

315 Malewicz, M. *et al.* Essential role for DNA-PK-mediated phosphorylation of NR4A nuclear orphan receptors in DNA double-strand break repair. *Genes Dev* **25**, 2031-2040, (2011).

316 Kim, H. *et al.* TRF2 functions as a protein hub and regulates telomere maintenance by recognizing specific peptide motifs. *Nat Struct Mol Biol* **16**, 372-379, (2009).

317 Ye, J. *et al.* TRF2 and apollo cooperate with topoisomerase 2alpha to protect human telomeres from replicative damage. *Cell* **142**, 230-242, (2010).

318 Halberg, N. *et al.* Hypoxia-inducible factor 1alpha induces fibrosis and insulin resistance in white adipose tissue. *Mol Cell Biol* **29**, 4467-4483, (2009).

319 Fernandez-Capetillo, O., Celeste, A. & Nussenzweig, A. Focusing on foci: H2AX and the recruitment of DNA-damage response factors. *Cell Cycle* **2**, 426-427, (2003).

320 Bunting, S. F. *et al.* 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243-254, (2010).

321 Polo, S. E., Kaidi, A., Baskcomb, L., Galanty, Y. & Jackson, S. P. Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4. *EMBO J* **29**, 3130-3139, (2010).

322 Ming, M. *et al.* Regulation of global genome nucleotide excision repair by SIRT1 through xeroderma pigmentosum C. *Proc Natl Acad Sci U S A* **107**, 22623-22628, (2010).

323 Raynard, S., Bussen, W. & Sung, P. A double Holliday junction dissolvasome comprising BLM, topoisomerase IIIalpha, and BLAP75. *J Biol Chem* **281**, 13861-13864, (2006).

324 Singh, T. R. *et al.* BLAP18/RMI2, a novel OB-fold-containing protein, is an essential component of the Bloom helicase-double Holliday junction dissolvasome. *Genes Dev* **22**, 2856-2868, (2008).

325 Bae, J. B. *et al.* Snm1B/Apollo mediates replication fork collapse and S Phase checkpoint activation in response to DNA interstrand cross-links. *Oncogene* **27**, 5045-5056, (2008).

326 Caldecott, K. W. XRCC1 and DNA strand break repair. *DNA Repair (Amst)* **2**, 955-969, (2003).

327 Zhang, X. *et al.* Artemis is a phosphorylation target of ATM and ATR and is involved in the G2/M DNA damage checkpoint response. *Mol Cell Biol* **24**, 9207-9220, (2004).

328 Poinsignon, C. *et al.* Phosphorylation of Artemis following irradiation-induced DNA damage. *Eur J Immunol* **34**, 3146-3155, (2004).

329 Jeong, J. *et al.* SIRT1 promotes DNA repair activity and deacetylation of Ku70. *Exp Mol Med* **39**, 8-13, (2007).

330 Peng, L. *et al.* SIRT1 negatively regulates the activities, functions, and protein levels of hMOF and TIP60. *Mol Cell Biol* **32**, 2823-2836, (2012).

331 Nimonkar, A. V. *et al.* BLM-DNA2-RPA-MRN and EXO1-BLM-RPA-MRN constitute two DNA end resection machineries for human DNA break repair. *Genes Dev* **25**, 350-362, (2011).

332 Ostermeier, G. C., Miller, D., Huntriss, J. D., Diamond, M. P. & Krawetz, S. A. Reproductive biology: delivering spermatozoan RNA to the oocyte. *Nature* **429**, 154, (2004).

333 Ciccia, A. *et al.* Identification of FAAP24, a Fanconi anemia core complex protein that interacts with FANCM. *Mol Cell* **25**, 331-343, (2007).

334 Svendsen, J. M. *et al.* Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. *Cell* **138**, 63-77, (2009).

335 Kelsall, I. R., Langenick, J., MacKay, C., Patel, K. J. & Alpi, A. F. The Fanconi anaemia components UBE2T and FANCM are functionally linked to nucleotide excision repair. *PLoS One* **7**, e36970, (2012).

336 Heo, J. I. *et al.* ATM mediates interdependent activation of p53 and ERK through formation of a ternary complex with p-p53 and p-ERK in response to DNA damage. *Mol Biol Rep* **39**, 8007-8014, (2012).

337 Gudjonsson, T. *et al.* TRIP12 and UBR5 suppress spreading of chromatin ubiquitylation at damaged chromosomes. *Cell* **150**, 697-709, (2012).

338 Esashi, F. *et al.* CDK-dependent phosphorylation of BRCA2 as a regulatory mechanism for recombinational repair. *Nature* **434**, 598-604, (2005).

339 Sorensen, C. S. *et al.* The cell-cycle checkpoint kinase Chk1 is required for mammalian homologous recombination repair. *Nat Cell Biol* **7**, 195-201, (2005).

340 Fitch, W. M. Homology a personal view on some of the problems. *Trends Genet* **16**, 227-231, (2000).

341 Korcsmaros, T. *et al.* Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS One* **6**, e19240, (2011).

342 Aravind, L., Anantharaman, V. & Venancio, T. M. Apprehending multicellularity: regulatory networks, genomics, and evolution. *Birth Defects Res C Embryo Today* **87**, 143-164, (2009).

343 Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* **8**, e1002514, (2012).

344 Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**, 1048-1059, (2002).

345 Weinstock, G. M. ENCODE: more genomic empowerment. *Genome Res* **17**, 667-668, (2007).

346 Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13**, 2229-2235, (2003).

347 Borenstein, E., Shlomi, T., Ruppin, E. & Sharan, R. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucleic Acids Res* **35**, e7, (2007).

348 Pagel, P., Wong, P. & Frishman, D. A domain interaction map based on phylogenetic profiling. *J Mol Biol* **344**, 1331-1346, (2004).

349 Wu, J., Hu, Z. & DeLisi, C. Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics* **7**, 80, (2006).

350 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, (2009).

351 Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**, 605-618, (2008).

352 Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-614, (2001).

353 Juan, D., Pazos, F. & Valencia, A. Co-evolution and co-adaptation in protein networks. *FEBS Lett* **582**, 1225-1230, (2008).

354 Yang, S., Doolittle, R. F. & Bourne, P. E. Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A* **102**, 373-378, (2005).

355 Vogel, C., Teichmann, S. A. & Pereira-Leal, J. The relationship between domain duplication and recombination. *J Mol Biol* **346**, 355-365, (2005).

356 Yang, S. & Bourne, P. E. The evolutionary history of protein domains viewed by species phylogeny. *PLoS One* **4**, e8378, (2009).

357 Dore, A. S. *et al.* Structure of an Xrcc4-DNA ligase IV yeast ortholog complex reveals a novel BRCT interaction mode. *DNA Repair (Amst)* **5**, 362-368, (2006).

358 Venancio, T. M., Balaji, S., Iyer, L. M. & Aravind, L. Reconstructing the ubiquitin network: cross-talk with other systems and identification of novel functions. *Genome Biol* **10**, R33, (2009).

359 Zmasek, C. M. & Godzik, A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol* **12**, R4, (2011).

360 Doucet-Chabeaud, G., Godon, C., Brutesco, C., de Murcia, G. & Kazmaier, M. Ionising radiation induces the expression of PARP-1 and PARP-2 genes in Arabidopsis. *Mol Genet Genomics* **265**, 954-963, (2001).

361 Costantini, S., Woodbine, L., Andreoli, L., Jeggo, P. A. & Vindigni, A. Interaction of the Ku heterodimer with the DNA ligase IV/Xrcc4 complex and its regulation by DNA-PK. *DNA Repair (Amst)* **6**, 712-722, (2007).

362 Devos, D. P. & Reynaud, E. G. Evolution. Intermediate steps. *Science* **330**, 1187-1188, (2010).

363 Reynaud, E. G. & Devos, D. P. Transitional forms between the three domains of life and evolutionary implications. *Proc Biol Sci* **278**, 3321-3328, (2011).

364 Devos, D. P. Regarding the presence of membrane coat proteins in bacteria: confusion? What confusion? *Bioessays* **34**, 38-39, (2012).

365 McInerney, J. O. *et al.* Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays* **33**, 810-817, (2011).

366 Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst Zool* **19**, 99-113, (1970).

367 Gnad, F. *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* **8**, R250, (2007).

368 Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834-840, (2009).

369 Zielinska, D. F., Gnad, F., Wisniewski, J. R. & Mann, M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* **141**, 897-907, (2010).

370 Minguez, P. *et al.* Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* **8**, 599, (2012).

371 Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* **1**, (2011).

372 Deribe, Y. L., Pawson, T. & Dikic, I. Post-translational modifications in signal integration. *Nat Struct Mol Biol* **17**, 666-672, (2010).

373 Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54-61, (2007).

374 Srivastava, M. *et al.* The Trichoplax genome and the nature of placozoans. *Nature* **454**, 955-960, (2008).

375 Cerutti, H. & Casas-Mollano, J. A. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* **50**, 81-99, (2006).

376 Zofall, M. & Grewal, S. I. RNAi-mediated heterochromatin assembly in fission yeast. *Cold Spring Harb Symp Quant Biol* **71**, 487-496, (2006).

377 Lorentzen, E. & Conti, E. The exosome and the proteasome: nano-compartments for degradation. *Cell* **125**, 651-654, (2006).

378 Fiedler, D. *et al.* Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-963, (2009).

379 Eichinger, L. *et al.* The genome of the social amoeba Dictyostelium discoideum. *Nature* **435**, 43-57, (2005).

380 Goldberg, J. M. *et al.* The dictyostelium kinome--analysis of the protein kinases from a simple model organism. *PLoS Genet* **2**, e38, (2006).

381 Shuman, S. & Glickman, M. S. Bacterial DNA repair by non-homologous end joining. *Nat Rev Microbiol* **5**, 852-861, (2007).

382 Sachadyn, P. Conservation and diversity of MutS proteins. *Mutat Res* **694**, 20-30, (2010).

383 Marti, T. M., Kunz, C. & Fleck, O. DNA mismatch repair and mutation avoidance pathways. *J Cell Physiol* **191**, 28-41, (2002).

384 Laguri, C. *et al.* Human mismatch repair protein MSH6 contains a PWWP domain that targets double stranded DNA. *Biochemistry* **47**, 6199-6207, (2008).

385 Mortusewicz, O., Fouquerel, E., Ame, J. C., Leonhardt, H. & Schreiber, V. PARG is recruited to DNA damage sites through poly(ADP-ribose)- and PCNA-dependent mechanisms. *Nucleic Acids Res* **39**, 5045-5056, (2011).

386 Sukhanova, M., Khodyreva, S. & Lavrik, O. Poly(ADP-ribose) polymerase 1 regulates activity of DNA polymerase beta in long patch base excision repair. *Mutat Res* **685**, 80-89, (2010).

387 Dantzer, F. *et al.* Functional interaction between poly(ADP-Ribose) polymerase 2 (PARP-2) and TRF2: PARP activity negatively regulates TRF2. *Mol Cell Biol* **24**, 1595-1607, (2004).

388 Bryant, H. E. *et al.* PARP is activated at stalled forks to mediate Mre11-dependent replication restart and recombination. *EMBO J* **28**, 2601-2615, (2009).

389 Huber, A., Bai, P., de Murcia, J. M. & de Murcia, G. PARP-1, PARP-2 and ATM in the DNA damage response: functional synergy in mouse development. *DNA Repair (Amst)* **3**, 1103-1108, (2004).

390 Schreiber, V. *et al.* Poly(ADP-ribose) polymerase-2 (PARP-2) is required for efficient base excision DNA repair in association with PARP-1 and XRCC1. *J Biol Chem* **277**, 23028-23036, (2002).

391 Otto, H. *et al.* In silico characterization of the family of PARP-like poly(ADP-ribosyl)transferases (pARTs). *BMC Genomics* **6**, 139, (2005).

392 Uchiyama, Y., Suzuki, Y. & Sakaguchi, K. Characterization of plant XRCC1 and its interaction with proliferating cell nuclear antigen. *Planta* **227**, 1233-1241, (2008).

393 Doetsch, P. W., Morey, N. J., Swanson, R. L. & Jinks-Robertson, S. Yeast base excision repair: interconnections and networks. *Prog Nucleic Acid Res Mol Biol* **68**, 29-39, (2001).

394 Legrand, M., Chan, C. L., Jauert, P. A. & Kirkpatrick, D. T. Analysis of base excision and nucleotide excision repair in Candida albicans. *Microbiology* **154**, 2446-2456, (2008).

395 Araujo, S. J. *et al.* Nucleotide excision repair of DNA with recombinant human proteins: definition of the minimal set of factors, active forms of TFIIH, and modulation by CAK. *Genes Dev* **14**, 349-359, (2000).

396 Petit, C. & Sancar, A. Nucleotide excision repair: from E. coli to man. *Biochimie* **81**, 15-25, (1999).

397 White, M. F. Archaeal DNA repair: paradigms and puzzles. *Biochem Soc Trans* **31**, 690-693, (2003).

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

399 Doherty, A. J., Jackson, S. P. & Weller, G. R. Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS Lett* **500**, 186-188, (2001).

400 Aravind, L. & Koonin, E. V. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* **11**, 1365-1374, (2001).

401 Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686-1689, (2002).

402 Gu, J. & Lieber, M. R. Mechanistic flexibility as a conserved theme across 3 billion years of nonhomologous DNA end-joining. *Genes Dev* **22**, 411-415, (2008).

403 Smith, P., Nair, P. A., Das, U., Zhu, H. & Shuman, S. Structures and activities of archaeal members of the LigD 3'-phosphoesterase DNA repair enzyme superfamily. *Nucleic Acids Res* **39**, 3310-3320, (2011).

404 Callebaut, I. *et al.* Cernunnos interacts with the XRCC4 x DNA-ligase IV complex and is homologous to the yeast nonhomologous end-joining factor Nej1. *J Biol Chem* **281**, 13857-13860, (2006).

405 Lieber, M. R. NHEJ and its backup pathways in chromosomal translocations. *Nat Struct Mol Biol* **17**, 393-395, (2010).

406 Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev* **72**, 642-671, Table of Contents, (2008).

407 Lin, Z., Kong, H., Nei, M. & Ma, H. Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci U S A* **103**, 10328-10333, (2006).

408 Hopkins, B. B. & Paull, T. T. The P. furiosus mre11/rad50 complex promotes 5' strand resection at a DNA double-strand break. *Cell* **135**, 250-260, (2008).

409 Ueno, M. *et al.* Molecular characterization of the Schizosaccharomyces pombe nbs1+ gene involved in DNA repair and telomere maintenance. *Mol Cell Biol* **23**, 6553-6563, (2003).

410 Bressan, D. A., Baxter, B. K. & Petrini, J. H. The Mre11-Rad50-Xrs2 protein complex facilitates homologous recombination-based double-strand break repair in Saccharomyces cerevisiae. *Mol Cell Biol* **19**, 7681-7687, (1999).

411 Hammet, A., Magill, C., Heierhorst, J. & Jackson, S. P. Rad9 BRCT domain interaction with phosphorylated H2AX regulates the G1 checkpoint in budding yeast. *EMBO Rep* **8**, 851-857, (2007).

412 Downs, J. A. *et al.* Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. *Mol Cell* **16**, 979-990, (2004).

413 Voss, T. S., Mini, T., Jenoe, P. & Beck, H. P. Plasmodium falciparum possesses a cell cycle-regulated short type replication protein A large subunit encoded by an unusual transcript. *J Biol Chem* **277**, 17493-17501, (2002).

414 Garcia-Dominguez, M., March-Diaz, R. & Reyes, J. C. The PHD domain of plant PIAS proteins mediates sumoylation of bromodomain GTE proteins. *J Biol Chem* **283**, 21469-21477, (2008).

415 Boudolf, V., Inze, D. & De Veylder, L. What if higher plants lack a CDC25 phosphatase? *Trends Plant Sci* **11**, 474-479, (2006).

416 De Schutter, K. *et al.* Arabidopsis WEE1 kinase controls cell cycle arrest in response to activation of the DNA integrity checkpoint. *Plant Cell* **19**, 211-225, (2007).

417 Huang, L. *et al.* Three tandem HRDC domains have synergistic effect on the RecQ functions in Deinococcus radiodurans. *DNA Repair (Amst)* **6**, 167-176, (2007).

GLOSSARY

List of abbreviations (in alphabetical order)

9-1-1	RAD9, RAD1 and HUS1 complex
A-NHEJ	Alternative non-homologous end-joining
AP site	Apurinic or apyrimidinic site
BBH	Best bidirectional hit
BER	Base excision repair
C-NHEJ	Classic non-homologous end-joining
CPDs	Cyclobutane pyrimidine dimers
DDR	DNA damage response
DNA-PKcs	DNA-dependent protein kinase catalytic subunit
DSBs	Double-strand breaks
GG-NER	Global genome nucleotide excision repair
GO	Gene Ontology
HGT	Horizontal gene transfer
HJs	Holliday junctions
HMM	Hidden Markov model
HR	Homologous recombination
ICLR	Interstrand crosslink repair
IR	Ionizing radiation
IRIF	Ionizing radiation-induced foci
LP-BER	Long-patch base excision repair
LSEs	Lineage-Specific Expansions
MMR	Mismatch repair
MPI	Message Passing Interface
MRN	Mre11–Rad50–Nbs1 complex
MSA	Multiple sequence alignment
MYA	Millions of years ago
ncRNAs	Non-coding RNAs
NER	Nucleotide excision repair
NHEJ	Non-homologous end-joining
PIKKs	Phosphatidylinositol 3-kinase-related kinases
PTMs	Post-translational modifications
RBPs	RNA-binding proteins
ROS	Reactive oxigen species
SCF	SKP1-cullin-F-box protein ligase complex
SP-BER	Short-patch base excision repair
SSBs	Single-strand breaks
SUMO	Small ubiquitin-like modifier
TC-NER	Transcription-coupled nucleotide excision repair
TFIIH	Transcription factor IIH
TLS	Translesion synthesis
UV	Ultraviolet

Concepts

Clade: A group of taxa that forms a monophyletic unit. It is applicable to any level of the taxonomical hierarchy.

Differential gene loss: The loss of opposite copies in a pair of paralogs between two cells that inherited the orthologs from a common genome duplication.

Holliday junction (HR): mobile junction between four strands of homologous DNA sequences. These highly conserved structures (from prokaryotes to mammals) are an intermediate in genetic recombination with an important role in maintaining genomic integrity.

Horizontal gene transfer (HGT): A process by which an organism incorporates genetic material from another organism that does not belong to its line of ancestry. This process is also called lateral gene transfer.

Ionizing radiation (IR): radiation composed of particles that carry enough kinetic energy to liberate an electron from an atom or molecule, ionizing it. Such an event can alter and break chemical bonds causing enormous biological damage.

Multiple sequence alignment (MSA): alignment of three or more DNA, RNA or protein sequences. Generally, from the resulting alignment, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origin.

Neo-functionalization: The acquisition of a novel function by a gene after mutational changes. This usually applies to one of the two paralogues that are produced from gene duplication.

Nucleomorph: A reduced eukaryotic nuclei found in certain plastids from cryptomonads and chlorarachniophytes. They are thought to be vestiges of primitive red and green algal nuclei and seem to be evidence of the evolutionary origin of plastids by endosymbiosis.

Phylum: taxonomic rank below kingdom and above class.

Post-translational modifications (PTMs): After translation, a protein can be modified to vary its range of functions by attaching biochemical functional groups (for example, a phosphate or acetyl group) to it, adding other proteins or peptides (e.g. SUMOylation or ubiquitination), changing the chemical nature of an amino acid (e.g. deamination), or making structural changes (e.g. formation of disulfide bridges).
PTMs modulate the function of most eukaryotic proteins by altering their activity state, localization, turnover, and interactions with other proteins.

Sub-functionalization: Functional specialization after mutational changes of the paralogues that are produced from a gene duplication.

Taxon: group of one or more populations of organisms that reflects evolutionary (phylogenetic) relationships.

PUBLICATIONS

San Martin-Uriz, P., Gómez, MJ., **Arcas, A**., Bargiela, RM., and Amils, R. "Draft Genome Sequence of the Electricigen Acidiphilium sp. strain PM (DSM 24941)". *Journal of Bacteriology*. 2011. Oct; 193(19):5585-6. (PMID: 21914891)

Moreno-Paz, M., Gómez, MJ., **Arcas, A.** and Parro, V. "Environmental transcriptome analysis reveals physiological differences between biofilm and planktonic modes of life of the iron oxidizer bacterium *Leptospirillum ferrooxidans* in its natural microbial community". *BMC Genomics.* 2010. Jun 24; 11: 404. (PMID: 20576116)

ANNEX

List of figures and tables in Annex

Gene-trees

Only proteins with representatives in early eukaryotes have been selected to conduct phylogenetic analyses. The figure below is a modification of Figure 17, where only selected trees are shown. Trees are classified by blocks.

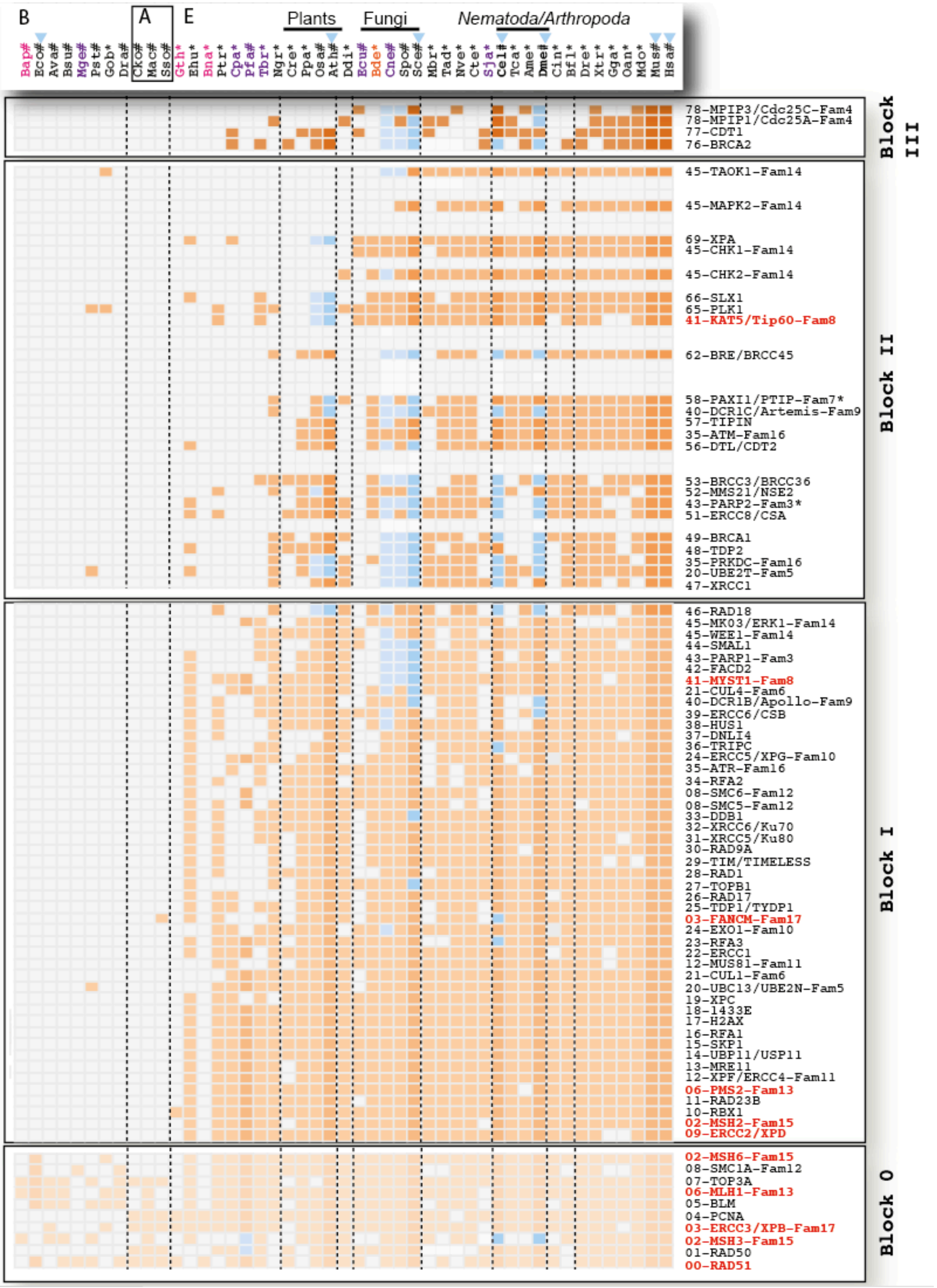
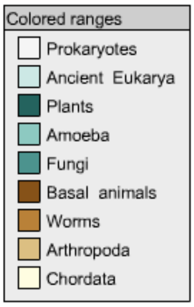


Figure SF1. Illustration indicating genes selected to conduct phylgenies.
Red names are genes/proteins explained and described in the main text.

The dots in the tree branches indicate clades with probability value higher than 80% or 0.8. In the case of differential domain architecture among orthologs, sequences are depicted showing their relative length, being the N-terminal region the closest displayed to the centre of the tree.



To easily distinguish the different organisms by taxonomic range, species are coloured using the following schema depicted at the left.

BLOCK 0:

01-RAD50

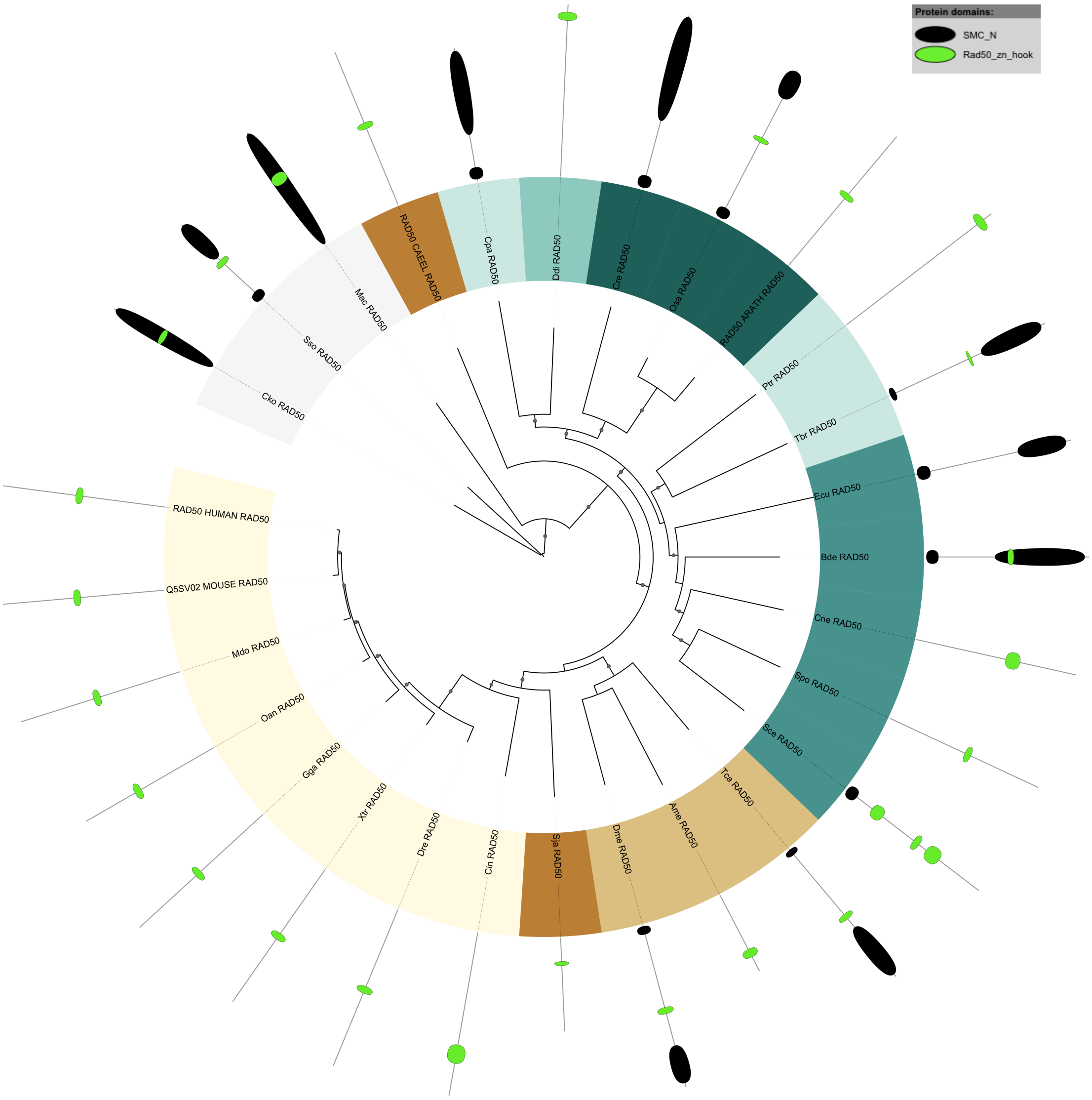


Figure STt1: phylogenetic tree of DNA repair protein **RAD50** (01). The *C. elegans* ortholog is misplaced according to its phylogenetic group, close to the base of the tree. Chordata miss the SMC_N termimal domain. In some cases domain boundaries are unclear as in the case of Prokaryotes.

04-PCNA

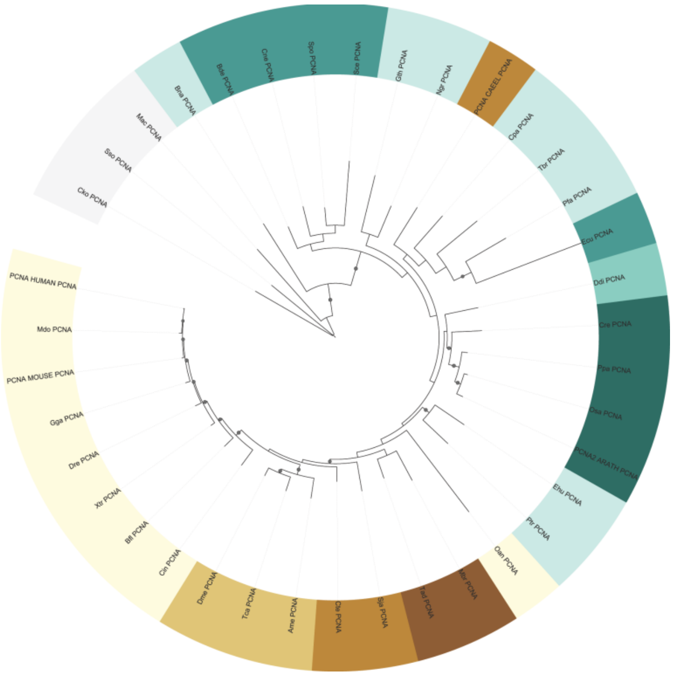


Figure STt2: phylogenetic tree of Proliferating cell nuclear antigen (**PCNA**) (04). *C. elegans* is found outside its phylogenetic group, among the ancient eukaryotes sequences. *O. anatinus* is unexpectedly placed prior to basal animals, probably due to the sequence having been incorrectly predicted.

05-BLM

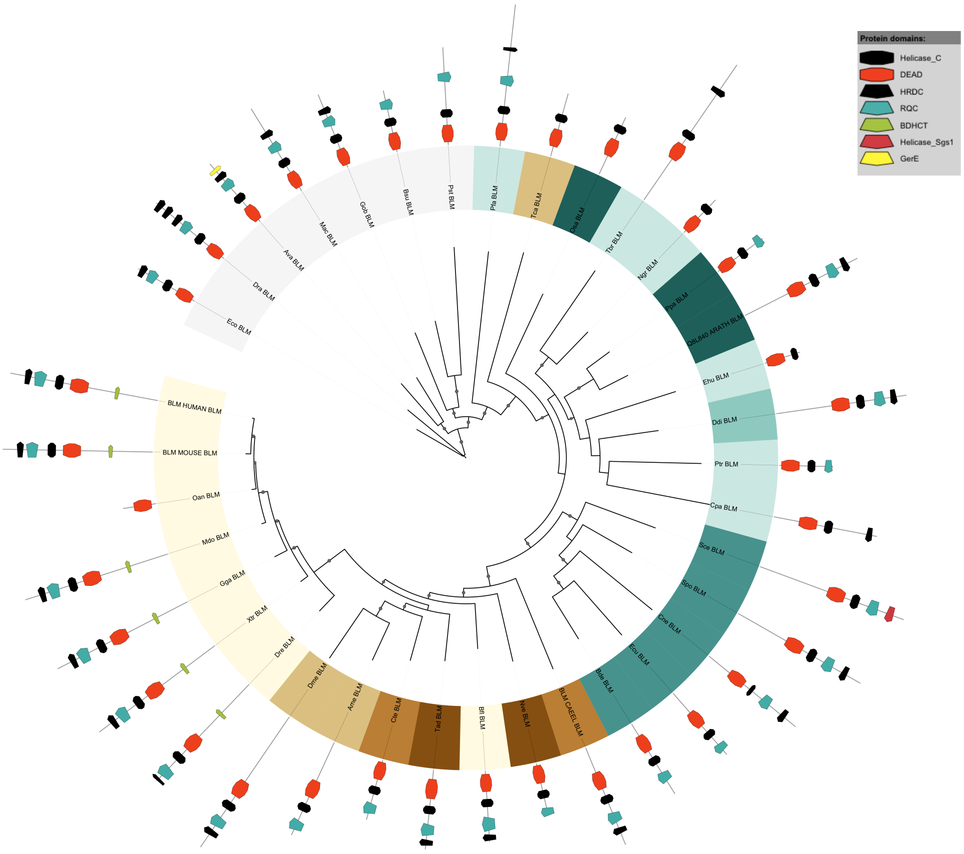


Figure STt3: phylogenetic tree of Bloom syndrome protein **BLM** (05). The *T. castaneum* sequence clusters before plants, probably due to the fact that the sequence is incomplete since it missing some domains. As in other trees, the problematic *C. elegans* is found before basal animals.

07-TOP3A

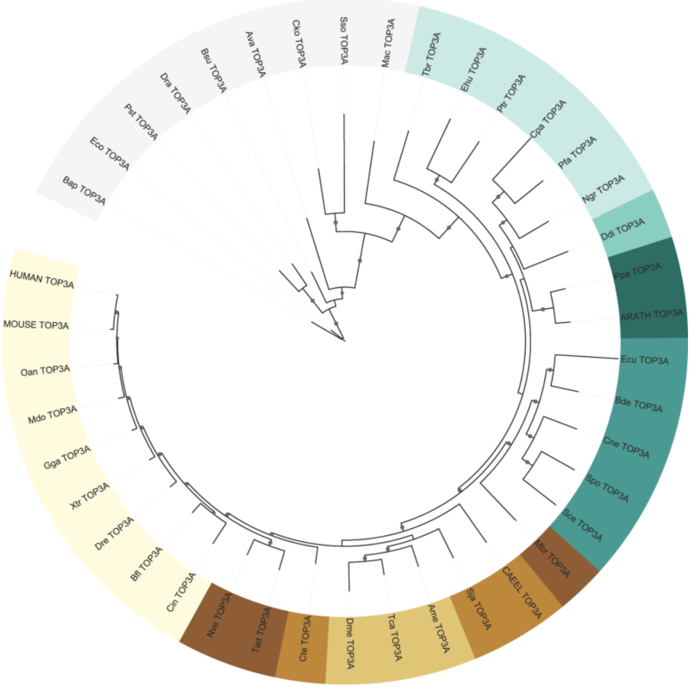


Figure STt4: phylogenetic tree of DNA topoisomerase 3-alpha **TOP3A** (07). *N. vectensis* and *T. adhaerens* cluster closer to *chordata* than with the other basal eukaryote *M. brevicollis*.

08-Family 12: SMC1A, SMC5 and SMC6

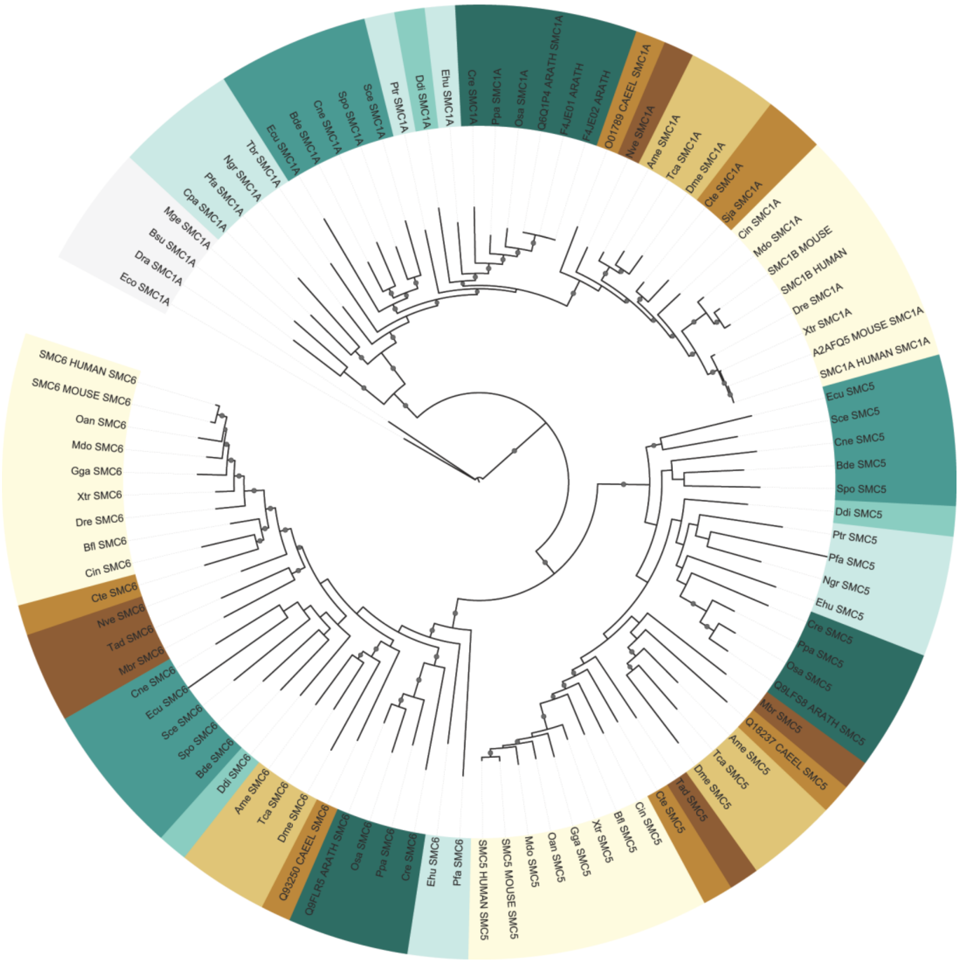


Figure STt5: phylogenetic tree of Family12 (08), comprising Structural maintenance of chromosomes proteins **SMC1A**, **SMC5** and **SMC6**. In SMC1A fungi cluster before plants, and *C. elegans* appears before basal animals. In SMC5, *T. adhaerens* and *C. teleta* group with *chordata* after *arthropoda*. In SMC6, *C. elegans* and *arthropoda* cluster between plants and fungi. These results suggest this family has likely suffered HGTs.

BLOCK I:

10-RBX1

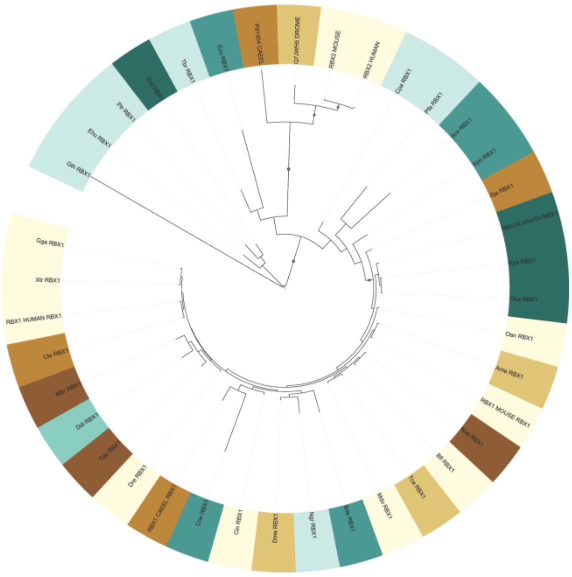


Figure STt6: phylogenetic tree of E3 ubiquitin-protein ligase **RBX1** (10). This tree is erroneous; most branches are unsupported and phylogenetic group are mixed. Either the orthologs have been incorrectly detected by InParanoid, or the MSA should be redone as well as the phylogenetic tree, which might need more sampling generations.

11-RAD23B

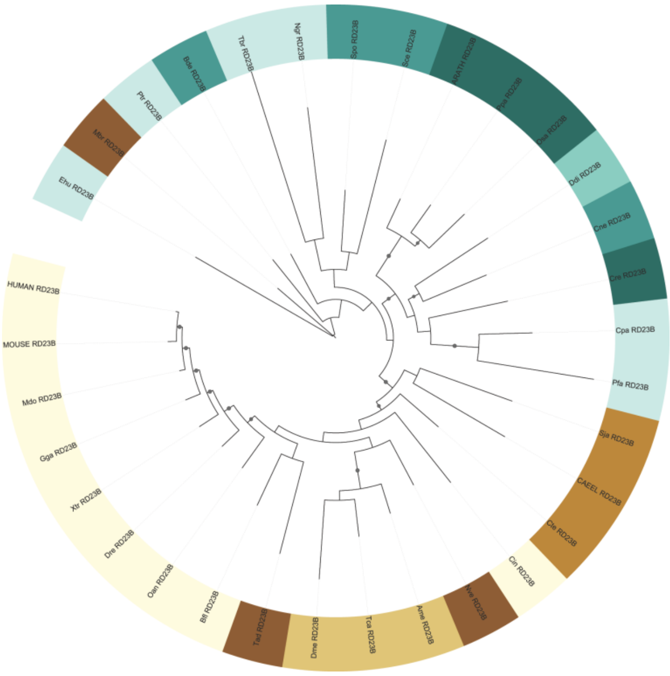


Figure STt7: phylogenetic tree of UV excision repair protein **RAD23** homolog **B** (11). The *M. brevicollis* ortholog is surprisingly placed at the base of the tree, which is well supported at the main taxa (plants, animals), but not at splits. Consequently, the data should be checked and more sampling is required.

12- Family11: XPF, MUS81 and EME1.

A tree with the orthologs of the three proteins in this family was constructed (data not shown), but despite being strongly supported, long branches in XPF were very different from the ones in the other two proteins, so we split the genes into individual trees. EME1 and MUS81 are more similar, and it seems that XPF is the most ancient sequence.

12-XPF

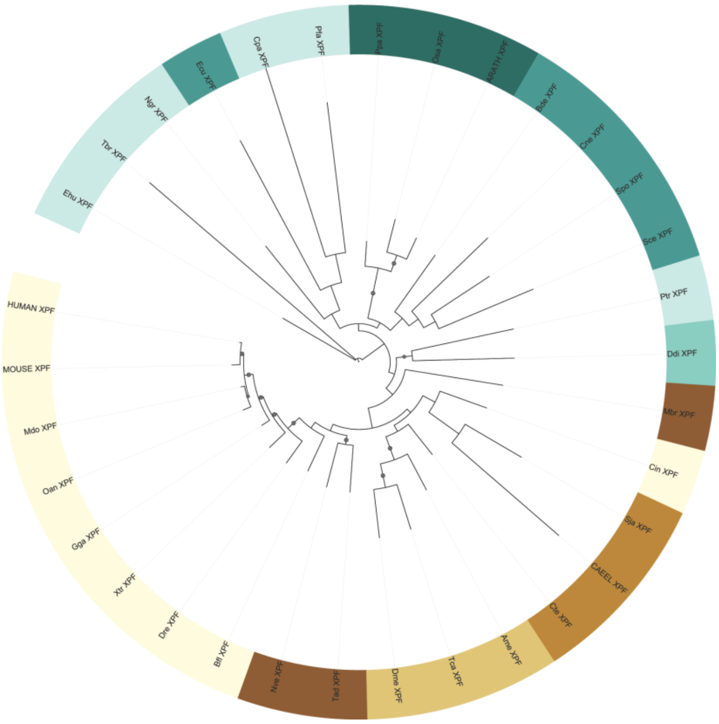


Figure STt8: phylogenetic tree of DNA repair endonuclease **XPF** (12). The tree is not well supported at split points. The *N. vectensis* and *T. adhaerens* sequences cluster with *chordata* instead of with *M. brevicollis* (like in the TOP3A tree), and the *C. intestinalis* ortholog groups with worms.

12-MUS81

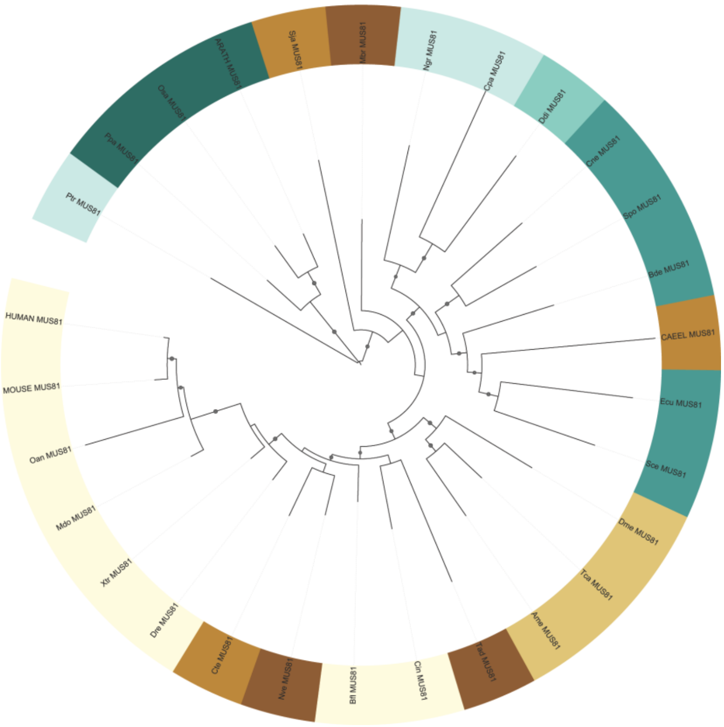


Figure STt9: phylogenetic tree of Crossover junction endonuclease **MUS81** (12). Though most branches are well supported, this tree shows several differences with the species-tree: *S. japonicum* and *M. brevicollis* are placed before some ancient eukaryotes, *C. elegans* clusters with fungi, arthropods appear before basal animals, and *N. vectensis* and *C. teleta* cluster among chordata.

12-EME1

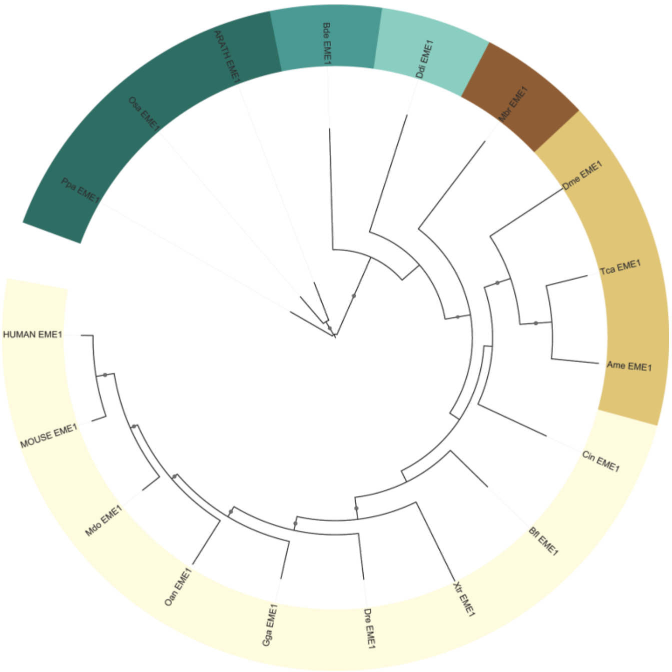


Figure STt10: phylogenetic tree of Crossover junction endonuclease **EME1** (12). This tree is well supported and follows the species-tree order.

13-MRE11

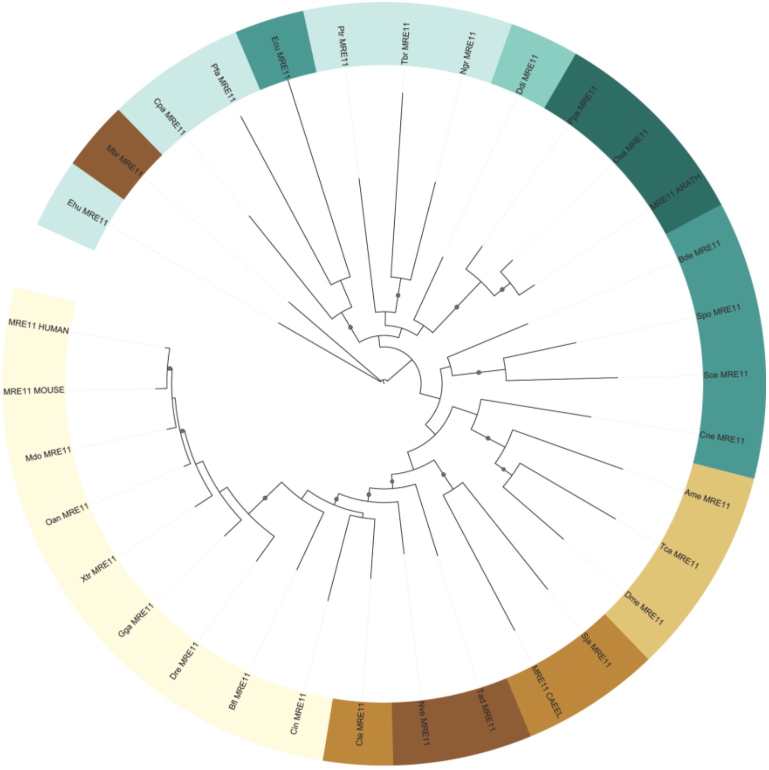


Figure STt11: phylogenetic tree of Double-strand break repair protein **MRE11A** (13). As in previous trees, the *M. brevicollis* sequence is placed with ancient eukaryotes, and arthropods cluster prior to worms and basal metazoa.

14-USP11

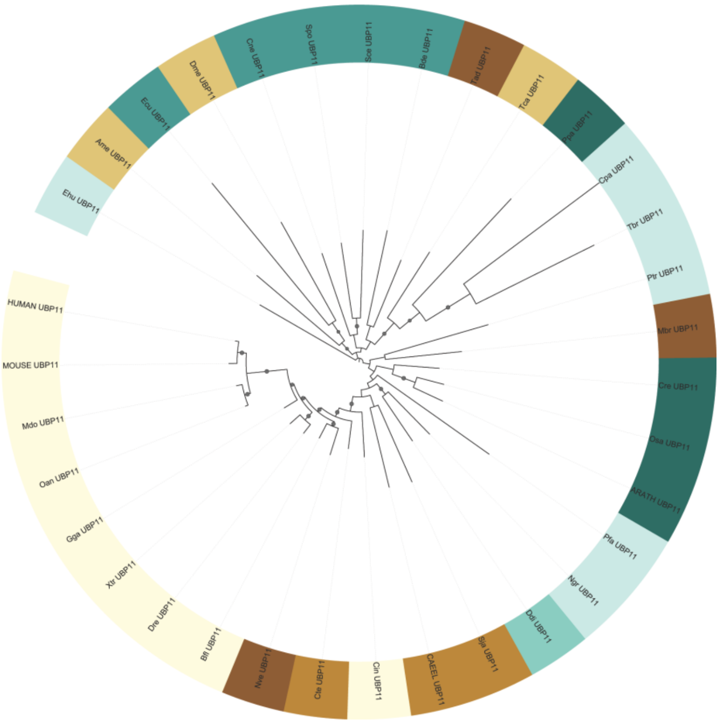


Figure STt12: phylogenetic tree of Ubiquitin carboxyl-terminal hydrolase 11 (**USP11**) (14). Though well supported in most branches, this tree is messy since phylogenetic group are mixed. Either the orthologs have been incorrectly detected by InParanoid, or the MSA should be redone as well as the phylogenetic tree, which might need more sampling generations.

15-SKP1

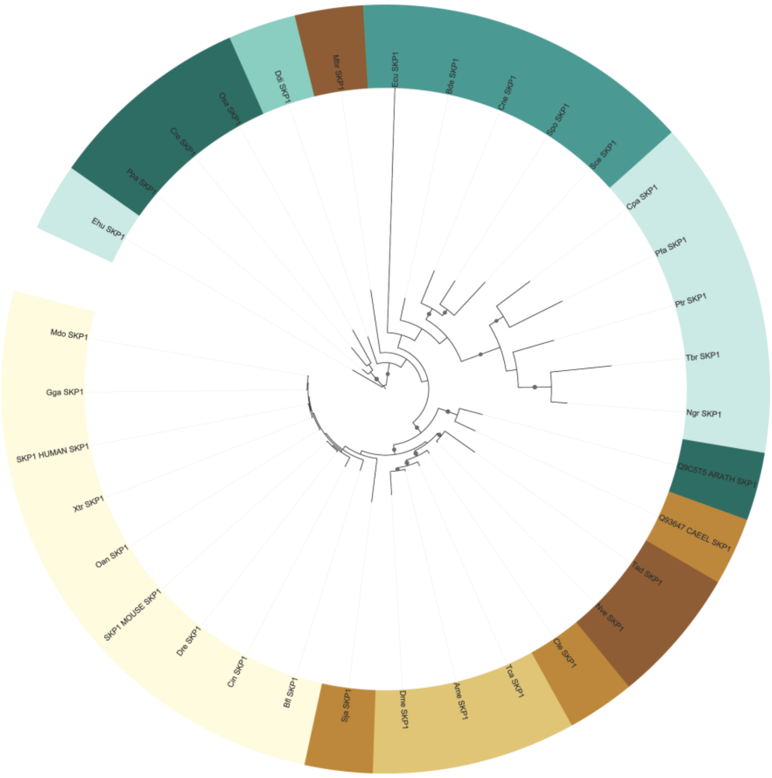


Figure STt13: phylogenetic tree of S-phase kinase-associated protein 1 (SKP1) (15). In this tree, the *M. brevicollis* sequence is placed again away from the rest of basal eukaryotes, and all plants but *A. thaliana* cluster prior to most ancient eukaryotes.

16-RFA1

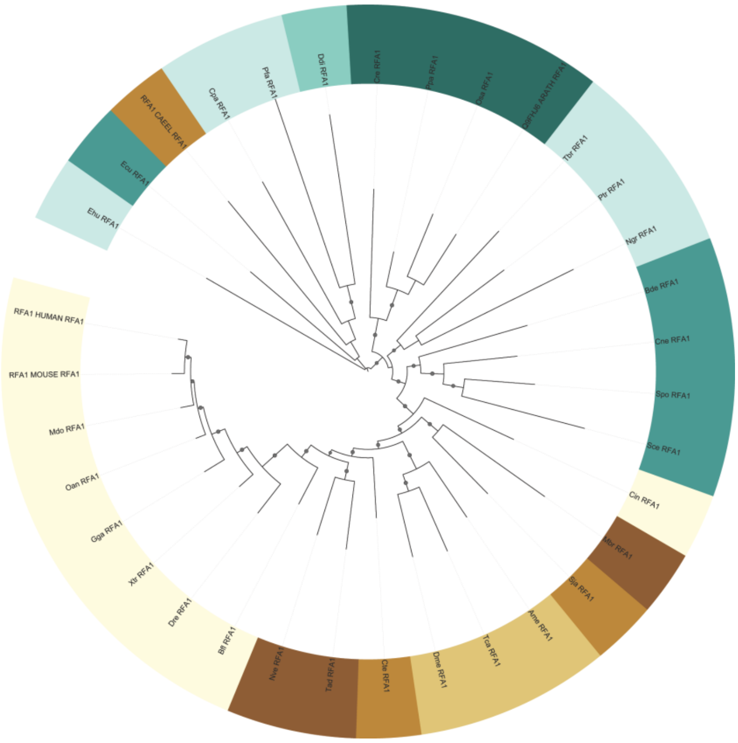


Figure STt14: phylogenetic tree of Replication protein A 70 kDa DNA-binding subunit (**RFA1**) (16). In this highly supported tree, the *C. elegans* and the *E. cuniculi* orthologs are placed with ancient eukaryotes, and the *C. intestinalis* sequence appears before basal eukaryotes. As in previous trees, *N. vectensis* and *T. adhaerens* cluster closer to *chordata* than to *M. brevicollis*.

17-H2AX

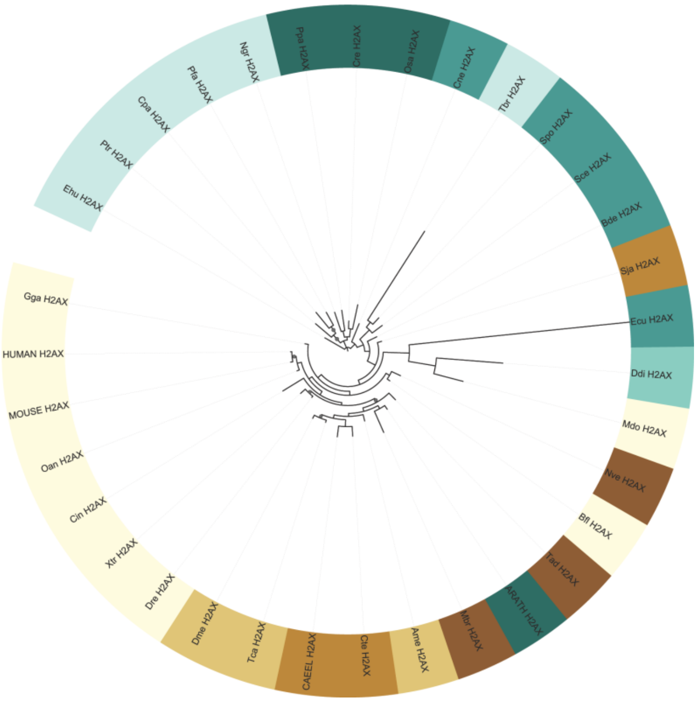


Figure STt15: phylogenetic tree of Histone **H2AX** (17). This tree is not supported but in few branches. Sequences should be revised and maybe more sampling is needed.

18- 14-3-3E

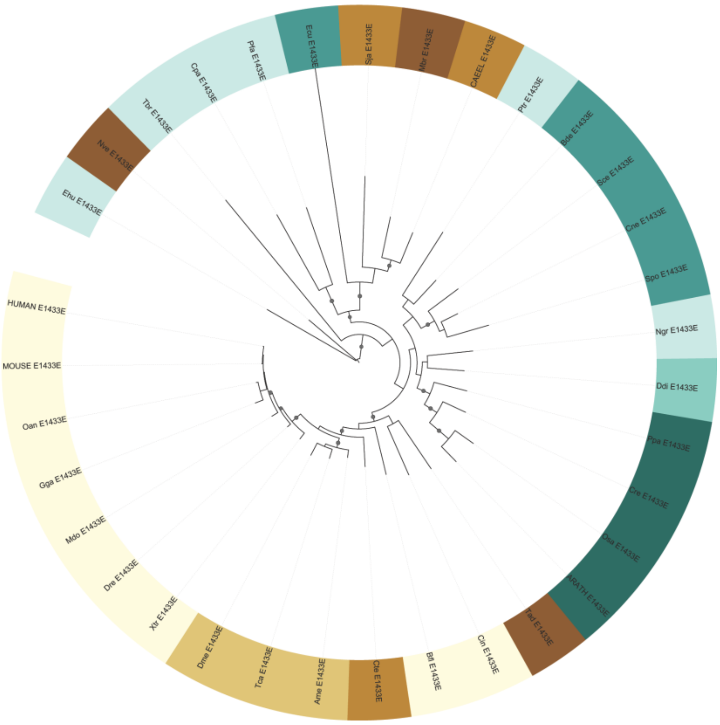


Figure STt16: phylogenetic tree of **14-3-3** protein epsilon (18). The only worm cladding as expected in this tree is the annelid *C. teleta*. The other worms, *N. vectensis* and *M. brevicollis* group with ancient eukaryotes. This might be due to HGT events or to the erroneus identification by InParanoid of proteins from the 14-3-3 family.

19-XPC

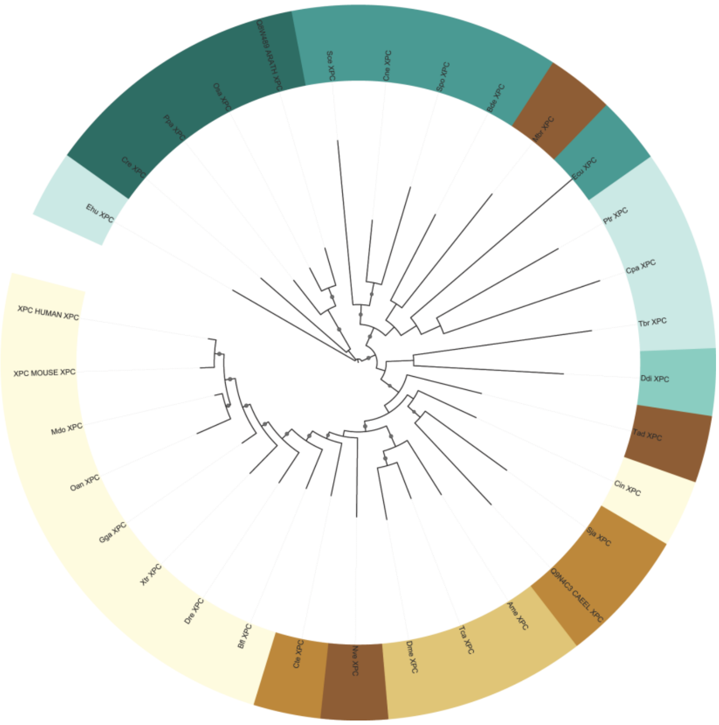


Figure STt17: phylogenetic tree of Xeroderma pigmentosum group C-complementing protein (**XPC**) (19). In this well supported tree, the *M. brevicollis* sequence is once more found away from the other basal eukaryotes, and *C. intestinalis* is found between *T. adhaerens* and *arthropoda*.

20- Family 5: UBE2N/T

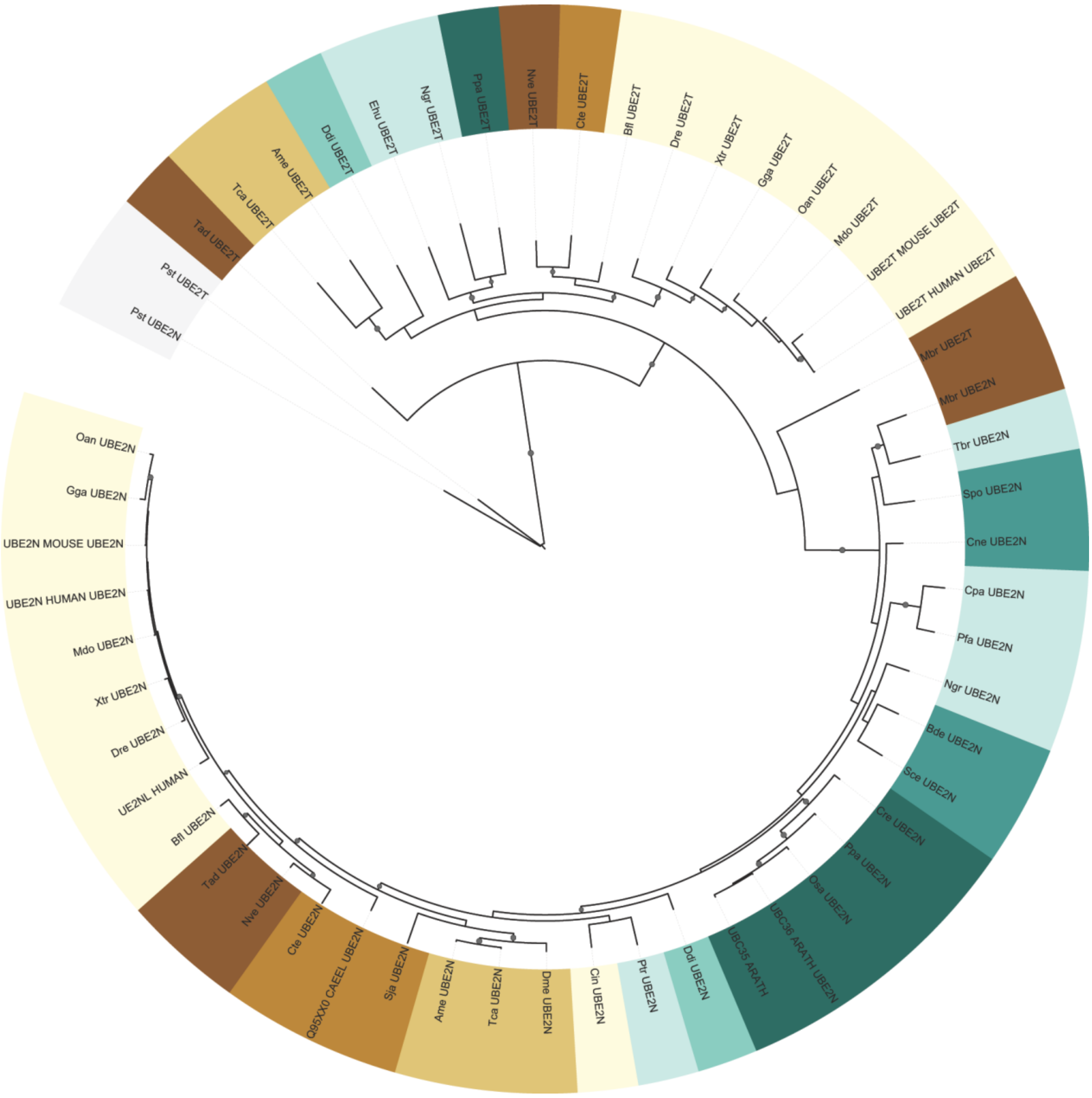


Figure STt18: phylogenetic tree of Family5 (20), comprising Ubiquitin-conjugating enzyme E2 proteins N (UBC13) and T. This tree is not well supported, and many phylogenetic groups are mixed in the clades. UBE2N from *P. staleyi* and *M. brevicollis* group with the UBE2T sequences instead of with the other UBE2N orthologs, which could be caused by this family having its origin in bacteria, from which the gene was transferred to other organisms by HGT.

21- Family 6: CUL1/4

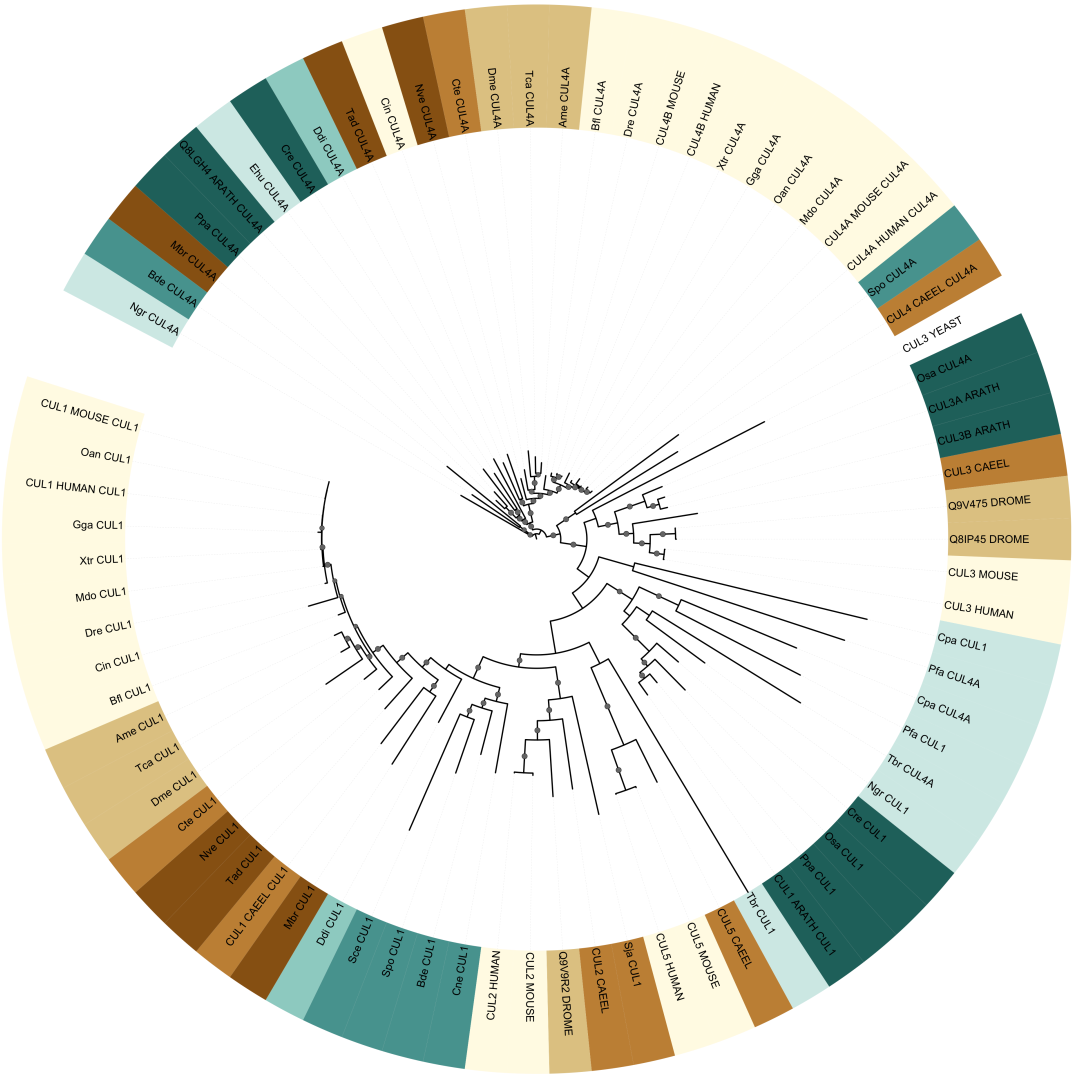


Figure STt19: phylogenetic tree of Family6 (21), comprising Cullin-1 and 4. Besides CUL1 and 4, this strongly supported tree includes in-paralogs from the Cullin family of proteins. CUL4 seems to be the ancient sequence, and CUL5 appears to be a duplication from CUL1. *S. japonicum* CUL1 groups with the CUL2 sequences, so this ortholog has probably been incorrectly detected by InParanoid. The CUL1 sequences follow almost perfectly the species-tree order.

22- ERCC1

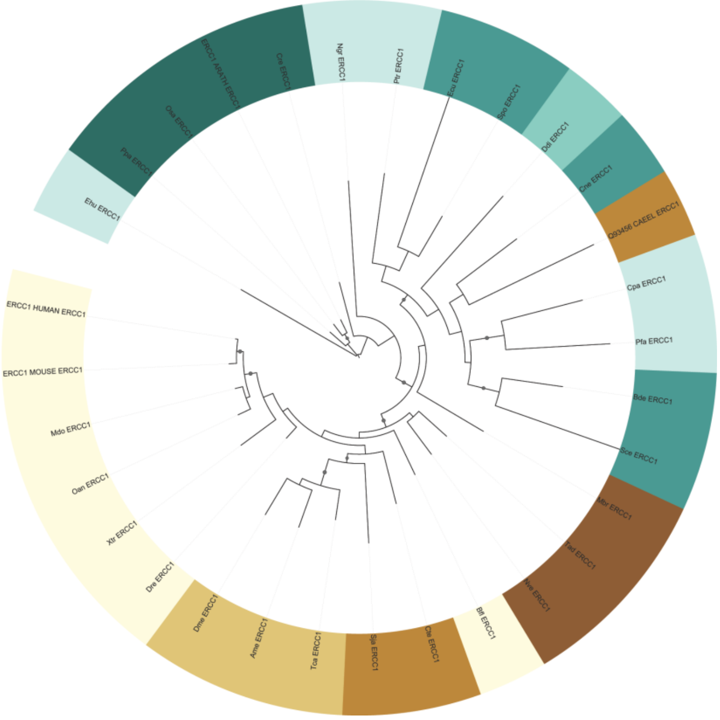


Figure STt20: phylogenetic tree of DNA excision repair protein **ERCC1** (22). This tree follows quite well the species-tree order, but for *C. elegans*, found cladding with fungi and ancient eukaryotes. Also, plants cluster at the base of the tree.

23-RFA3

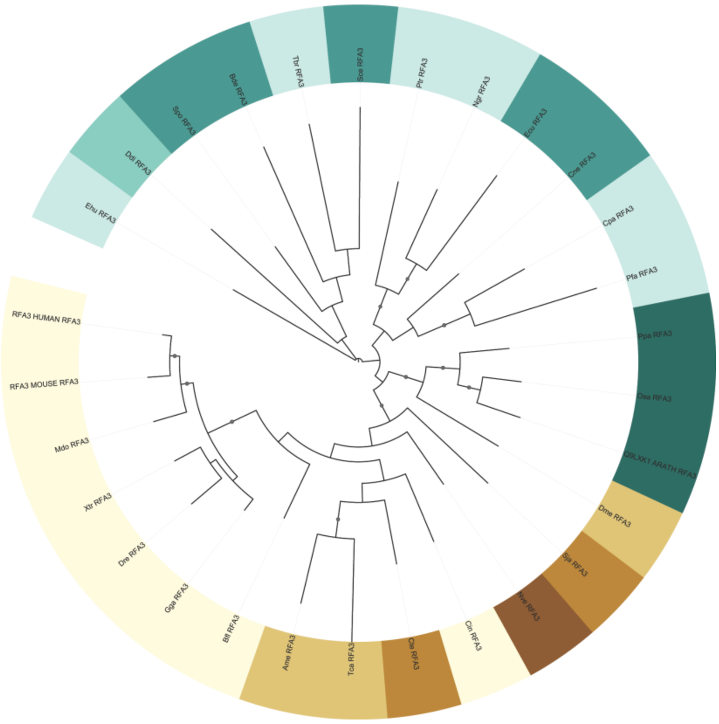


Figure STt21: phylogenetic tree of Replication protein A 14 kDa subunit (**RFA3**) (23). This tree is poorly supported and would need more sampling since the fungi and ancient eukaryotes sequences are mixed. Also, the *D. melanogaster* sequence wrongly clusters with plants.

24-ERCC5

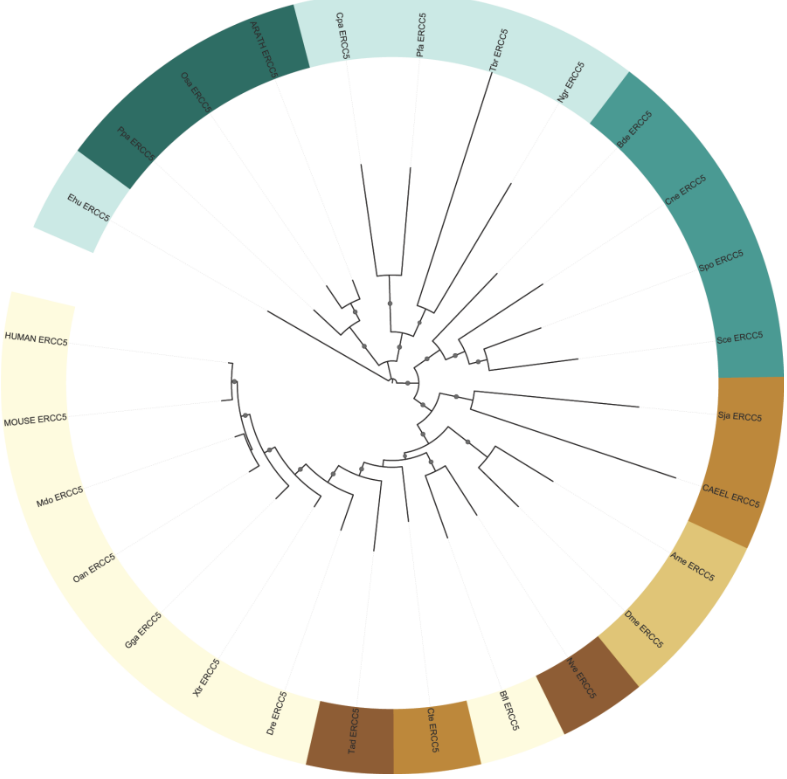


Figure STt22: phylogenetic tree of DNA excision repair protein **ERCC5** (24). In this well supported tree plants are cladding before most ancient eukaryotes, and worms and *arthropoda* prior to basal animals.

25-TYDP1

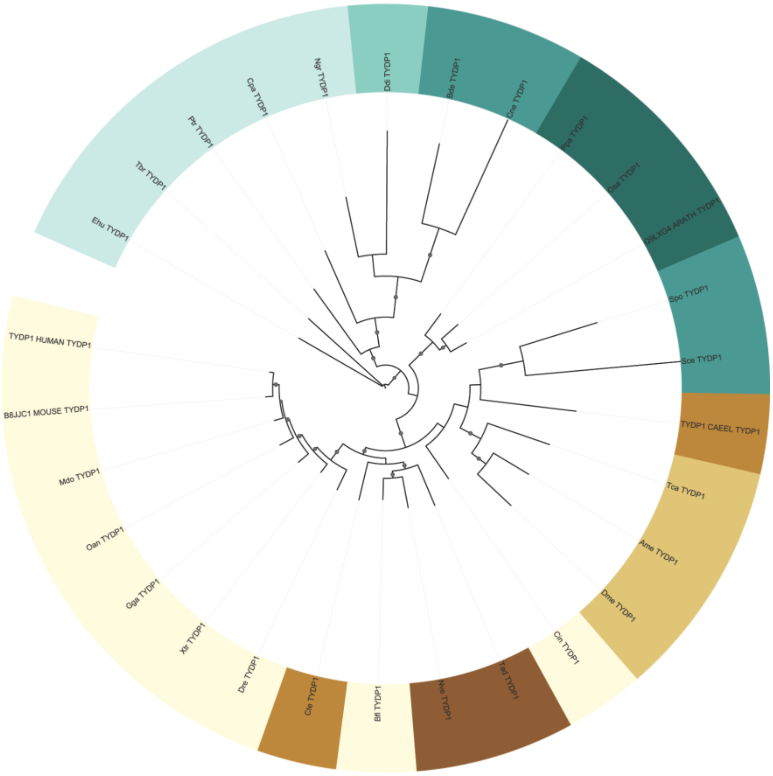


Figure STt23: phylogenetic tree of Tyrosyl-DNA phosphodiesterase 1 (**TYDP1**) (25). This tree is well supported at deep branches, but plants clade between fungi and yeast, the worms *N. vectensis* and *T. adhaerens* cluster among *chordata* (as in some previous trees), and *C. teleta* is placed also with *chordata*, instead of with *C. elegans*.

26-RAD17

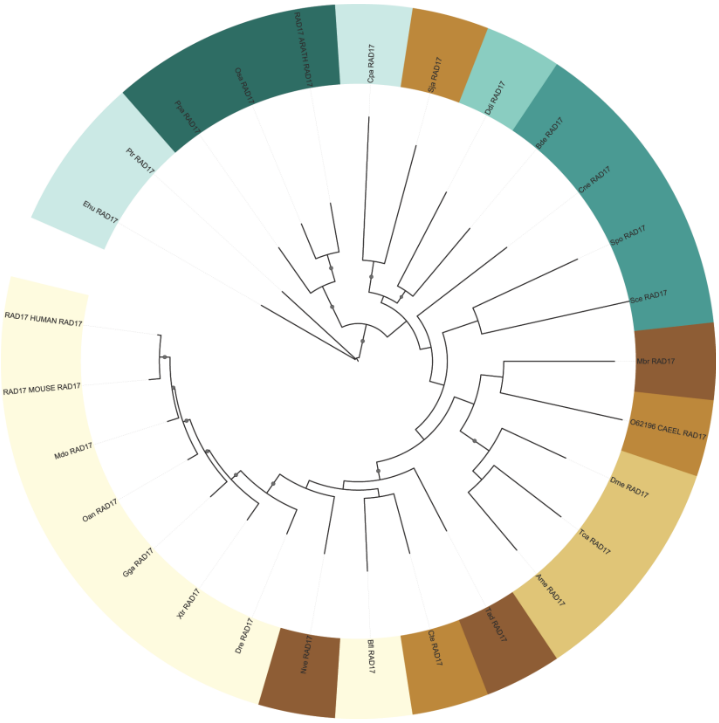


Figure STt24: phylogenetic tree of Cell cycle checkpoint protein **RAD17** (26). This tree is poorly supported at deep branches and might need more sampling since basal metazoa group after arthropoda. Besides, the *S. japonicum* sequence clades with ancient eukaryotes. Nevertheless, this gene-tree is similar to the one obtained for Topb1, being both proteins part of the same complex at stalled replication forks, which might have evolutionary significance.

27-TOPB1

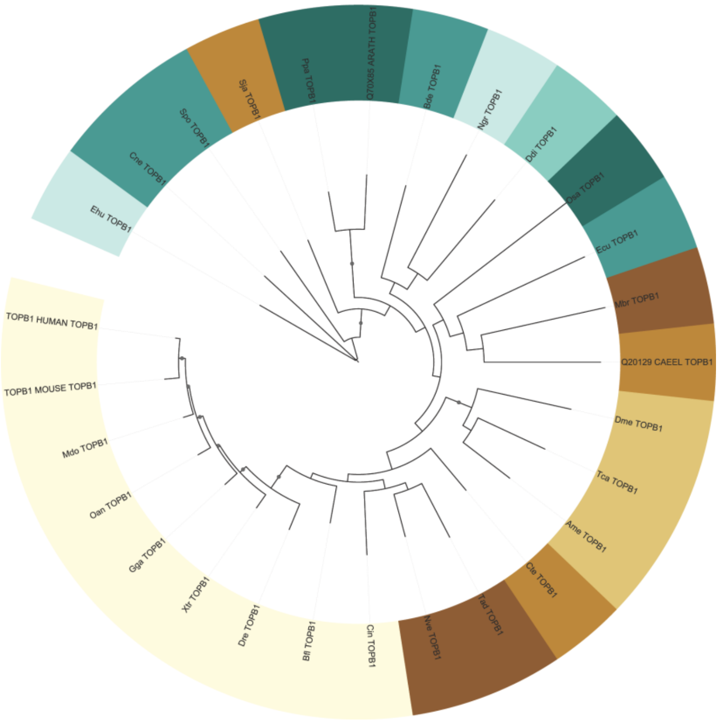


Figure STt25: phylogenetic tree of DNA topoisomerase 2-binding protein 1 (**Topb1**) (27). As in the previous tree, deep branches are poorly supported, and plant and fungi sequences are mixed. Besides, like in the case of RAD17, the *S. japonicum* sequence clades with ancient eukaryotes, and the distribution of basal animals to *chordata* is alike in both trees. That both trees show similar inconsistencies might show correlated evolution or common evolutionary pressure.

28-RAD1

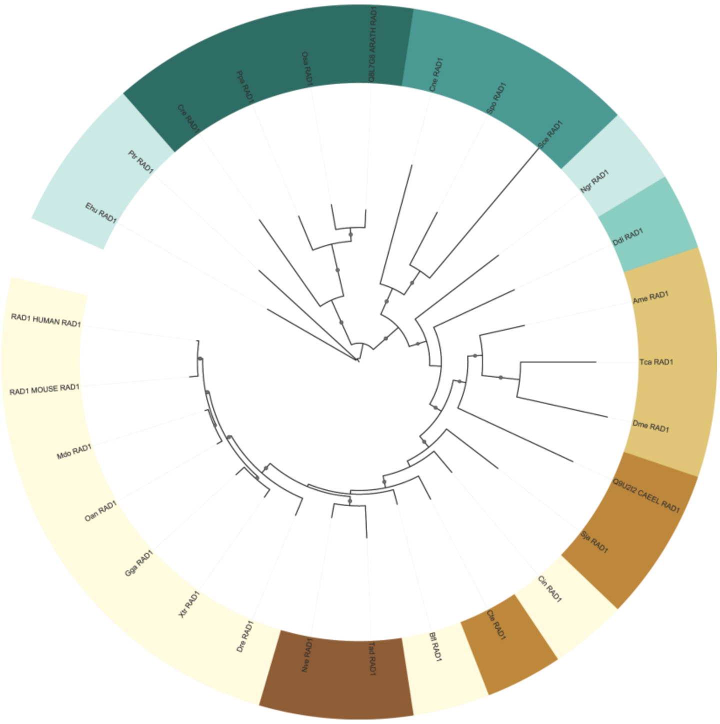


Figure STt26: phylogenetic tree of Cell cycle checkpoint protein **RAD1** (28). In this well supported tree, the arthropods clade prior to worms, and basal animals group with *chordata*. Also, the *N. gruberi* sequence clusters after fungi, and the worms *N. vectensis* and *T. adhaerens* cluster among *chordata.*

29-TIM

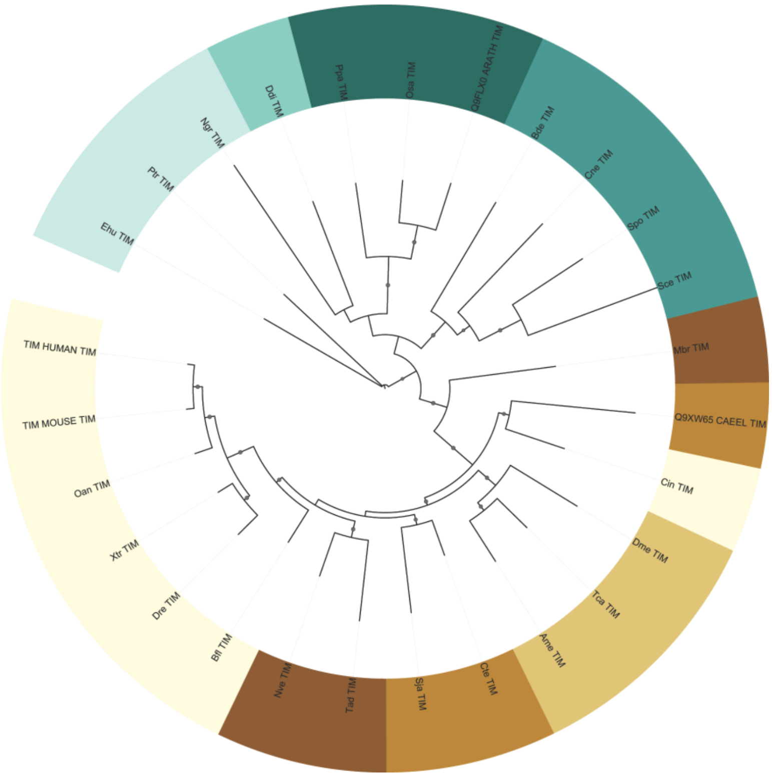


Figure STt27: phylogenetic tree of Protein timeless homolog (**TIM**) (29). The C. intestinalis sequence is placed before arthropoda, and as in previous cases, the worms *N. vectensis* and *T. adhaerens* cluster with *chordata.*

30-RAD9

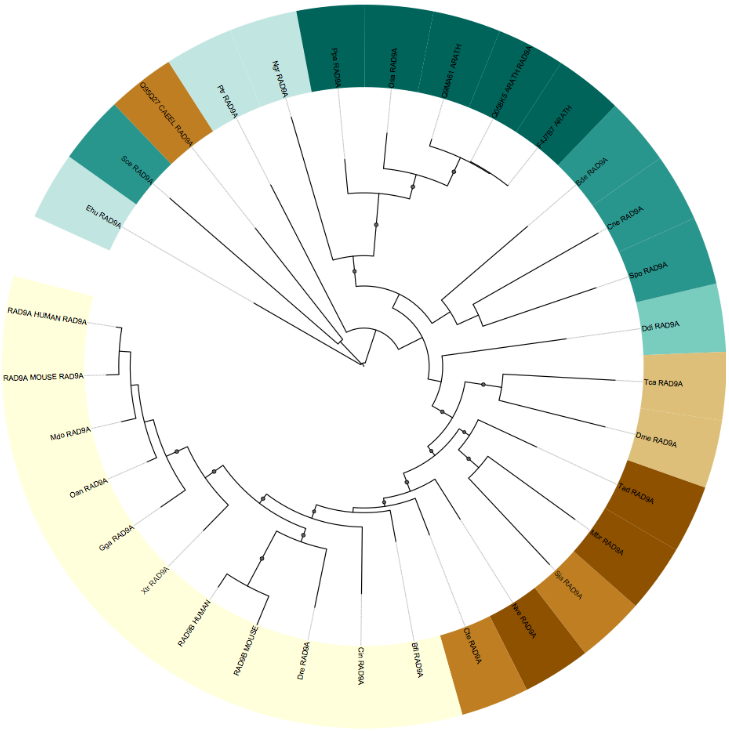


Figure STt28: phylogenetic tree of Cell cycle checkpoint control protein **RAD9A** (30). In this tree, the *C. elegans* and the *S. cerevisiae* sequences clade with ancient eukaryotes (though deep branches are not well supported), and arthropods are placed prior to basal eukaryotes.

31-XRCC5

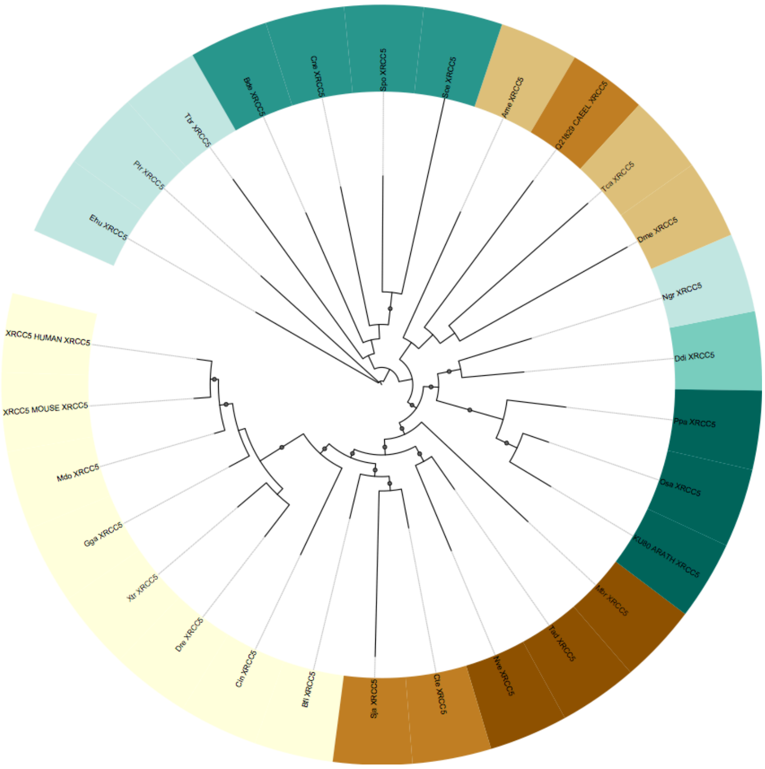


Figure STt29: phylogenetic tree of X-ray repair cross-complementing protein 5 (**XRCC5**) (31). In this tree, plants and fungi have inverted positions, and arthropods and *C. elegans* clade between them, though this invertebrates’ branch is not well supported.

32-XRCC6

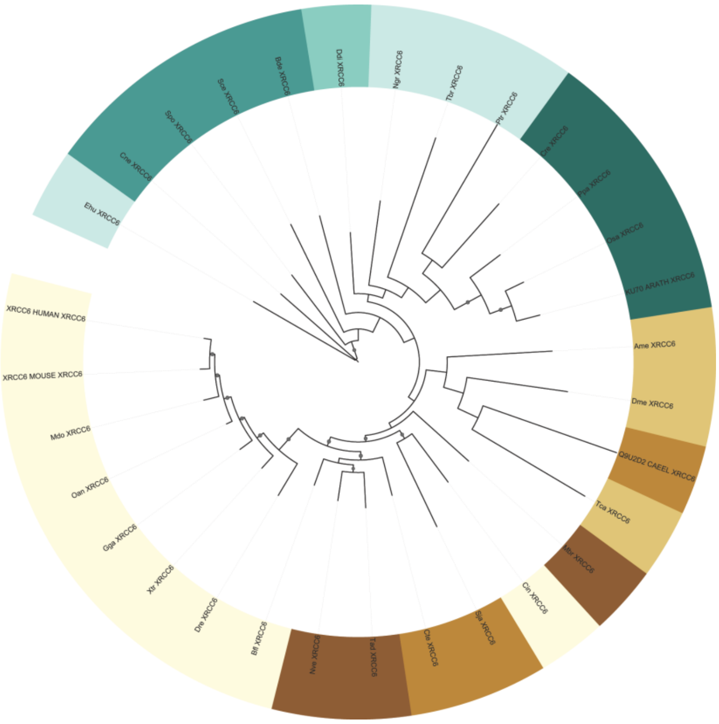


Figure STt30: phylogenetic tree of X-ray repair cross-complementing protein 6 (**XRCC6**) (32). As in the previous case, in this tree deep branches are not supported, and fungi and plants have their location inverted. Besides, arthropods and *C. elegans* clade together before basal eukaryotes. Both XRCC5 and XRCC6 are part of the same protein complex, and share some inconsistencies in their gene-trees, which might point towards correlated evolution or common evolutionary pressure.

33-DDB1

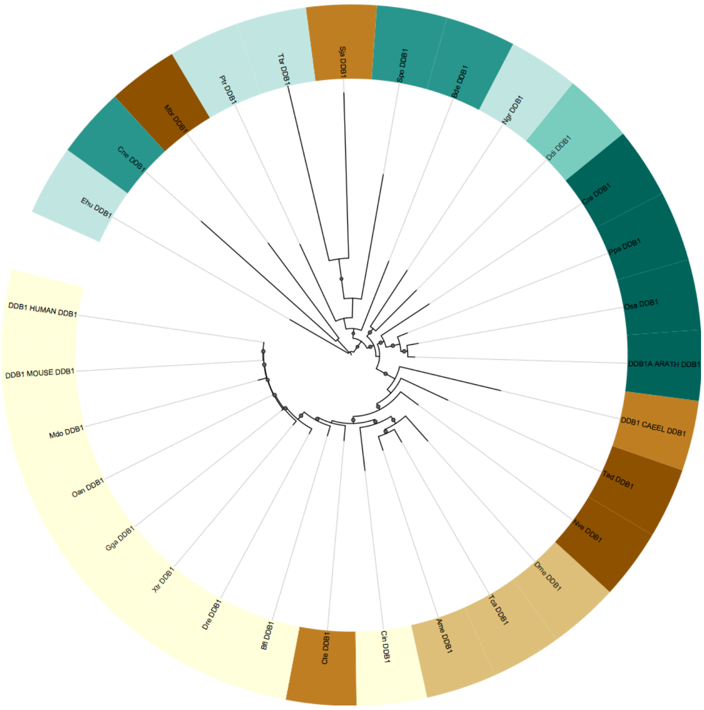


Figure STt31: phylogenetic tree of DNA damage-binding protein 1 (**DDB1**) (33). As in previous cases, The *M. brevicollis* and *S. japonicum* sequences clade with ancient eukaryotes, fungi and plants have their positions inverted, and *C. elegans* is positioned before basal *metazoa*. Also, *C. teleta* clusters with *chordata*.

34-RFA2

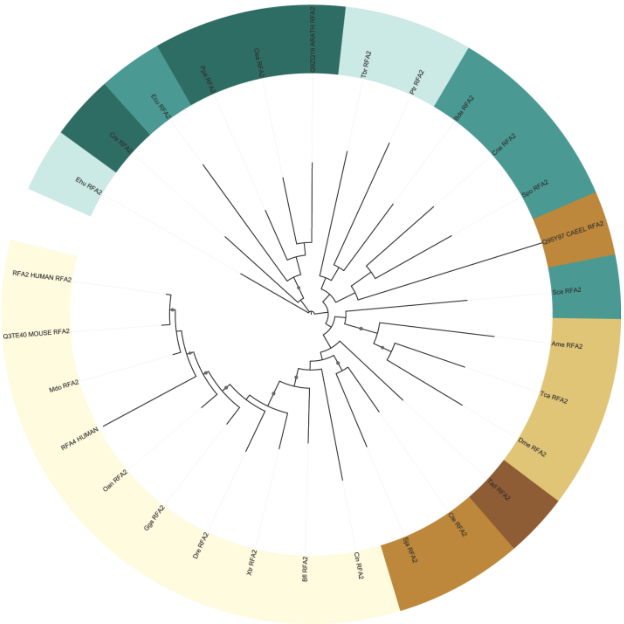


Figure STt32: phylogenetic tree of Replication protein A 32 kDa subunit (**RFA2**) (34). In this poorly supported tree, the *E. cuniculi* ortholog clades with plants, and all of them among ancient eukaryotes. Also, *C elegans* groups with fungi, and arthropods are located prior to *T. adhaerens*, the only basal animal in this tree.

35-Family 16: ATM/ATR/PKRCD

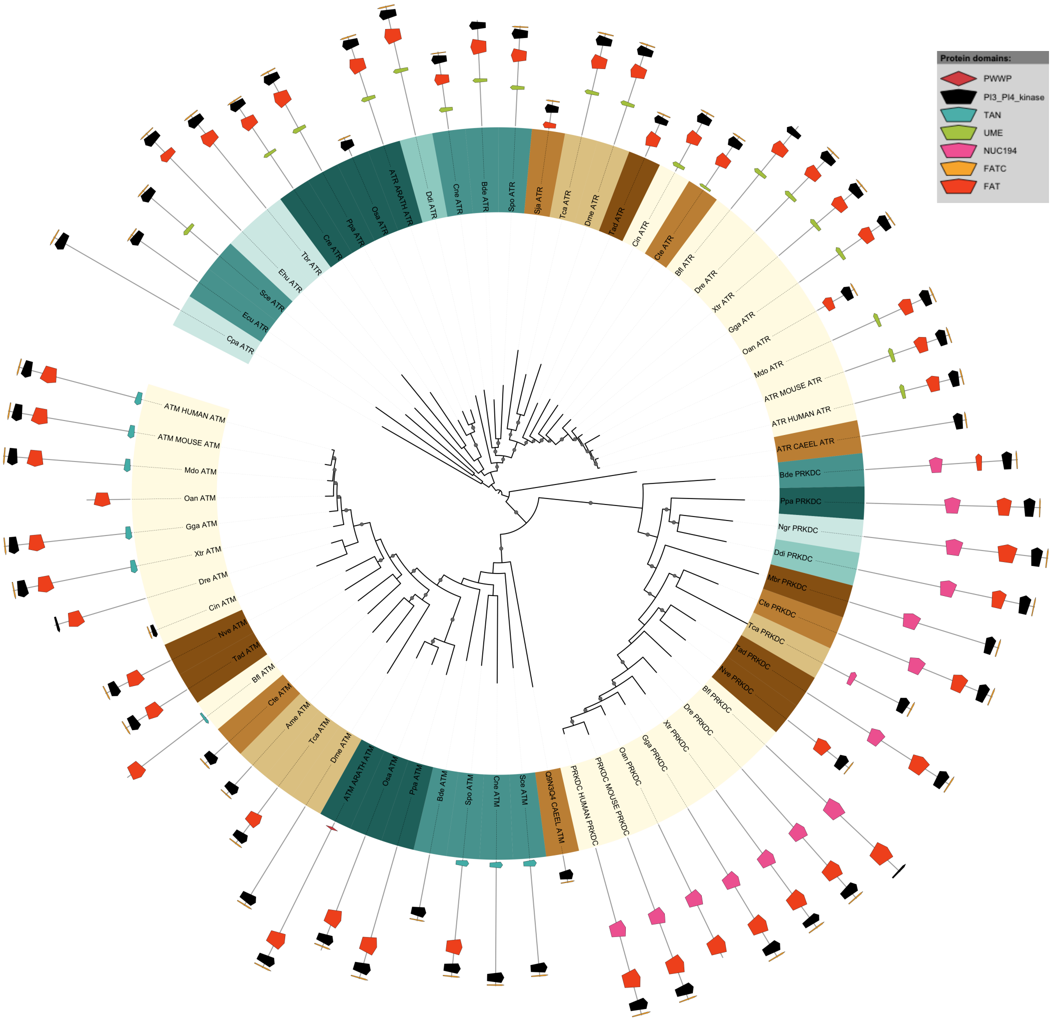


Figure STt33: phylogenetic tree of Family 16 (35), comprising PIK-related kinases **ATM**, **ATR** and **PKRCD**. ATR seems to be the most ancient sequence, in both PKRDC and ATM, the worms *N. vectensis* and *T.*

adhaerens cluster with *chordata;* and in the three proteins, fungi and plants have inverted positions. The *C. elegans* ATR clades at the base of the PRKDC and ATM sequences, which may be because the sequence is actually not ATR, or might also be that the sequence has not been correctly predicted.

36-TRIPC

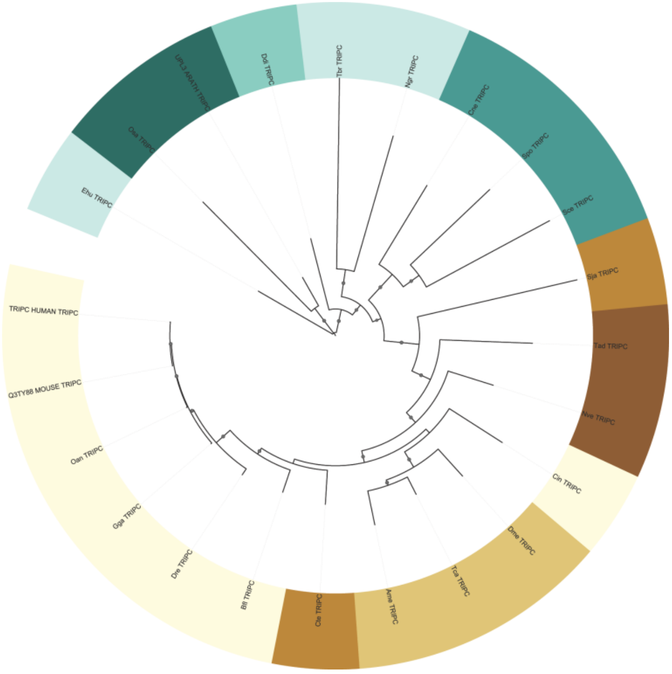


Figure STt34: phylogenetic tree of E3 ubiquitin-protein ligase TRIP12 (**TRIPC**) (36). In this well supported tree, the *S. japonicum* sequence clades prior to basal *metazoa*, and the *C. intestinalis* ortholog groups with arthropoda, while *C. teleta* clusters with *chordata*.

37-DNLI4

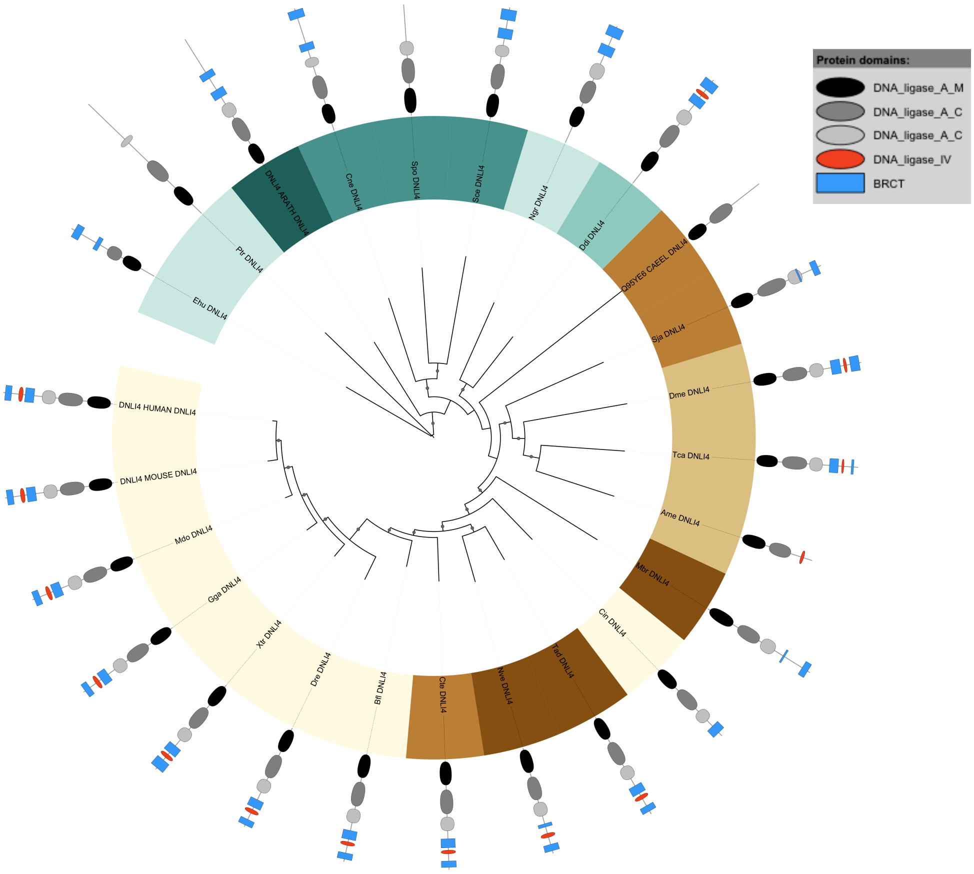


Figure STt35: phylogenetic tree of DNA ligase 4 (**DNLI4**) (37). In this tree, basal eukaryotes (among which clusters the chordate C. intestinalis) clade after worms (but *C. teleta*) and *arthropoda*.

38-HUS1

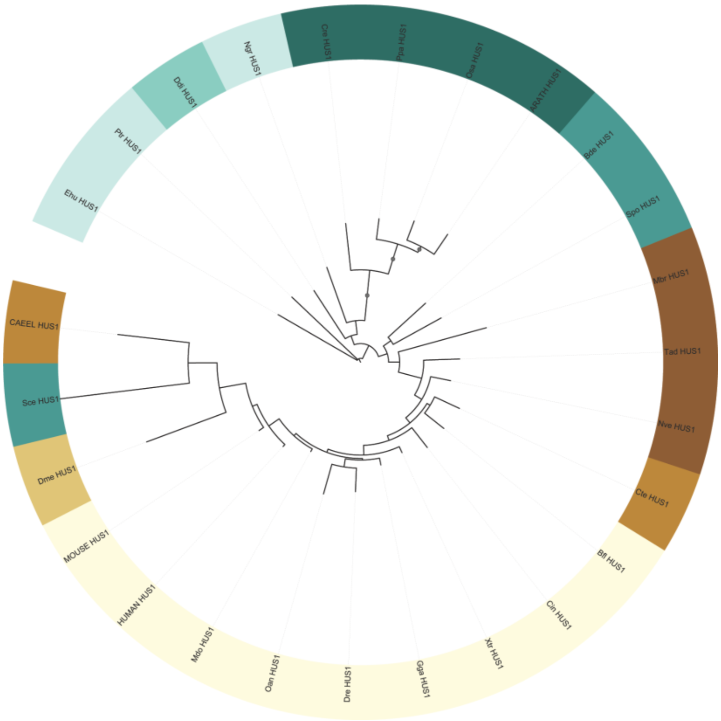


Figure STt36: phylogenetic tree of Checkpoint protein **HUS1** (38). This tree is not supported and shows strong inconsistencies, being the *C. elegans*, *S. cerevisiae* and *D. melanogaster* sequences after chordata. Thus, more sampling is needed.

39-ERCC6

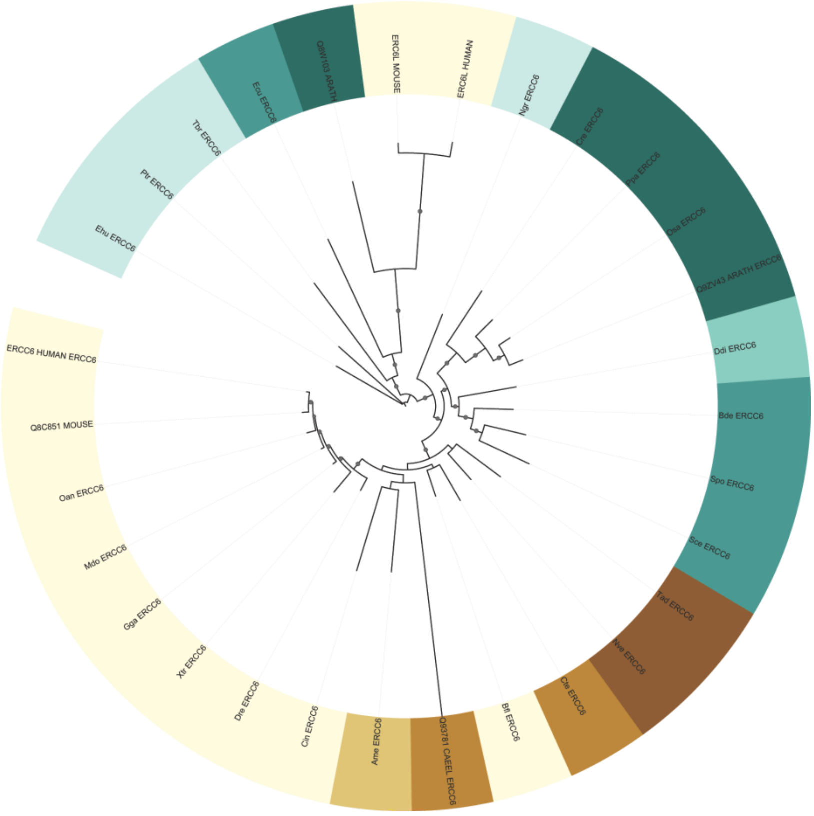


Figure STt37: phylogenetic tree of DNA excision repair protein **ERCC6** (39). In this tree, the sequences of three inparalogs (ERCC6L from mouse and human, and maybe the equivalent in *A. thaliana*) were included. Most proteins follow the phylogenetic tree order, with few exceptions like the *E. cuniculi* sequence clustering with ancient eukaryotes, or the *B. floridae* cladding with worms.

40-Family 9: Artemis-Apollo

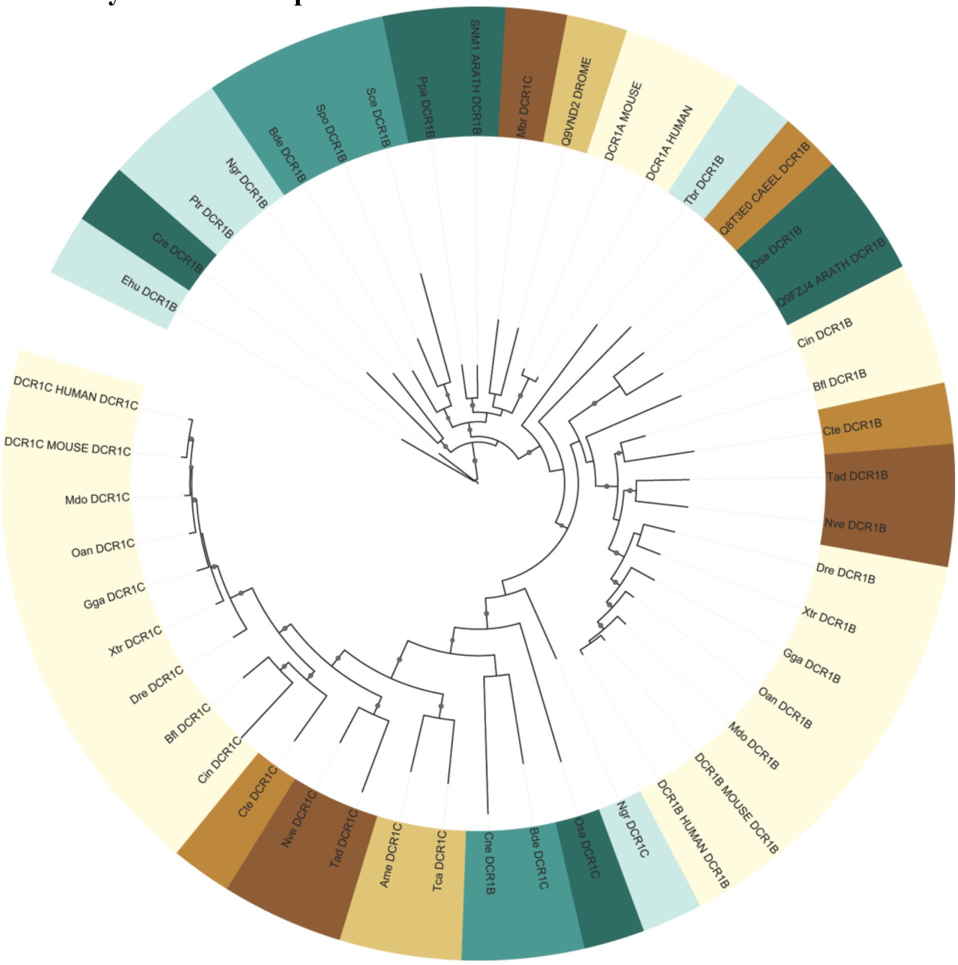


Figure STt38: phylogenetic tree of Family 9 (40), comprising the DNA cross-link repair 1C protein Artemis (**DCR1C**) and the 5' exonuclease Apollo (**DCR1B**). This family also includes DCR1A, which is involved in DDR too. The phylogeny establishes that the DCR1B in fungi, *M. brevicollis*, *D. melanogaster* and some plants (top of the tree) are probably in fact DCR1A, and have been erroneously detected as DCR1B by InParanoid. Also according to the tree, DCR1B was lost in fungi and invertebrates but *C. teleta* and *C. elegans*, while DCR1C is present in all lineages.

42-FACD2

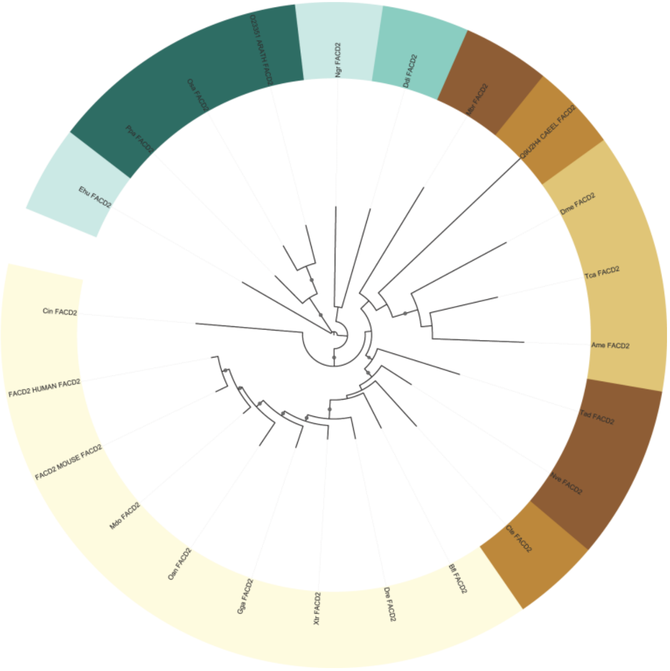


Figure STt39: phylogenetic tree of Fanconi anemia group D2 protein (**FACD2**) (42). As in previous trees, *C. elegans* groups with *M. brevicollis* and the basal *metazoa* cluster closer to *chordata* than arthropods.

43-Family 3: PARPs

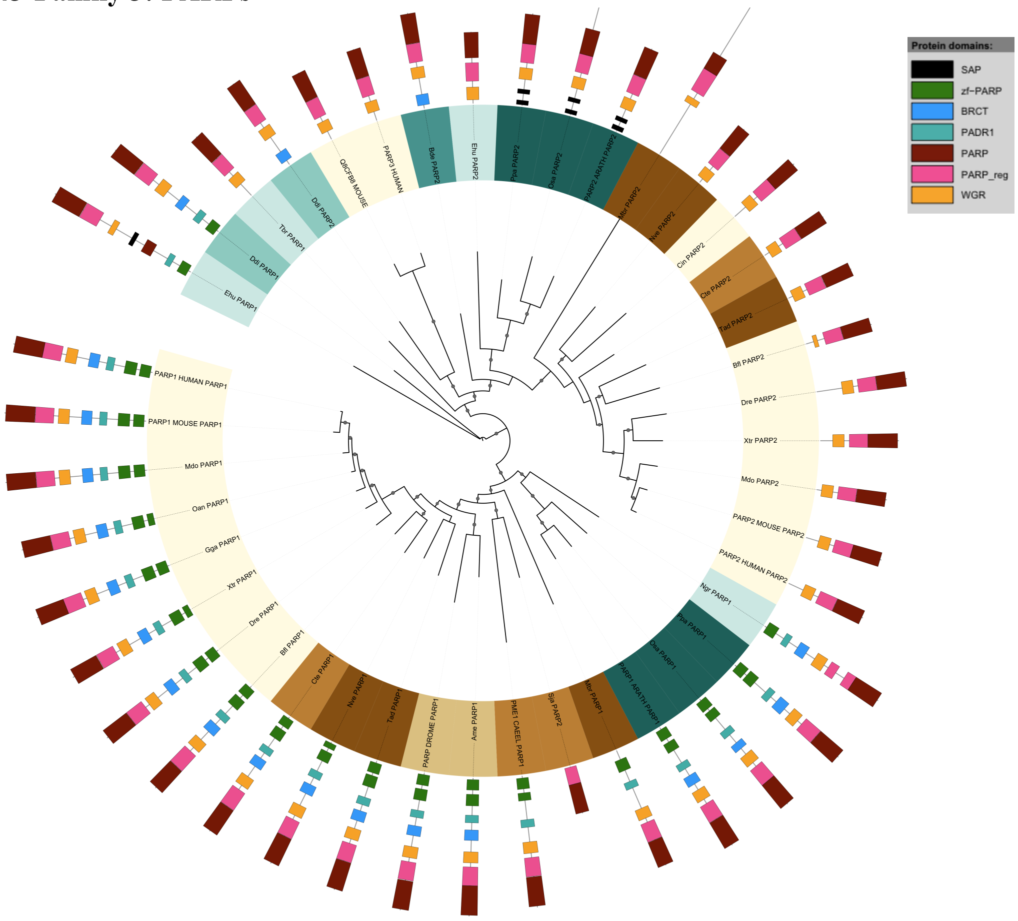


Figure STt40: phylogenetic tree of Family 3 (43), comprising Poly [ADP-ribose] polymerases **PARP1** and **PARP2**. This well supported tree indicates that PARPs duplication is an old event occurring in ancient eukaryotes. The tree also includes PARP3, other member of the PARPs family. According to the phylogeny, *S. japonicum* PARP2 is actually PARP1. As in other trees, PARP1 in basal *metazoa* cluster closer to *chordata* than arthropods.

44-SMAL1

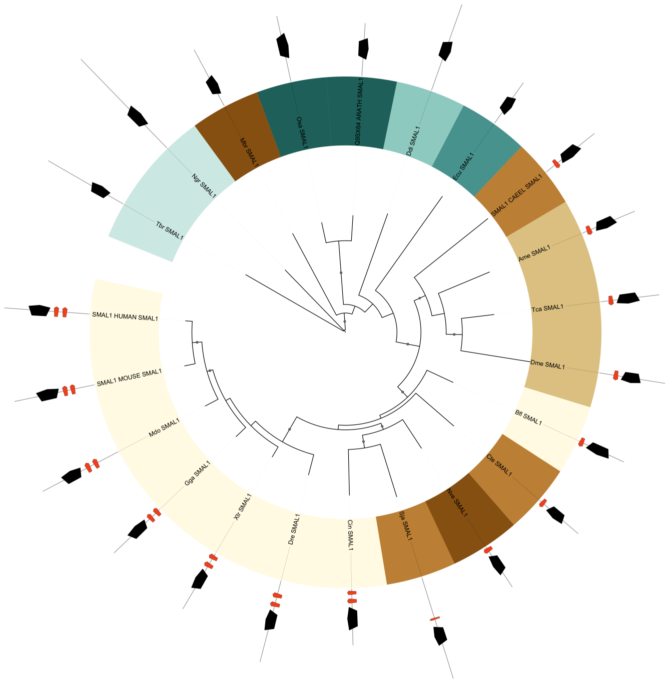


Figure STt41: phylogenetic tree of SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 (**SMARCAL1**) (44). *M. brevicollis*, as in other trees, clades with ancient eukaryotes (maybe due to the domain composition of the protein), and worms (but *C. elegans*) and arthropods have inverted their position. Also, the *B. floridae* sequence groups with invertebrates.

45-Family 14: Kinases

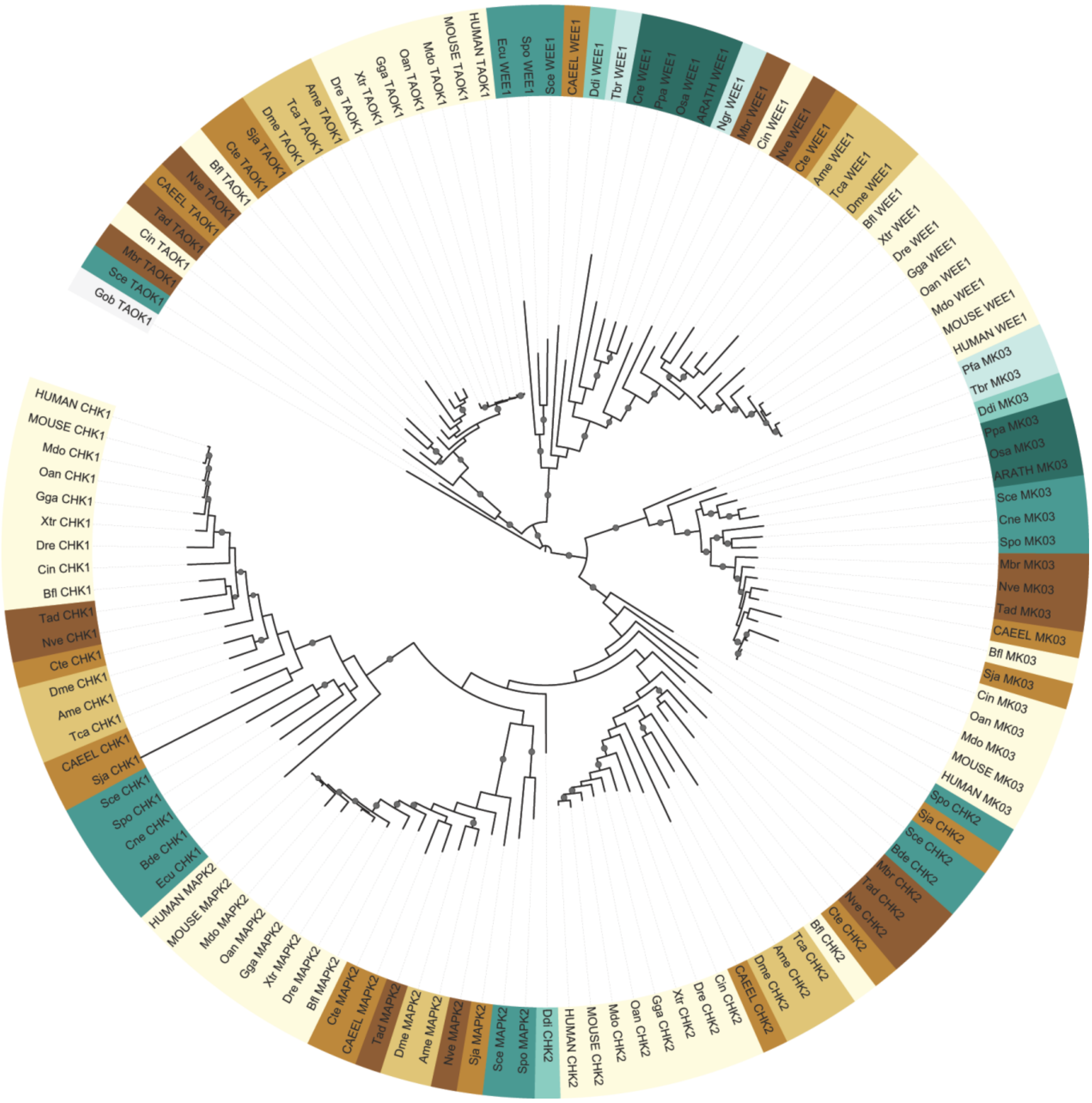


Figure STt42: phylogenetic tree of Family15 (45), comprising kinases **CHK1**, **CHK2**, **MK2**, **MK03**, **Tao** and **WEE1**.

This well supported tree shows the most recent kinases are CHK1 and MAPK2, while the most ancient is TAOK1, which was already present in some *Planctomycetes*. Regarding this kinase, the only fungus containing it in our study is *S. cerevisiae*, and it is also remarkable that the *C. intestinalis* sequence groups with basal eukaryotes. In the case of WEE1, fungi and plants have their positions inverted and the *C. elegans* ortholog clades between them. Also, ancient eukaryotes cluster with plants or prior to basal eukaryotes. As with TAOK1, the *C. intestinalis* WEE1 sequence groups with basal eukaryotes. Regarding MK03, the *S. japonicum* ortholog clades between worms. The sequences of this parasite seems to frequently clade away from the other worms in this study, since in CHK2 it clusters among fungi. Also in CHK2, the *B. floridae* sequence groups between worms and *arthropoda*, instead of with the other chordates. The *D. dictiostellum* CHK2 sequence groups at the base of the MAPK2 branch, suggesting the orthology inference by InParanoid might be erroneous. Again in MAPK2, the *S. japonicum* sequence clades before basal metazoa, which also happens in the case of CHK1, this time also with the *C. elegans* ortholog.

BLOCK II:

46-RAD18

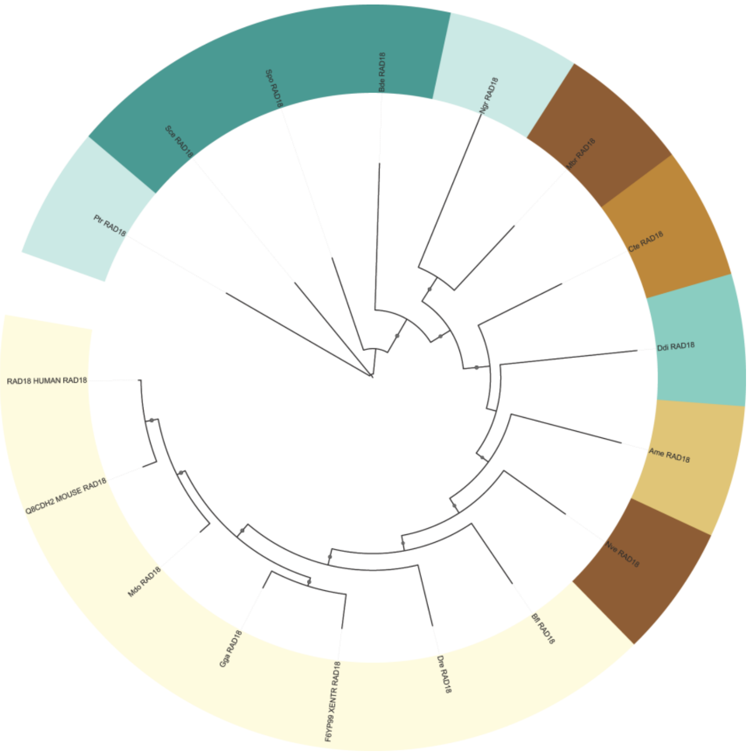


Figure STt43: phylogenetic tree of E3 ubiquitin-protein ligase **RAD18** (46). In this tree, fungi clade among ancient eukaryotes, and *N. vectensis* groups closer to *chordata* than to basal eukaryotes.

47-XRCC1

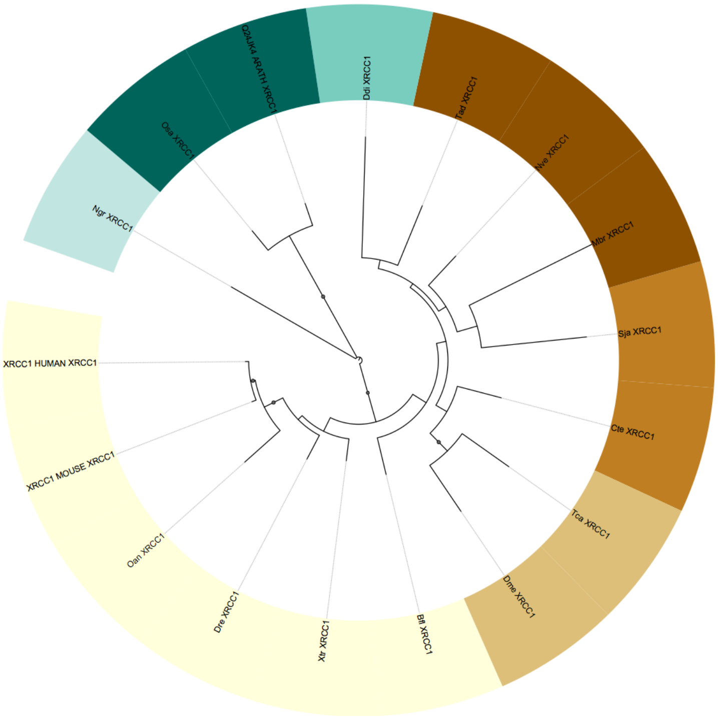


Figure STt44: phylogenetic tree of DNA repair protein **XRCC1** (47). This tree represents quite well the species phylogenetic order.

48-TDP2

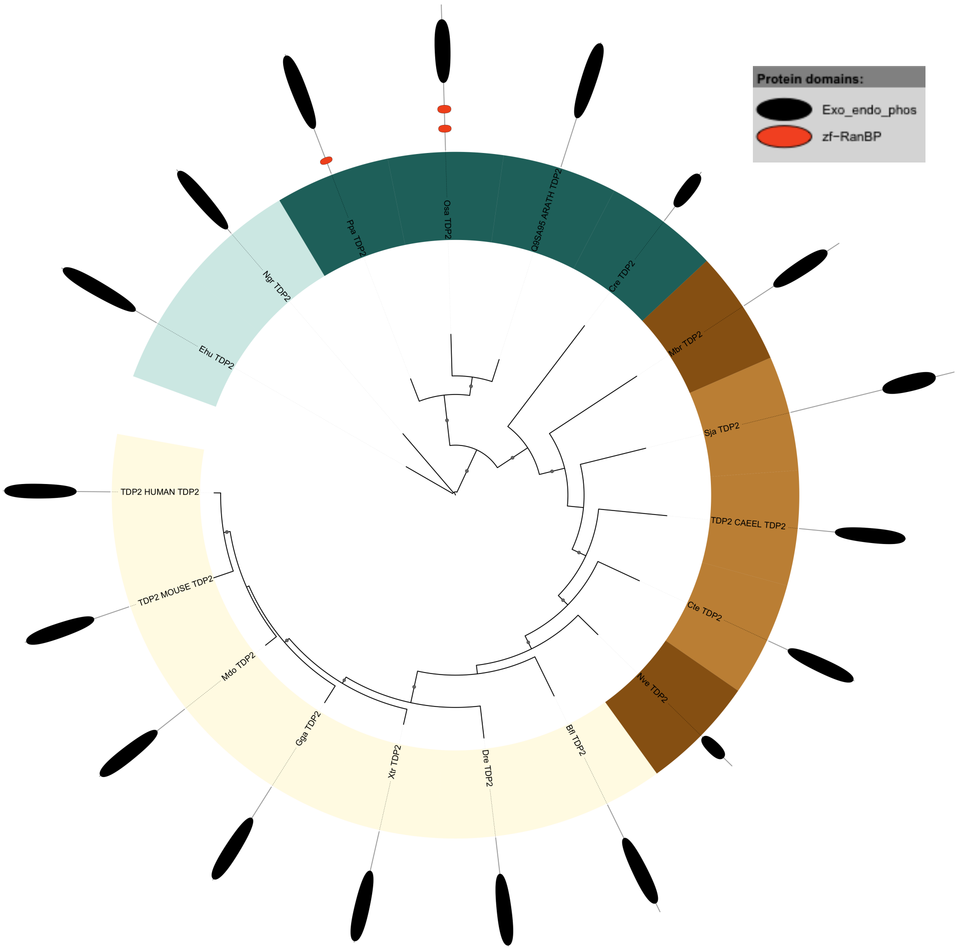


Figure STt45: phylogenetic tree of Tyrosyl-DNA phosphodiesterase 2 (**TDP2**) (48). This tree represents quite well the species phylogenetic order, but for the *N. vectensis* sequence, that as in previous trees, clades closer to *chordata* than to basal eukaryotes.

49-BRCA1

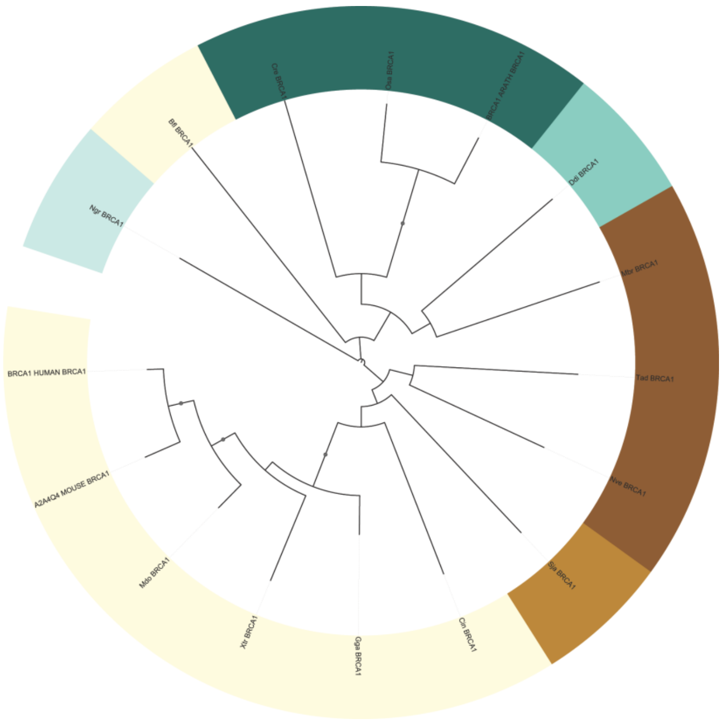


Figure STt46: phylogenetic tree of Breast cancer type 1 susceptibility protein (**BRCA1**) (49). This tree is poorly supported, and shows the *B. floridae* sequence cladding away from the other *chordata*.

51-ERCC8

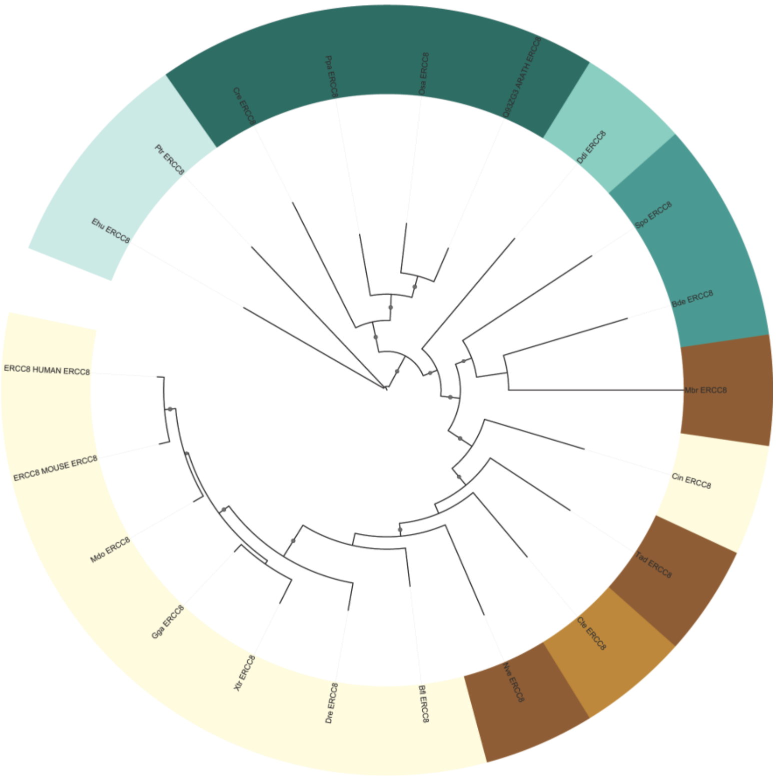


Figure STt47: phylogenetic tree of DNA excision repair protein **ERCC8** (51). But for the *C. intestinalis* sequence cladding with basal eukaryotes, this tree follows the species phylogenetic order.

52-NSE2 (MMS21)

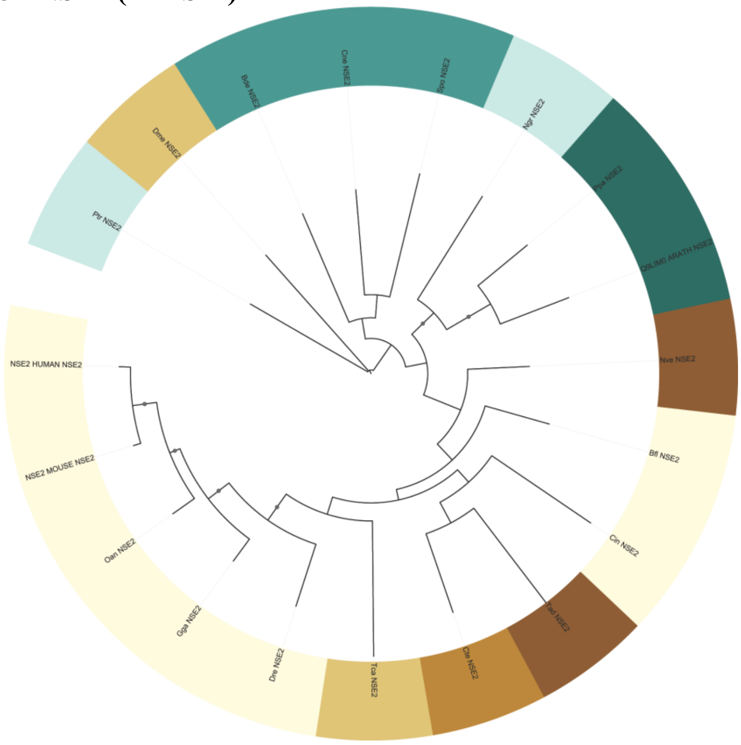


Figure STt48: phylogenetic tree of E3 SUMO-protein ligase **NSE2** (52). This poorly supported tree shows some inconsistencies, like the *D. melanogaster* sequence cladding at the base of the tree, or the *B. floridae* and *C intestinalis* orthologs clustering among basal eukaryotes.

53-BRCC3

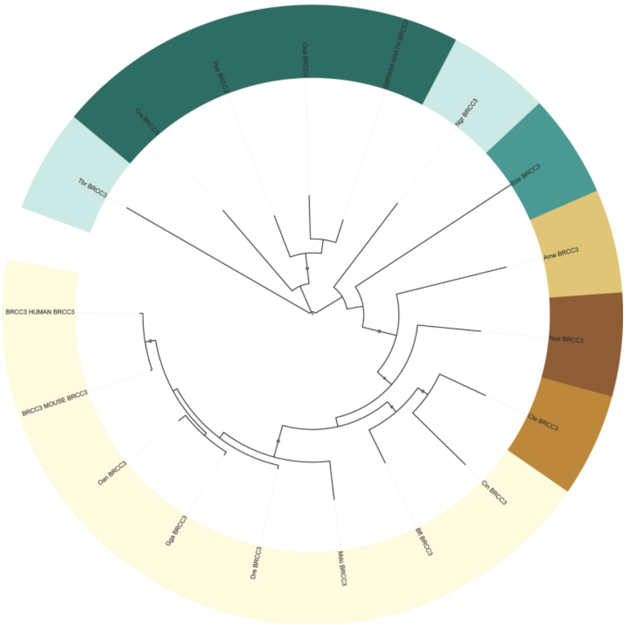


Figure STt49: phylogenetic tree of Lys-63-specific deubiquitinase **BRCC36** (53). This tree, follows the species phylogenetic order but for *A. mellifera* cladding prior to basal metazoa.

55-Family 1: PIAS

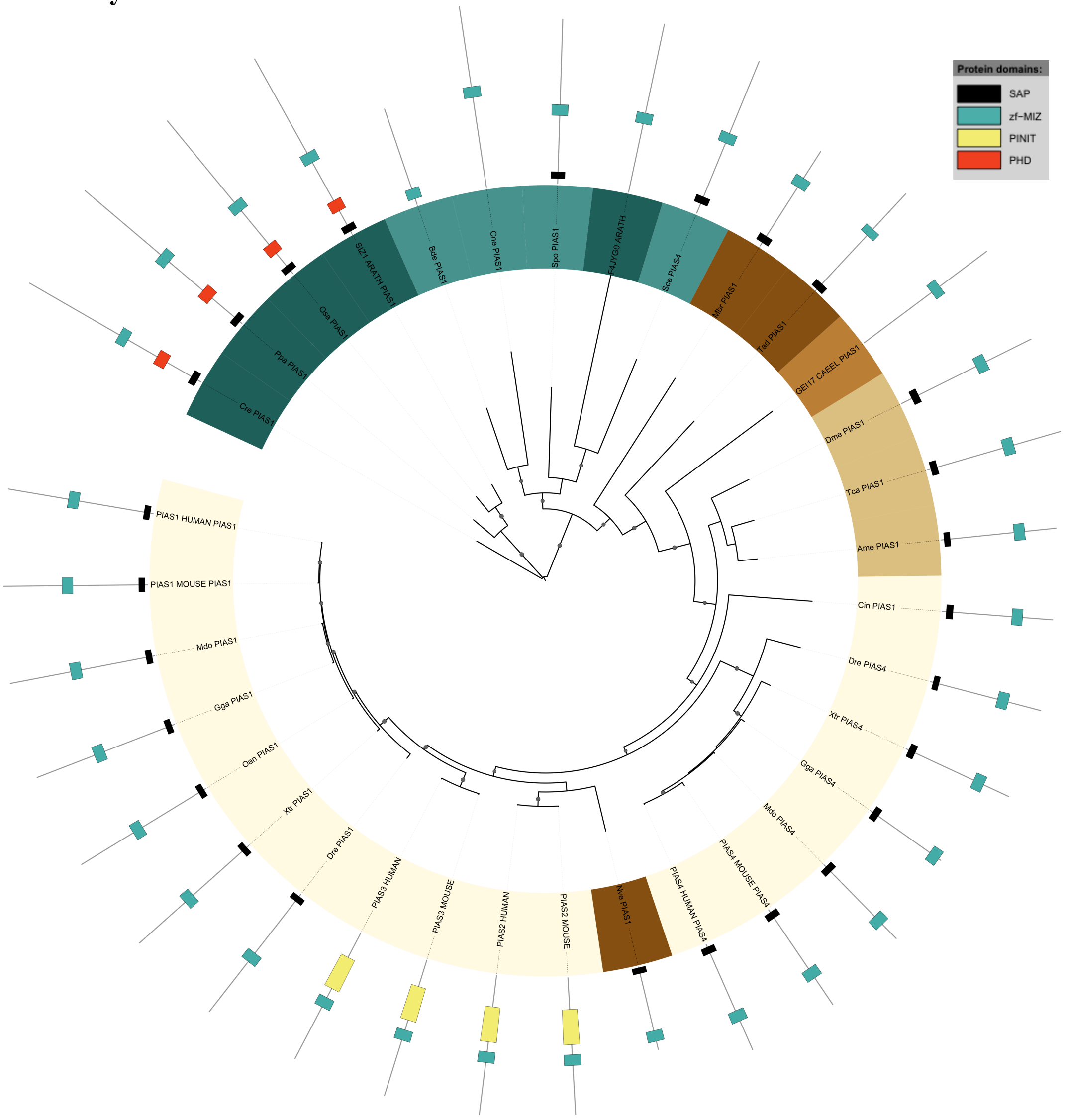


Figure STt50: phylogenetic tree of Family 1 (55), comprising E3 SUMO-protein ligases **PIAS1** to **PIAS4**. This well supported tree shows that *N. vectensis* PIAS1 might actually be PIAS2 (though it does not contain the PINIT domain), and that *S. cerevisiae* PIAS4 might be in fact PIAS1, being both incorrect predictions by InParanoid.

56-DTL

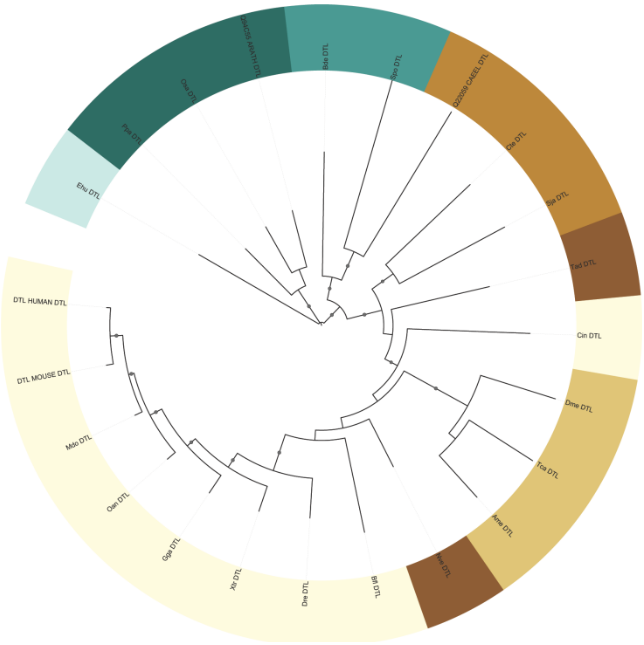


Figure STt51: phylogenetic tree of Denticleless protein homolog (**DTL**) (56). As in other trees, the *N. vectensis* ortholog clusters with chordata and the *C. intestinalis* sequence groups away from these.

58-PAXI1 – MDC1

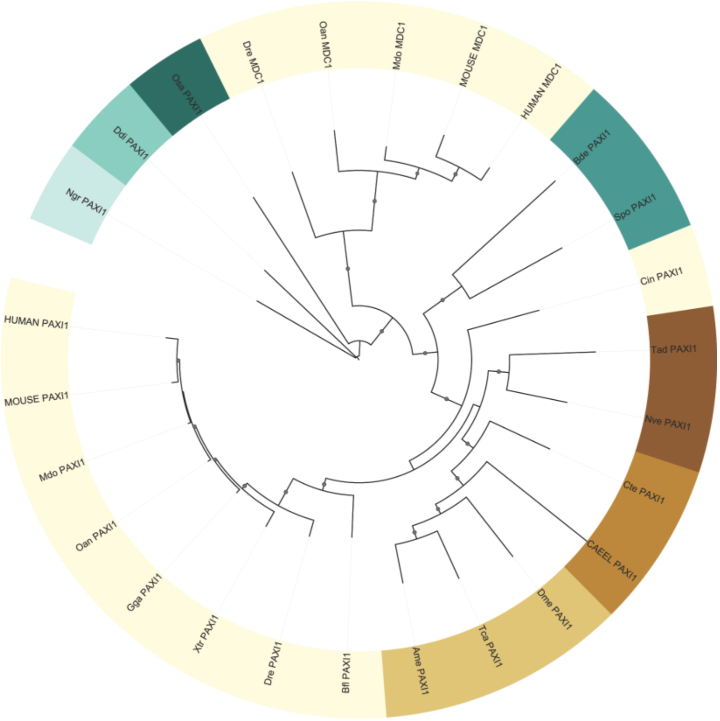


Figure STt52: phylogenetic tree of PAX-interacting protein 1 (PAXI1/PTIP) and Mediator of DNA damage checkpoint protein 1 (MDC1). In this tree, the PAXI1 orthologs follow the taxonomic order except for *C. intestinalis*, which is close to basal eukaryotes. On the other hand, the MDC1 sequences group among the plant and fungal PAXI1 orthologs. PAXI1 and MDC1 were initially grouped as a family because Ensemble COMPARA considered them as related, but given the MSA obtained with these sequences, the differences in their domain composition (they only share the promiscuous BRCT domain) and the clustering in the phylogenetic tree, we finally decided not to consider PAXI1 and MDC1 as members of the same family.

65-PLK1

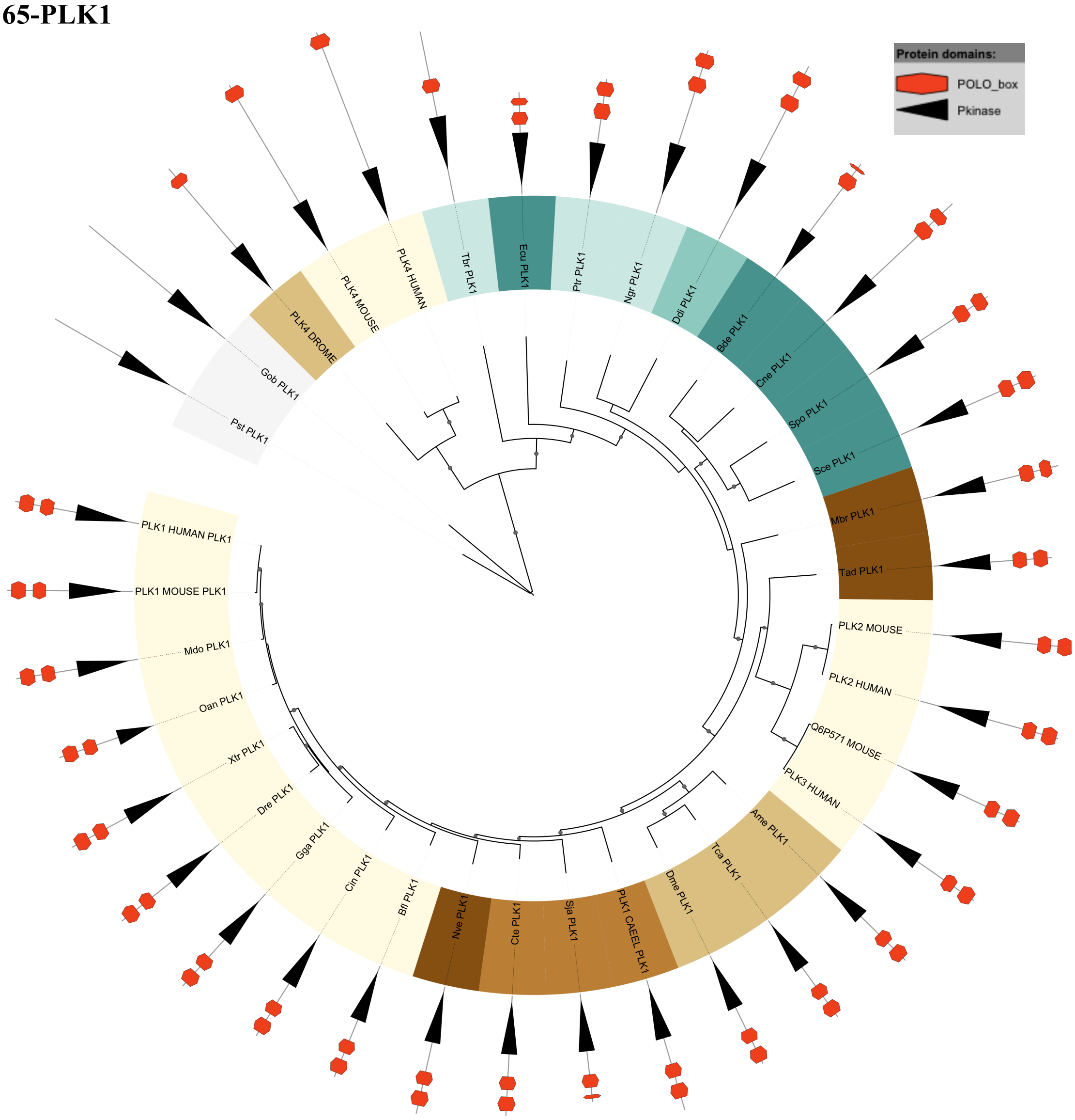


Figure STt53: phylogenetic tree of Serine/threonine-protein kinase **PLK1** (65). In this well supported tree, other sequences from the Polo kinase family have been included. According to the phylogeny, PLK1 seem to be the ancient sequence, and then generated PLK4, and finally PLK2 and 3. The *N. vectensis* sequence clades close to *chordata* instead of with basal eukaryotes, and worms and arthropoda have inverted positions compared to the species-tree.

66-SLX1

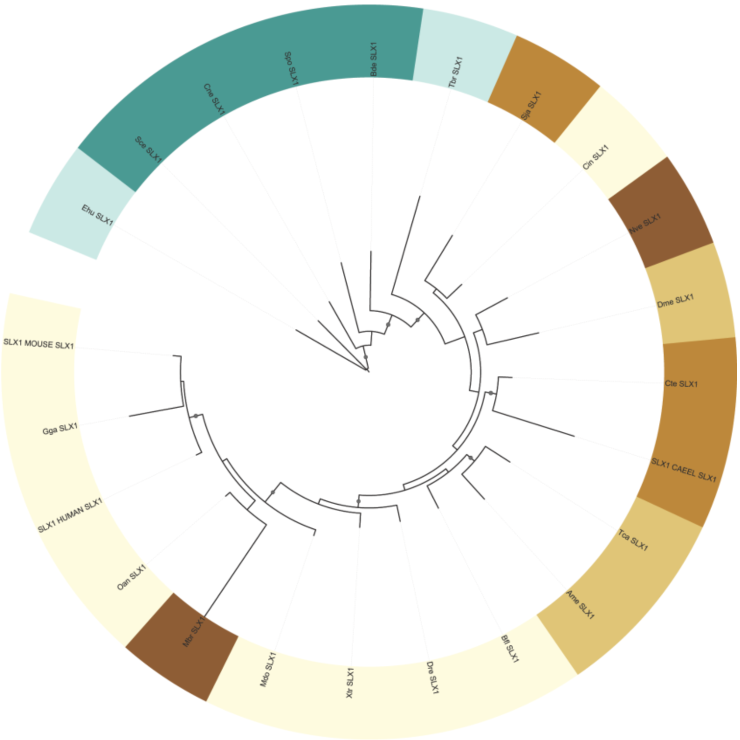


Figure STt54: phylogenetic tree of Structure-specific endonuclease subunit **SLX1** (66). This tree is poorly supported and species from different phylogenetic groups are mixed. M. *brevicollis* clades among *chordata*, and *C. intestinalis* groups with *S. japonicum* and close to *N. vectensis*. More sampling is requiered for this tree.

69-XPA

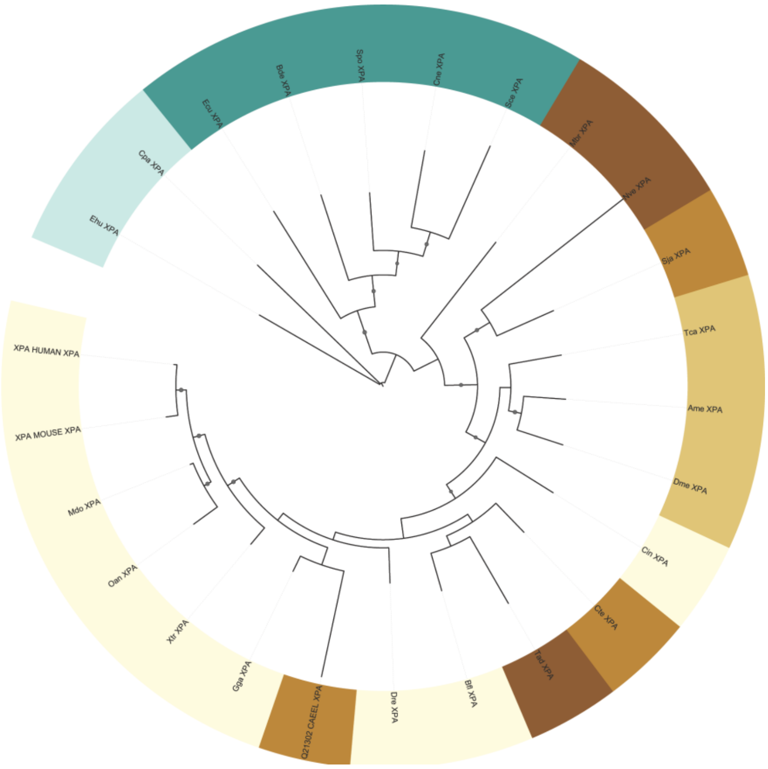


Figure STt55: phylogenetic tree of DNA repair protein complementing **XPA** cells (69). The *C. elegans* sequence clades among *chordata*, while *C. intestinalis* is placed before *C. teleta* and *T. adhaerens*, which does not cluster with the other basal eukaryotes.

BLOCK III:

76-BRCA2

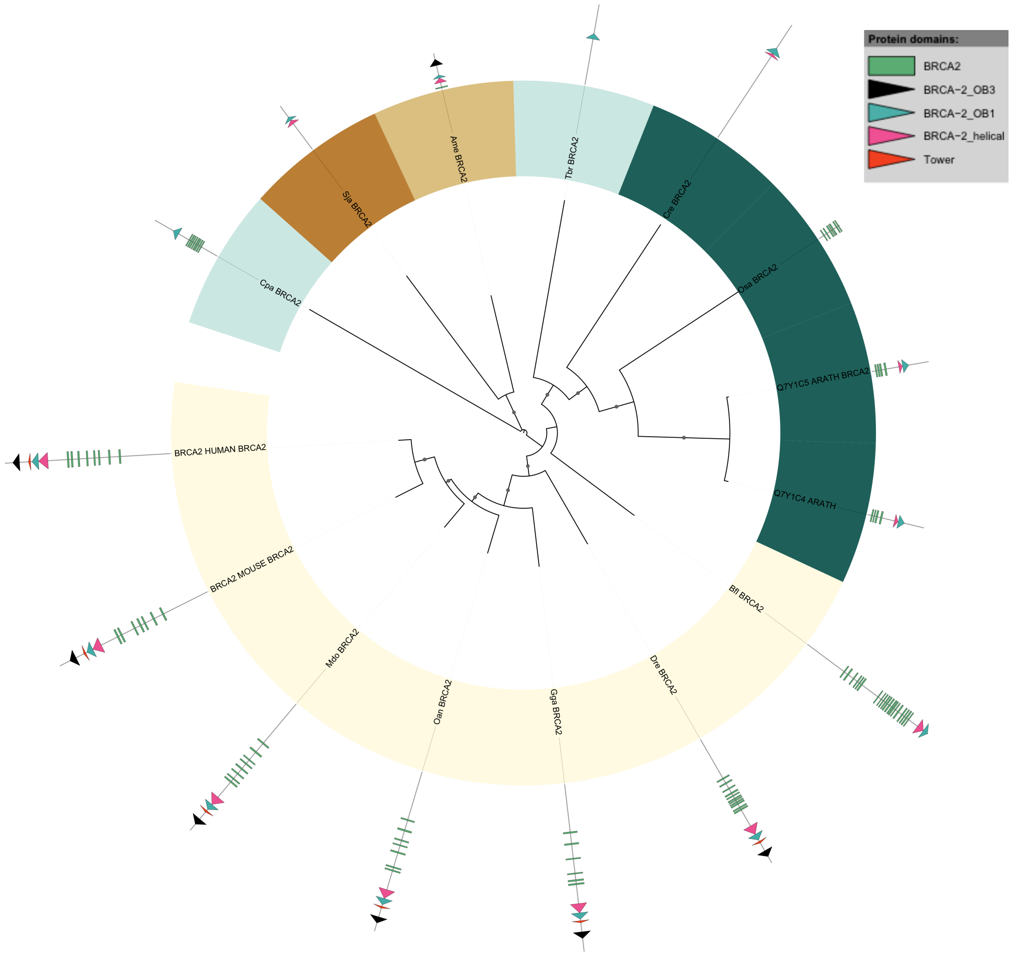


Figure STt56: phylogenetic tree of Breast cancer type 2 susceptibility protein (**BRCA2**) (76). *S. japonicum* and *A. mellifera* clade at the base of the tree instead of closer to *chordata*, which may be due to high divergence in these sequences or to an incorrect prediction of these protein sequences.

77-CDT1

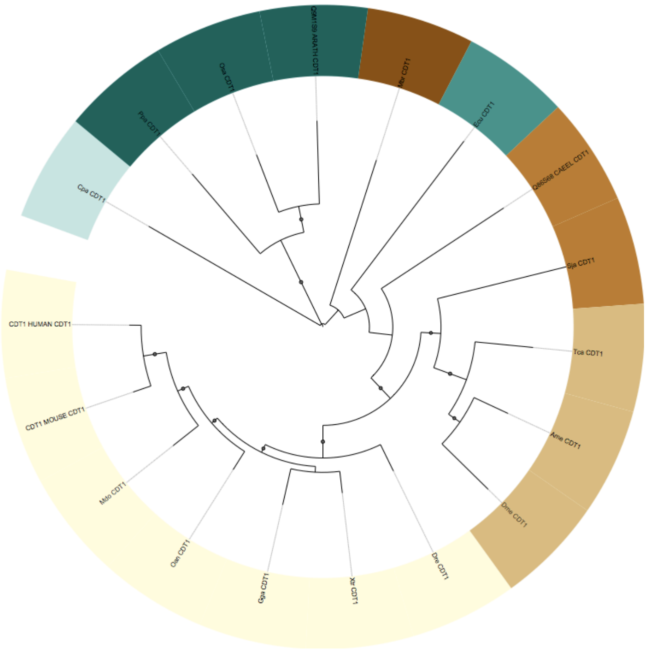


Figure STt57: phylogenetic tree of DNA replication factor **Cdt1** (77). This tree follows the species-tree order but for *E. cuniculi*, that clades after *M. brevicollis*, a basal eukaryote.

78-Family 4: MPIPs

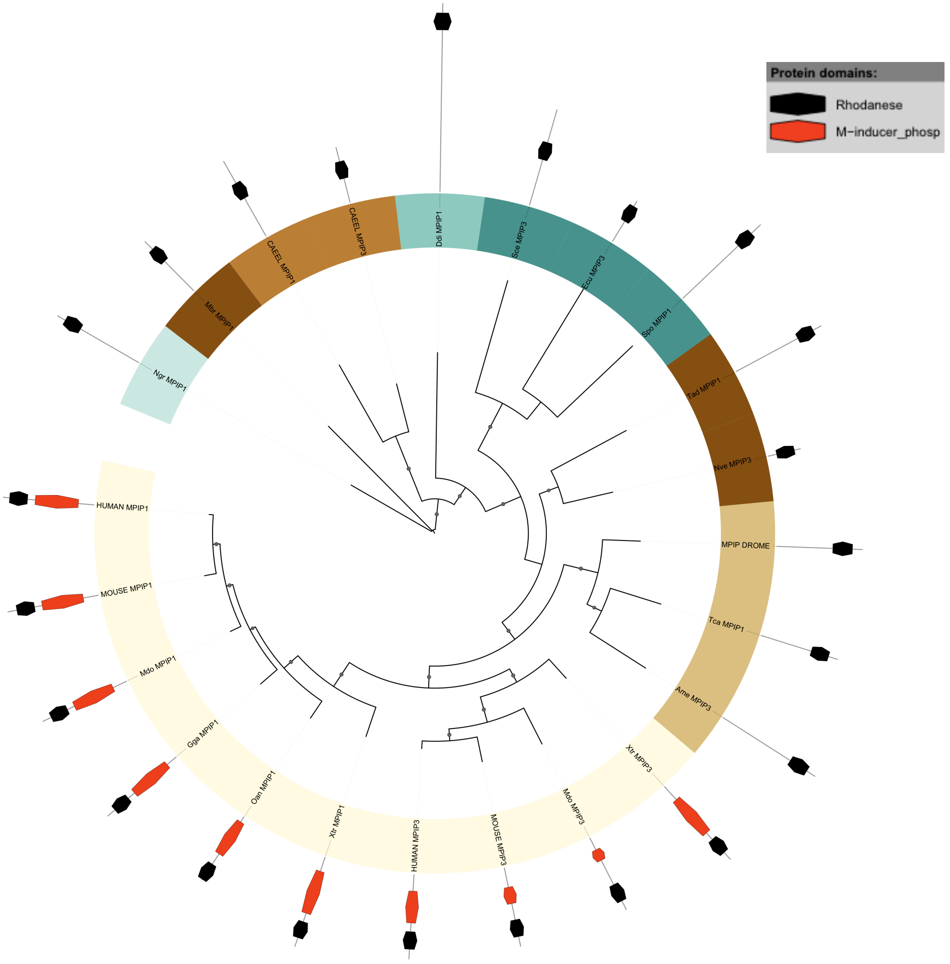


Figure STt58: phylogenetic tree of Family 4 (78), comprising M-phase inducer phosphatases 1 and 3 (**MPIP1 /CDC25A** and **MPIP3/ CDC25C**). As *C. elegans* seems to have both MPIP1 and 3, it is unclear if duplications were transmitted ancestrally or not, since basal eukaryotes have only one MPIP (*N. vectensis* MPIP3 will probably be MPIP1 instead). Fungi seem to have MPIP3 (unless this is an incorrect detection by InParanoid). Either both basal eukaryotes and arthropods lost one MPIP (as *C. elegans* has two), or *C. elegans* acquired one copy by HGT.

82-RNF8

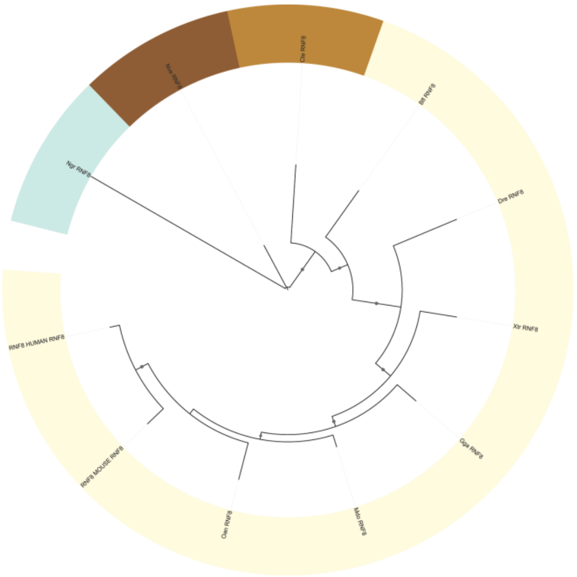


Figure STt59: phylogenetic tree of E3 ubiquitin-protein ligase **RNF8** (82). This tree represents quite well the species phylogenetic order.

ST1a. *E. coli* DDR proteins extracted from bibliography

DDR protein	UniProt ID	Entrez_ID	Annotation	References (PMID)	
CHO	P76213	16129695	Excinuclease cho	11818552	
CLPX	P0A6H1	16128423	ATP-dependent Clp protease ATP-binding subunit ClpX	16630889	21576225
DINF	P28303	162135919	DNA-damage-inducible protein F	22523558	
DINI	P0ABR1	16129024	DNA-damage-inducible protein I	15954802	
DNAA	P03004	16131570	Chromosomal replication initiator protein DnaA	1779750	16132081
DNLJ	P15042	16130337	DNA ligase	17938628	
DPO1	P00582	16131704	DNA polymerase I	11352575	21622737
DPO2	P21189	16128054	DNA polymerase II	16000023	
DPO3A	P10443	16128177	DNA polymerase III subunit alpha	12215643	
DPO3B	P0A988	16131569	DNA polymerase III subunit beta	16132081	
DPO3E	P03007	16128202	DNA polymerase III subunit epsilon	1575709	21576225
DPO4	Q47155	16128217	DNA polymerase IV	16000023	
EX5A	P04993	16130723	Exodeoxyribonuclease V alpha chain	19052323	
EX5B	P08394	16130724	Exodeoxyribonuclease V beta chain	19052323	
EX5C	P07648	16130726	Exodeoxyribonuclease V gamma chain	19052323	
FPG	P05523	16131506	Formamidopyrimidine-DNA glycosylase	1689309	14607836
HELD	P15038	16128929	Helicase IV	19451222	
HOLE	P0ABS8	16129795	DNA polymerase III subunit theta	8505306	
LEXA	P0A7C2	16131869	LexA repressor	21576225	
LON	P0A9M0	16128424	Lon protease	8995294	
MUTH	P06722	16130735	DNA mismatch repair protein mutH	7859291	
MUTL	P23367	16131992	DNA mismatch repair protein mutL	7859291	16132081
MUTS	P23909	16130640	DNA mismatch repair protein MutS	7859291	16132081
RECA	P0A7G6	16130606	Protein RecA	19052323	
RECF	P0A7H0	16131568	DNA replication and repair protein RecF	19052323	
RECG	P24230	16131523	ATP-dependent DNA helicase recG	19052323	
RECJ	P21893	16130794	Single-stranded-DNA-specific exonuclease recJ	19052323	
RECN	P05824	49176247	DNA repair protein recN	19451222	
RECO	P0A7H3	16130490	DNA repair protein RecO	19451222	
RECQ	P15043	162135918	ATP-dependent DNA helicase recQ	19052323	
RECR	P0A7H6	16128456	Recombination protein RecR	19451222	
RECX	P33596	16130605	Regulatory protein RecX	16000023	
RUVA	P0A809	16129814	Holliday junction ATP-dependent DNA helicase RuvA	9442895	
RUVB	P0A812	16129813	Holliday junction ATP-dependent DNA helicase RuvB	9442895	
RUVC	P0A814	16129816	Crossover junction endodeoxyribonuclease RuvC	9442895	
SBCC	P13458	16128382	Nuclease sbcCD subunit C	9927737	
SBCD	P0AG76	16128383	Nuclease sbcCD subunit D	9927737	
SSB	P0AGE0	16131885	Single-stranded DNA-binding protein	19052323	
SULA	P0AFZ5	16128925	Cell division inhibitor SulA	16630889	
UMUC	P04152	16129147	Protein umuC	16000023	
UMUD	P0AG11	16129146	Protein umuD	16000023	
UVRA	P0A698	16131884	UvrABC system protein A	16000023	
UVRB	P0A8F8	16128747	UvrABC system protein B	16000023	
UVRC	P0A8G0	90111354	UvrABC system protein C	16000023	
UVRD	P03018	16131665	DNA helicase II	16000023	
YOAA	P76257	16129762	Probable ATP-dependent helicase yoaA	InParanoid: hERCC2 homolog	

ST1b. *A. thaliana* DDR proteins extracted from bibliography

DDR protein	UniProt ID	Annotation	References (PMID)
1433E	Q01525	14-3-3-like protein GF14 omega	InParanoid: h1433E ortholog
3MG	Q39147	DNA-3-methyladenine glycosylase	20646326
ALKBH	Q9SA98	Alpha-ketoglutarate-dependent dioxygenase alkB	20646326
APE1L	Q9STM2	Putative uncharacterized protein T29H11_60	20646326
APE2	Q8W4I0	Endonuclease 2	20646326
ATM	Q9M3G7	Serine/threonine-protein kinase ATM	20646326
ATR	Q9FKS4	Serine/threonine-protein kinase ATR	20646326
ATRIP	C8KI33	ATR interacting protein	19619159
BARD1	Q3E7F4	Putative uncharacterized protein At1g04020.2	20646326
BH140	Q9M041	Transcription factor bHLH140	20646326
BLM (RecQl4)	Q8L840	ATP-dependent DNA helicase Q-like 4A	20646326
BRCA1	Q8RXD4	Protein BREAST CANCER SUSCEPTIBILITY 1 homolog	20646326
BRCA2	Q7Y1C4	Breast cancer 2 susceptibility protein	20646326
BRCC3	Q8RW94	At1g80210/F18B13_28	InParanoid: hBRCC3 ortholog
BRE	Q5XF81	At5g42470	InParanoid: hBRE ortholog
CCH11	Q8W5S1	Cyclin-H1-1	20646326
CDKD1	Q9C9U2	Cyclin-dependent kinase D-1	20646326
CDPKL	Q9ZSA2	Calcium-dependent protein kinase 21	20646326
CDT1	Q9M1S9	CDT1-like protein b	15928083
CIPK3	Q2V452	CBL-interacting serine/threonine-protein kinase 3	20646326
CRY1	Q43125	Cryptochrome-1	20646326
CRY2	Q96524	Cryptochrome-2	20646326
CRYD	Q84KJ5	Cryptochrome DASH, chloroplastic/mitochondrial	20646326
CUL4	Q8LGH4	Cullin-4	20646326
DDB1A	Q9M0V3	DNA damage-binding protein 1a	20646326
DDB2	Q6NQ88	Protein DAMAGED DNA-BINDING 2	20646326
DET1	P48732	Light-mediated development protein DET1	20646326
DMC1	Q39009	Meiotic recombination protein DMC1 homolog	20646326
DNLI1	Q42572	DNA ligase 1	20646326
DNLI4	Q9LL84	DNA ligase 4	20646326
DPOLA	Q9FHA3	DNA polymerase alpha catalytic subunit	21867786
DR100	Q00874	DNA-damage-repair/toleration protein DRT100	21867786
DTL	Q94C55	Protein denticleless	InParanoid: hDTL ortholog
EME1	Q9SJ19	Putative uncharacterized protein At2g21800	20646326
ERCC1	Q9MA98	DNA excision repair protein ERCC-1	20646326
ERCC2	Q8W4M7	DNA repair helicase UVH6	20646326
ERCC6 (CSB)	Q9ZV43	DNA excision repair protein E	20646326
ERCC8 (CSA)	Q93ZG3	At1g27840/F28L5_15	20646326
EXO1	Q8L6Z7	Exonuclease 1	20646326
FACD2	O23351	Putative uncharacterized protein	InParanoid: hFACD2 ortholog
FPG	O80358	At1g52500	20646326
GR1	Q9ZRT1	Protein gamma response 1	21867786
GTF2H1	Q3ECP0	Probable RNA polymerase II transcription factor B subunit 1-1	20646326
GTF2H2	Q9ZVN9	P44/SSL1-like protein	20646326
H2AX	Q8GUH3	Histone H2A	InParanoid: hH2AX ortholog
HMGB1	O49595	High mobility group B protein 1	20646326
HUS1	Q9SSQ5	F6D8.25	20646326
KU70	Q9FQ08	ATP-dependent DNA helicase 2 subunit KU70	20646326
KU80	Q9FQ09	ATP-dependent DNA helicase 2 subunit KU80	20646326
MAGLP	Q9C5J2	DNA-3-methyladenine glycosylase II	20646326

MBD4	Q0IGK1	At3g07930	20646326
MLH1	Q9ZRV4	DNA mismatch repair protein MLH1 (Fragment)	20646326
MND1	Q8GYD2	Meiotic nuclear division protein 1 homolog	20646326
MRE11	Q9XGM2	Double-strand break repair protein MRE11	20646326
MRI40	O82638	At4g32960	InParanoid: hMRI40 ortholog
MSH2	O24617	DNA mismatch repair protein Msh2	20646326
MSH3	O65607	DNA mismatch repair protein Msh3	20646326
MSH6	O04716	DNA mismatch repair protein Msh6-1	20646326
MUS81	O65562	Putative uncharacterized protein AT4g30870	20646326
MUTY	Q9SU12	Adenine DNA glycosylase like protein	20646326
MYST1	Q9LXD7	MYST-like histone acetyltransferase 2	InParanoid: hMYST1 ortholog
NBS1	Q0H8D7	NBS1	20646326
NTH	Q94EJ4	At1g05900/T20M3_15	20646326
PARP1	Q9ZP54	Poly [ADP-ribose] polymerase 1	20646326
PARP2	Q11207	Poly [ADP-ribose] polymerase 2	20646326
PARP3	Q9FK91	Poly [ADP-ribose] polymerase 3	20646326
PCNA	Q9M7Q7	Proliferating cellular nuclear antigen 1	20646326
PHR2	Q8LB72	Blue-light photoreceptor PHR2	20646326
PMS1	Q941I6	DNA mismatch repair protein	20646326
PNKP	Q9LKB5	Putative uncharacterized protein	20646326
PR19A	Q94BR4	Pre-mRNA-processing factor 19 homolog 1	20646326
PRD1	O23277	Protein PRD1	20646326
RAD1	Q8L7G8	At4g17760	20646326
RAD17	Q9MBA3	Cell cycle checkpoint protein RAD17	20646326
RAD50	Q9SL02	DNA repair protein RAD50	20646326
RAD51	P94102	DNA repair protein RAD51 homolog 1	20646326
RAD54L	Q0PCS3	RAD54-like protein	20646326
RAD9	Q058K5	At3g05480	20646326
RBX1A	Q940X7	RING-box protein 1a	12381738
RD23B	Q84L32	Putative DNA repair protein RAD23-2	20646326
RecG	Q9ZVG0	Putative ATP-dependent DNA helicase RECG	20646326
RFA1	Q9FHJ6	Replication factor-A protein 1-like protein	20646326
RFC1	Q9C587	At5g22010	20646326
RFC2	Q9CAM7	At1g63160	20646326
RFC3	Q8VXX4	Putative replication factor C	20646326
RFC4	Q93ZX1	At1g21690	20646326
RFC5	Q9CAQ8	Putative replication factor C	20646326
RPA1	Q9SKI4	Replication factor A1	20646326
RPA2	Q9ZQ19	DNA replication protein A2 subunit	20646326
RPA3	Q9LXK1	Nucleic acid-binding, OB-fold-like protein	20646326
SCC12	Q9FQ20	Sister chromatid cohesion 1 protein 2	21867786
SIZ1	Q680Q4	E3 SUMO-protein ligase SIZ1	17905899
SKP1	Q9C5T5	SKP1-like protein ASK10 (Fragment)	12381738
SM3L2	Q9FNI6	Putative SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 3-like 2	20646326
SMAL1	Q9SX64	F11A17.14 protein	InParanoid: hSMAL1 ortholog
SMC1A	Q6Q1P4	Structural maintenance of chromosomes 1 protein	20646326
SNM1	Q38961	DNA cross-link repair protein SNM1	20646326
SPO11	Q9M4A2	Meiotic recombination protein SPO11-1	20646326
SSBP	Q84J78	Single-stranded DNA-binding protein, mitochondrial	20646326
SSRP1	Q05153	FACT complex subunit SSRP1	20646326
TIM	Q9FLX0	Timeless family protein	InParanoid: hTIM ortholog
TIPIN	Q8GW91	At3g02820	InParanoid: hTIPIN ortholog
TOP3	Q9LVP1	DNA topoisomerase	20646326
TOPB1	Q70X85	MEI1 protein	InParanoid: hTOPBP1 ortholog

UBC1	P25865	Ubiquitin-conjugating enzyme E2 1	20646326
UBC36	Q9FZ48	Ubiquitin-conjugating enzyme E2 36	20646326
UBP11	Q9MAQ3	Putative ubiquitin carboxyl-terminal hydrolase 11	InParanoid: hUBP11 ortholog
UEV1B	Q9CAB6	Ubiquitin-conjugating enzyme E2 variant 1B	20646326
ULA1	P42744	NEDD8-activating enzyme E1 regulatory subunit	20646326
UNG	Q9LIH6	At3g18630	20646326
UVH3	Q9ATY5	DNA repair protein UVH3	20646326
UVR3	O48652	(6-4)DNA photolyase	20646326
WEE1	Q8L4H0	Wee1-like protein kinase	17209125
WEX	Q84LH3	Werner Syndrome-like exonuclease	20646326
XPB1	Q38861	DNA repair helicase XPB1	20646326
XPC	Q8W489	At5g16630	20646326
XPF	Q9LKI5	DNA repair endonuclease UVH1	20646326
XRCC1	Q24JK4	At1g80420	20646326
XRCC2	Q682D3	DNA repair protein XRCC2 homolog	20646326
XRCC3	Q9FKM5	DNA repair protein XRCC3 homolog	20646326
XRCC4	Q682V0	DNA repair protein XRCC4	20646326
XRI1	Q6NLW5	Protein XRI1	21867786

ST1c. *S. cerevisiae* DDR proteins extracted from bibliography				
DDR protein	UniProt ID	Entrez_ID	Annotation	References (PMID)
1433E	P34730	6320304	Protein BMH2	Inparanoid: h1433E ortholog
ATM	P38110	6319383	Serine/threonine-protein kinase TEL1	19464966
ATR	P38111	6319612	Serine/threonine-protein kinase MEC1	19464966
CHK1	P38147	6319751	Serine/threonine-protein kinase CHK1	19464966
COM1	P46946	6321263	DNA endonuclease SAE2	9215887
CSM3	Q04659	6323692	Chromosome segregation in meiosis protein 3	21127252
CTF18	P49956	6323724	Chromosome transmission fidelity protein 18	21127252
CTF8	P38877	6321985	Chromosome transmission fidelity protein 8	21127252
CUL1	Q12018	6320070	Cell division control protein 53	Inparanoid: hCUL1 ortholog
CUL8	P47050	6322414	Regulator of Ty1 transposition protein 101	22353182
DCC1	P25559	10383774	Sister chromatid cohesion protein DCC1	21127252
Ddc1	Q08949	6325062	DNA damage checkpoint protein 1	19464966
DMA2	P53924	6324213	E3 ubiquitin-protein ligase DMA2	15146058
DNLI4	Q08387	6324578	DNA ligase 4	16388993
DPO4	P25615	10383782	DNA polymerase IV	12235149
DUN1	P39009	6320102	DNA damage response protein kinase DUN1	10357828
H2AX	P04912	6319470	Histone H2A.2	11140636
KAT5	Q08649	6324818	Histone acetyltransferase ESA1	12353039
KU70	P32807	6323940	ATP-dependent DNA helicase II subunit 1	21127252
KU80	Q04437	6323753	ATP-dependent DNA helicase II subunit 2	21127252
LCD1	Q04377	6320707	DNA damage checkpoint protein LCD1	11060031
LIF1	P53150	6321348	Ligase-interacting factor 1	16388993
MEC3	Q02574	6323319	DNA damage checkpoint control protein MEC3	19464966
MEK1	P24719	6324927	Meiosis-specific serine/threonine-protein kinase MEK1	1741279
MET18	P40469	6322063	DNA repair/transcription protein MET18/MMS19	8943333
MLH1	P38920	6323819	DNA mismatch repair protein MLH1	9368761
MMS1	Q06211	6325422	Methyl methanesulfonate-sensitivity protein 1	21127252
MMS2	P53152	6321351	Ubiquitin-conjugating enzyme variant MMS2	11440714
MMS22	Q06164	6323352	Methyl methanesulfonate-sensitivity protein 22	15718301
MPH1	P40562	6322192	ATP-dependent DNA helicase MPH1	15126389
MPIP3	P23748	6323679	M-phase inducer phosphatase	Inparanoid: hMPIP3 ortholog
MRC1	P25588	10383754	Mediator of replication checkpoint protein 1	21127252
MRE11	P32829	6323880	Double-strand break repair protein MRE11	20348017
MSH2	P25847	6324482	DNA mismatch repair protein MSH2	9368761
MSH3	P25336	157285763	DNA mismatch repair protein MSH3	9368761
MSH6	Q03834	6320302	DNA mismatch repair protein MSH6	8723353
NEJ1	Q06148	6323295	Non-homologous end-joining protein 1	16388993
PCNA	P15873	6319564	Proliferating cell nuclear antigen	20981145
PIAS4	Q12216	6324730	E3 SUMO-protein ligase SIZ2	Inparanoid: hPIAS4 ortholog
PLK1	P32562	6323643	Cell cycle serine/threonine-protein kinase CDC5/MSD2	15920482
PMS2	P14242	46562124	DNA mismatch repair protein PMS1	20981145
PP4R2	P38193	6319425	Serine/threonine-protein phosphatase 4 regulatory subunit 2	21127252
PP4R3	P40164	6324128	Serine/threonine-protein phosphatase 4 regulatory subunit 3	16299494
PSY3	Q12318	6323408	Platinum sensitivity protein 3	21127252
RAD1	P06777	6325235	DNA repair protein RAD1	8479526
RAD10	P06838	6323543	DNA repair protein RAD10	20981145
RAD14	P28519	6323857	DNA repair protein RAD14	20981145
RAD16	P31244	6319590	DNA repair protein RAD16	20348017
RAD17	P48581	6324944	DNA damage checkpoint control protein RAD17	19464966

RAD18	P10862	6319911	Postreplication repair E3 ubiquitin-protein ligase RAD18	15388802	
RAD2	P07276	6321697	DNA repair protein RAD2	20981145	
RAD23	P32628	6320798	UV excision repair protein RAD23	20981145	
RAD24	P32641	6321021	Checkpoint protein RAD24	15369670	19464966
RAD25	Q00578	6322048	DNA repair helicase RAD25	20981145	
RAD26	P40352	6322495	DNA repair and recombination protein RAD26	12024048	
RAD3	P06839	6321019	DNA repair helicase RAD3	8631896	
RAD4	P14736	6321010	DNA repair protein RAD4	9837874	
RAD5	P32849	6323060	DNA repair protein RAD5	16224103	
RAD50	P12753	6324079	DNA repair protein RAD50	20981145	
RAD51	P25454	6320942	DNA repair protein RAD51	12235149	
RAD52	P06778	27808713	DNA repair and recombination protein RAD52	21127252	
RAD53	P22216	6325104	Serine/threonine-protein kinase RAD53	12724400	19464966
RAD54	P32863	6321275	DNA repair and recombination protein RAD54	15369670	
RAD55	P38953	6320281	DNA repair protein RAD55	15369670	
RAD57	P25301	6320207	DNA repair protein RAD57	15369670	
RAD59	Q12223	6320144	DNA repair protein RAD59	15369670	
RAD7	P06779	6322512	DNA repair protein RAD7	15177043	
RAD9A	P14737	6320423	DNA repair protein RAD9	19464966	
RBX1	Q08273	6324438	RING-box protein HRT1	10579999	
RCK1	P38622	6321280	Serine/threonine-protein kinase RCK1	10778743	19230643
RCK2	P38623	6323277	Serine/threonine-protein kinase RCK2	10778743	19230643
RFA1	P22336	6319321	Replication factor A protein 1	20981145	
RFA2	P26754	6324017	Replication factor A protein 2	20981145	
RFA3	P26755	6322288	Replication factor A protein 3	20981145	
RFX1	P48743	6323205	RFX-like DNA-binding protein RFX1	16107689	
RT109	Q07794	6323027	Histone acetyltransferase RTT109	17272722	
SGS1	P35187	6323844	ATP-dependent helicase SGS1	20981145	
SKP1	P52286	6320535	Suppressor of kinetochore protein 1	8706132	
SLX5	P32828	6320191	E3 ubiquitin-protein ligase complex SLX5-SLX8 subunit SLX5	17591698	21127252
SLX8	P40072	6320962	E3 ubiquitin-protein ligase complex SLX5-SLX8 subunit SLX8	17591698	
SMC1A	P32908	14318514	Structural maintenance of chromosomes protein 1	11983169	
SMC5	Q08204	6324539	Structural maintenance of chromosomes protein 5	15738391	
SMC6	Q12749	6323415	Structural maintenance of chromosomes protein 6	15738391	
SWE1	P32944	6322274	Mitosis inhibitor protein kinase SWE1	15920482	
SWI4	P25302	6320957	Regulatory protein SWI4	12724400	
TAOK1	P38692	6321894	Serine/threonine-protein kinase KIC1	Inparanoid: hTAOK1 ortholog	
TOF1	P53840	6324056	Topoisomerase 1-associated factor 1	21127252	
UBE2N	P52490	6320297	Ubiquitin-conjugating enzyme E2 13	21127252	
UBC4	P15731	6319556	Ubiquitin-conjugating enzyme E2 4	18202552	
UBP11	P39538	6322264	Ubiquitin carboxyl-terminal hydrolase 12	Inparanoid: hUBP11 ortholog	
XRS2	P33301	6320577	DNA repair protein XRS2	20981145	

ST1d. *H. sapiens* DDR proteins extracted from bibliography

DDR protein	UniProt ID	Annotation	References (PMID)		Uni/multi-gene family
1433E	P62258	14-3-3 protein epsilon	18082599	20413225	18
ATM	Q13315	Serine-protein kinase ATM [2.7.1.11]	17525332	18082599	35 / Fam16
ATR	Q13535	Serine/threonine-protein kinase ATR [2.7.1.11]	18082599	20947357	35 / Fam16
ATRIP	Q8WXE1	ATR-interacting protein	18082599	18594563	64
BARD1	Q99728	BRCA1-associated RING domain protein 1 [6.3.2.-]	18082599	20029420	50
BLM	P54132	Bloom syndrome protein [3.6.1.-]	15257300	19843584	5
BRCA1	P38398	Breast cancer type 1 susceptibility protein [6.3.2.-]	20029420	21203981	49
BRCA2	P51587	Breast cancer type 2 susceptibility protein	19268590	21203981	76
BRCC3 (BRCC36)	P46736	Lys-63-specific deubiquitinase BRCC36 [3.1.2.15]	18082599		53
BRE (BRCC45)	Q9NXR7	BRCA1-A complex subunit BRE	19261748		62
CDT1	Q9H211	DNA replication factor Cdt1	18082599		77
CHK1	O14757	Serine/threonine-protein kinase Chk1 [2.7.11.1]	19230643	21088254	45 / Fam14
CHK2	O96017	Serine/threonine-protein kinase Chk2 [2.7.11.1]	19230643	19473886	45 / Fam14
CLSPN	Q9HAW4	Claspin	18082599		70
CUL1	Q13616	Cullin-1	19903939		21 / Fam6
CUL4	Q13619	Cullin-4	18082599		21 / Fam6
DCR1B (Apollo)	Q9H816	5' exonuclease Apollo	18469862	19411856	40 / Fam9
DCR1C (Artemis)	Q96SD1	Protein artemis [3.1.-.-]	20543526		40 / Fam9
DDB1	Q16531	DNA damage-binding protein 1, UV-damaged DNA-binding factor, DDB p127 subunit	18082599		33
DNA2L (DNA2)	P51530	DNA2-like helicase	19487465	21325134	63
DNLI4 (LIG4)	P49917	DNA ligase 4 [6.5.1.1]	11357144	21329706	37
DTL (CDT2)	Q9NZJ0	Denticleless protein homolog	18082599		56
EME1 (MMS4)	Q96AY2	Crossover junction endonuclease EME1 (MMS4)	12686547		59
ERCC1	P07992	DNA excision repair protein ERCC-1	16855787	18166977	22
ERCC2 (XPD)	P18074	TFIIH basal transcription factor complex helicase XPD subunit [3.6.4.12]	16855787	18166977	9
ERCC3 (XPB)	P19447	TFIIH basal transcription factor complex helicase XPB subunit [3.6.4.12]	16855787	18166977	3 / Fam17
ERCC5 (XPG)	P28715	DNA repair protein complementing XP-G cells	16855787	18166977	24 / Fam10
ERCC6 (CSB)	Q03468	DNA excision repair protein ERCC-6 [3.6.4.-]	16855787	18166977	39
ERCC8 (CSA)	Q13216	DNA excision repair protein ERCC-8	16855787	18166977	51
EXO1	Q9UQ84	Exonuclease 1	14676842	18048416	24 / Fam10
F175A (Abraxas)	Q6UWZ7	BRCA1-A complex subunit Abraxas	18082599		94
FACD2	Q9BXW9	Fanconi anemia group D2 protein; Protein FACD2	18082599		42
FANCM	Q8IYD8	Fanconi anemia group M protein	22615860		3 / Fam17
FBW1A (BTRCP)	Q9Y297	F-box/WD repeat-containing protein 1A, F-box and WD repeats protein beta-TrCP	18082599		67

FBX31	Q5XUX0	F-box only protein 31	18082599	19412162	86
H2AX	P16104	Histone H2A.x; H2a/x	18082599		17
HERC2	O95714	E3 ubiquitin-protein ligase HERC2 [6.3.2.-]	20023648	20406985	72
HNRPK	P61978	Heterogeneous nuclear ribonucleoprotein K	19579069		75
HUS1	O60921	Checkpoint protein HUS1	18082599		38
KAT5 (Tip60)	Q92993	Histone acetyltransferase KAT5 [2.3.1.48]	18082599	20160506	41 / Fam8
MAPK2	P49137	MAP kinase-activated protein kinase 2 [2.7.11.1]	18082599	19230643	45 / Fam14
MDC1	Q14676	Mediator of DNA damage checkpoint protein 1	18082599	21482717	89 / Fam7*
MDM2	Q00987	E3 ubiquitin-protein ligase Mdm2 [6.3.2.-]; p53-binding protein Mdm2	18082599		88 / Fam2
MDM4	O15151	Protein Mdm4	18082599		88 / Fam2
MK03 (ERK1)	P27361	Extracellular signal-regulated kinase 1 (Mitogen-activated protein kinase 3)	22576881		45 / Fam14
MLH1	P40692	DNA mismatch repair protein Mlh1	16464007	16612326	6 / Fam13
MMS21 (NSMCE2)	Q96MF7	E3 SUMO-protein ligase NSE2 (NSMCE2)	16055714		52
MPIP1 (Cdc25A)	P30304	M-phase inducer phosphatase 1 [3.1.3.48]	18082599		78 / Fam4
MPIP3 (Cdc25C)	P30307	M-phase inducer phosphatase 3 [3.1.3.48]	18082599		78 / Fam4
MRE11	P49959	Double-strand break repair protein MRE11A	18082599	21252998	13
MRI40	Q9NWV8	BRCA1-A complex subunit MERIT40	19261746	20029420	54
MSH2	P43246	DNA mismatch repair protein Msh2; hMSH2	16464007	16612326	2 / Fam15
MSH3	P20585	DNA mismatch repair protein Msh3; hMSH3	16464007	16612326	2 / Fam15
MSH6	P52701	DNA mismatch repair protein Msh6; hMSH6	16464007	16612326	2 / Fam15
MTA2	O94776	Metastasis-associated protein MTA2	12920132	IRB DDR Meeting, 2012. V. Costanzo	79
MUS81	Q96NY9	Crossover junction endonuclease MUS81	12686547		12 / Fam11
MYST1	Q9H7Z6	Probable histone acetyltransferase MYST1 [2.3.1.48]	20479123		41 / Fam8
NBN (NBS1)	O60934	Nibrin	18082599	21252998	61
NR4A2	Q6NXU0	Nuclear receptor subfamily 4 group A member 2	21979916		73
PALB2	Q86YC2	Partner and localizer of BRCA2	19268590	21203981	93
PARP1	P09874	Poly [ADP-ribose] polymerase 1, PARP-1 [2.4.2.30]	18082599		43 / Fam3
PARP2	Q9UGN5	Poly [ADP-ribose] polymerase 2, PARP-2 [2.4.2.30]	19847258		43 / Fam3
PAXI1 (PTIP)	Q6ZW49	PAX-interacting protein 1	21035408		58 / Fam7*
PCNA	P12004	Proliferating cell nuclear antigen	18082599	20074041	4
PIAS1	B3KSY9	E3 SUMO-protein ligase PIAS1	20016603	20074042	55 / Fam1
PIAS4	Q8N2W9	E3 SUMO-protein ligase PIAS4	20016603	20074042	55 / Fam1
PLK1	P53350	Serine/threonine-protein kinase PLK1 [2.7.11.21]	18082599	20126263	65
PMS2	P54278	Mismatch repair endonuclease PMS2 [3.1.-.-]	16464007	16612326	6 / Fam13
PRKDC (DNA-PK)	P78527	DNA-dependent protein kinase catalytic subunit, DNA-PKcs [2.7.11.1]	11357144	21487018	35 / Fam16
RAD1	O60671	Cell cycle checkpoint protein RAD1; hRAD1 [3.1.11.2]	18082599		28
RAD17	O75943	Cell cycle checkpoint protein RAD17	18082599		26
RAD18	Q9NS91	E3 ubiquitin-protein ligase RAD18 [6.3.2.-]	18082599	20971043	46
RAD23B	P54727	UV excision repair protein RAD23 homolog B	16855787	18166977	11
RAD50	Q92878	DNA repair protein RAD50; hRAD50 [3.6.-.-]	18082599	21252998	1

RAD51	Q06609	DNA repair protein RAD51 homolog 1	18082599		0
RAD9A	Q99638	Cell cycle checkpoint control protein RAD9A, hRAD9 [3.1.11.2]	18082599		30
RBBP8 (CTIP)	Q99708	Retinoblastoma-binding protein 8	17965729		83
RBX1	P62877	RING-box protein 1	18082599	21115485	10
RMI1	Q9H9A7	RecQ-mediated genome instability protein 1	15775963	16595695	80
RNF168	Q8IYW5	E3 ubiquitin-protein ligase RNF168 [6.3.2.-]	19203579		90
RNF8	O76064	E3 ubiquitin-protein ligase RNF8 [6.3.2.-]	18082599		82
RPA1	P27694	Replication protein A 70 kDa DNA-binding subunit, RP-A p70	18082599	18166977	16
RPA2	P15927	Replication protein A 32 kDa subunit; RP-A p32	18082599	18166977	34
RPA3	P35244	Replication protein A 14 kDa subunit; RP-A p14	18082599	18166977	23
SIR1 (SIRT1)	Q96EB6	NAD-dependent protein deacetylase sirtuin-1 (Regulatory protein SIR2 homolog 1)	22586264		68
SKP1	P63208	S-phase kinase-associated protein 1; Cyclin-A/CDK2-associated protein p19; Transcription elongation factor B	19903939		15
SLX1	Q9BQ83	Structure-specific endonuclease subunit SLX1	19596236		66
SLX4	Q8IY92	Structure-specific endonuclease subunit SLX4	19596235	19596236	85
SMAL1	Q9NZC9	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 [3.6.1.-]	19793861	19841479	44
SMC1A	Q14683	Structural maintenance of chromosomes protein 1A	19842212	21139141	8 / Fam12
SMC5	Q8IY18	Structural maintenance of chromosomes protein 5	17589526		8 / Fam12
SMC6	Q96SB8	Structural maintenance of chromosomes protein 6	17589526		8 / Fam12
SOX4	Q06945	Transcription factor SOX-4	19234109		87
TAOK1	Q7L7X3	Serine/threonine-protein kinase TAO1 [2.7.11.1]	18082599		45 / Fam14
TDP1 (TYDP1)	Q9NUW8	Tyrosyl-DNA phosphodiesterase 1 [3.1.4.-]	16141202		25
TDP2 (TYDP2)	O95551	Tyrosyl-DNA phosphodiesterase 2 [3.1.4.-]	19794497		48
TERF2 (TRF2)	Q15554	Telomeric repeat-binding factor 2	20655466		92
TIF1B (KAP1)	Q13263	Transcription intermediary factor 1-beta (KRAB-associated protein 1)	17056014	17178852	81
TIMELESS	Q9UNS1	Protein timeless	20233725		29
TIPIN	Q9BVW5	TIMELESS-interacting protein	20233725		57
TOP3A	Q13472	DNA topoisomerase 3-alpha	16595695		7
TOPB1	Q92547	DNA topoisomerase 2-binding protein 1	18082599	19464966	27
TP53B	Q12888	Tumor suppressor p53-binding protein 1	18082599		84
TRIPC (TRIP12)	Q14669	Probable E3 ubiquitin-protein ligase TRIP12	17525332	22884692	36
UBE2N (UBC13)	P61088	Ubiquitin-conjugating enzyme E2 N [6.3.2.19]	18082599		20 / Fam5
UBE2T	Q9NPD8	Ubiquitin-conjugating enzyme E2 T [6.3.2.19]	22615860		20 / Fam5
UBP11 (USP11)	P51784	Ubiquitin carboxyl-terminal hydrolase 11 [3.1.2.15]	20233726		14
UBR5	O95071	E3 ubiquitin-protein ligase UBR5	11714696	22884692	74
UIMC1 (RAP80)	Q96RL1	BRCA1-A complex subunit RAP80	17525341	18082599	91
WEE1	P30291	Wee1-like protein kinase, WEE1hu [2.7.10.2]	18082599		45 / Fam14
XLF (NHEJ1)	Q9H9Q4	Non-homologous end-joining factor 1; Protein cernunnos; XRCC4-like factor	17038309	21329706	71

XPA	P23025	DNA repair protein complementing XP-A cells; Xeroderma pigmentosum group A-complementing protein	16855787	18166977	69
XPC	Q01831	DNA repair protein complementing XP-C cells	16855787	18166977	19
XPF (ERCC4)	Q92889	DNA repair endonuclease XPF; DNA excision repair protein ERCC-4	16855787	18166977	12 / Fam11
XRCC1	P18887	DNA repair protein XRCC1	18971944		47
XRCC4	Q13426	DNA repair protein XRCC4	11357144	21329706	60
XRCC5 (Ku80)	P13010	ATP-dependent DNA helicase 2 subunit 1 [3.6.1.-]	19260023		31
XRCC6 (Ku70)	P12956	ATP-dependent DNA helicase 2 subunit 2 [3.6.1.-]	19260023		32

ST2. Organisms and databases

ORGANISM					DATABASE			No. of PROTEINS
Organism ID	Species	Kingdom	Phylum	Observation	Status/Seq. version	Source	Download date	
Ame	*Apis mellifera*	*Eukaryota*	*Metazoa*		draft assembly	NCBI	Feb 5, 2010	9257
Ath	*Arabidopsis thaliana*	*Eukaryota*	*Viridiplantae*		complete	EBI	May 1, 2010	36628
Ava	*Anabaena variabilis ATCC 29413*	*Bacteria*	*Cyanobacteria*		complete	NCBI	Feb 3, 2010	5661
Bap	*Buchnera aphidicola strain 5A*	*Bacteria*	*Proteobacteria*	Endosymbiont	complete	NCBI	Apr 19, 2010	555
Bde	*Batrachochytrium dendrobatidis*	*Eukaryota*	*Fungi*	Pathogen	draft assembly 1.0	Broad Institute	Feb 9, 2010	8818
Bfl	*Branchiostoma floridae*	*Eukaryota*	*Metazoa*		draft assembly 1.0	JGI	Feb 9, 2010	50817
Bna	*Bigelowiella natans*	*Eukaryota*	*Rhizaria*	Secondary endosymbiosis	draft assembly	NCBI	Feb 9, 2010	344
Bsu	*Bacillus subtilis*	*Bacteria*	*Firmicutes*		complete	NCBI	Feb 3, 2010	4176
Cel	*Caenorhabditis elegans*	*Eukaryota*	*Metazoa*		complete	EBI	May 1, 2010	24243
Cin	*Ciona intestinalis*	*Eukaryota*	*Metazoa*		draft assembly 2.0	JGI	Feb 3, 2010	14002
Cko	*Candidatus Korarchaeum cryptofilum OPF8*	*Archaea*	*Korarchaeota*		complete	NCBI	Feb 3, 2010	1602
Cne	*Cryptococcus neoformans*	*Eukaryota*	*Fungi*	Pathogen	complete	NCBI	Feb 4, 2010	6475
Cpa	*Cryptosporidium parvum*	*Eukaryota*	*Alveolata*	Parasite	draft assembly	NCBI	Feb 4, 2010	3805
Cre	*Chlamydomonas reinhardtii*	*Eukaryota*	*Viridiplantae*		draft assembly 4.0	JGI	Feb 5, 2010	16709
Cte	*Capitella teleta*	*Eukaryota*	*Metazoa*		draft assembly 1.0	JGI	Feb 8, 2010	32415
Ddi	*Dictyostelium discoideum*	*Eukaryota*	*Amoebozoa*		draft assembly	NCBI	Feb 4, 2010	13377
Dme	*Drosophila melanogaster*	*Eukaryota*	*Metazoa*		complete	NCBI	Feb 5, 2010	21603
Dra	*Deinococcus radiodurans R1*	*Bacteria*	*Deinococcus-Thermus*		complete	NCBI	Feb 3, 2010	3167
Dre	*Danio rerio*	*Eukaryota*	*Metazoa*		draft assembly	NCBI	Feb 5, 2010	26709
Eco	*Escherichia coli K12*	*Bacteria*	*Proteobacteria*		complete	NCBI	Feb 3, 2010	4149
Ecu	*Encephalitozoon cuniculi*	*Eukaryota*	*Fungi*	Parasite	complete	NCBI	Feb 4, 2010	1996
Ehu	*Emiliania huxleyi 1516*	*Eukaryota*	*Haptophyceae*		draft assembly 1.0	JGI	Feb 3, 2010	39125
Gga	*Gallus gallus*	*Eukaryota*	*Metazoa*		draft assembly	EBI	May 1, 2010	22194
Gob	*Gemmata obscuriglobus UQM 2246*	*Bacteria*	*Planctomycetes*		draft assembly	NCBI	Feb 3, 2010	7989
Gth	*Guillardia theta*	*Eukaryota*	*Cryptophyta*	Secondary endosymbiosis	draft assembly	NCBI	Feb 5, 2010	632
Hsa	*Homo sapiens*	*Eukaryota*	*Metazoa*		complete	EBI	May 1, 2010	64611
Mac	*Methanosarcina acetivorans C2A*	*Archaea*	*Euryarchaeota*		complete	NCBI	Feb 3, 2010	4540
Mbr	*Monosiga brevicollis*	*Eukaryota*	*Choanoflagellida*		draft assembly 1.0	JGI	Feb 3, 2010	9196
Mdo	*Monodelphis domestica*	*Eukaryota*	*Metazoa*		draft assembly	EBI	Feb 5, 2010	32541
Mge	*Mycoplasma genitalium G37*	*Bacteria*	*Tenericutes*	Parasite	complete	NCBI	Feb 3, 2010	475
Mus	*Mus musculus*	*Eukaryota*	*Metazoa*		complete	EBI	Feb 5, 2010	43905
Ngr	*Naegleria gruberi*	*Eukaryota*	*Heterolobosea*		draft assembly 1.0	JGI	Feb 3, 2010	15753
Nve	*Nematostella vectensis*	*Eukaryota*	*Metazoa*		draft assembly 1.0	JGI	Feb 8, 2010	27273
Oan	*Ornithorhynchus anatinus*	*Eukaryota*	*Metazoa*		draft assembly	EBI	Feb 5, 2010	26836
Osa	*Oryza sativa*	*Eukaryota*	*Viridiplantae*		complete	NCBI	Feb 5, 2010	26777
Pfa	*Plasmodium falciparum*	*Eukaryota*	*Alveolata*	Parasite	complete	NCBI	Feb 4, 2010	5265
Ppa	*Physcomitrella patens subsp patens*	*Eukaryota*	*Viridiplantae*		draft assembly 1.1	JGI	Feb 9, 2010	35938
Pst	*Pirellula staleyi DSM 6068*	*Bacteria*	*Planctomycetes*		draft assembly	NCBI	Feb 3, 2010	4717
Ptr	*Phaeodactylum tricornutum*	*Eukaryota*	*Stramenopiles*		draft assembly 2.0	JGI	Feb 5, 2010	10402
Sce	*Saccharomyces cerevisiae*	*Eukaryota*	*Fungi*		complete	NCBI	Feb 4, 2010	5880
Sja	*Schistosoma japonicum*	*Eukaryota*	*Metazoa*	Parasite	draft assembly	CHGC	Feb 8, 2010	13469
Spo	*Schizosaccharomyces pombe*	*Eukaryota*	*Fungi*		complete	NCBI	Feb 4, 2010	5003
Sso	*Sulfolobus solfataricus P2*	*Archaea*	*Crenarchaeota*		complete	NCBI	Feb 3, 2010	2977
Tad	*Trichoplax adhaerens*	*Eukaryota*	*Metazoa*		draft assembly 1.0	JGI	Jan 26, 2010	11520
Tbr	*Trypanosoma brucei*	*Eukaryota*	*Euglenozoa*	Parasite	draft assembly	NCBI	Feb 4, 2010	9279
Tca	*Tribolium castaneum*	*Eukaryota*	*Metazoa*		draft assembly	NCBI	Feb 5, 2010	9833
Xtr	*Xenopus tropicalis*	*Eukaryota*	*Metazoa*		draft assembly 4.1	NCBI	Feb 3, 2010	27916

ST3. Human DDR orthologs obtained using other seed organisms

UniProt ID	Gene Name	Protein Name	Seed organism
P29372	3MG	DNA-3-methyladenine glycosylase	*A. thaliana*
Q13686	ALKB1	Alkylated DNA repair protein alkB homolog 1	*A. thaliana*
Q9UBZ4	APEX2	DNA-(apurinic or apyrimidinic site) lyase 2	*A. thaliana*
P51946	CCNH	Cyclin-H	*A. thaliana*
P50613	CDK7	Cyclin-dependent kinase 7	*A. thaliana*
O76031	CLPX	ATP-dependent Clp protease ATP-binding subunit clpX-like, mitochondrial	*E. coli*
Q16526	CRY1	Cryptochrome-1	*A. thaliana*
P0CG13	CTF8	Chromosome transmission fidelity protein 8 homolog	*S. cerevisiae*
Q9BVC3	DCC1	Sister chromatid cohesion protein DCC1	*S. cerevisiae*
Q6PJP8	DCR1A	DNA cross-link repair 1A protein	*A. thaliana*
Q92466	DDB2	DNA damage-binding protein 2	*A. thaliana*
Q7L5Y6	DET1	DET1 homolog	*A. thaliana*
Q14565	DMC1	Meiotic recombination protein DMC1/LIM15 homolog	*E. coli / A. thaliana*
P28340	DPOD1	DNA polymerase delta catalytic subunit	*E. coli*
P09884	DPOLA	DNA polymerase alpha catalytic subunit	*A. thaliana*
Q9UGP5	DPOLL	DNA polymerase lambda	*S. cerevisiae*
Q9UKC9	FBXL2	F-box/LRR-repeat protein 2	*S. cerevisiae*
Q14527	HLTF	Helicase-like transcription factor	*S. cerevisiae*
P09429	HMGB1	High mobility group protein B1	*A. thaliana*
P36776	LONM	Lon protease homolog, mitochondrial	*E. coli*
Q96T76	MMS19	MMS19 nucleotide excision repair protein homolog	*S. cerevisiae*
Q9BWT6	MND1	Meiotic nuclear division protein 1 homolog	*A. thaliana*
Q9UIF7	MUTYH	A/G-specific adenine DNA glycosylase	*A. thaliana*
Q96FI4	NEIL1	Endonuclease 8-like 1	*A. thaliana*
P78549	NTHL1	Endonuclease III-like protein 1	*A. thaliana*
Q6IN85	P4R3A	Serine/threonine-protein phosphatase 4 regulatory subunit 3A	*S. cerevisiae*
Q9UBT6	POLK	DNA polymerase kappa	*E. coli*
Q9NY27	PP4R2	Serine/threonine-protein phosphatase 4 regulatory subunit 2	*S. cerevisiae*
Q9UMS4	PRP19	Pre-mRNA-processing factor 19	*A. thaliana*
P43351	RAD52	DNA repair protein RAD52 homolog	*S. cerevisiae*
Q92698	RAD54	DNA repair and recombination protein RAD54-like	*A. thaliana / S. cerevisiae*
P46063	RECQ1	ATP-dependent DNA helicase Q1	*S. cerevisiae*
Q9UBZ9	REV1	DNA repair protein REV1	*A. thaliana*
P35251	RFC1	Replication factor C subunit 1	*A. thaliana*
P35250	RFC2	Replication factor C subunit 2	*A. thaliana*
P40938	RFC3	Replication factor C subunit 3	*A. thaliana*
P35249	RFC4	Replication factor C subunit 4	*A. thaliana*
P40937	RFC5	Replication factor C subunit 5	*A. thaliana*
P48378	RFX2	DNA-binding protein RFX2	*S. cerevisiae*
Q04837	SSBP	Single-stranded DNA-binding protein, mitochondrial	*E. coli / A. thaliana*
Q08945	SSRP1	FACT complex subunit SSRP1	*A. thaliana*
Q6P1K8	T2H2L	General transcription factor IIH subunit 2-like protein	*A. thaliana*
P32780	TF2H1	General transcription factor IIH subunit 1	*A. thaliana*
Q9Y2X8	UB2D4	Ubiquitin-conjugating enzyme E2 D4	*S. cerevisiae*
Q13404	UB2V1	Ubiquitin-conjugating enzyme E2 variant 1	*A. thaliana*
Q15819	UB2V2	Ubiquitin-conjugating enzyme E2 variant 2	*S. cerevisiae*
P49459	UBE2A	Ubiquitin-conjugating enzyme E2 A	*A. thaliana*
Q13564	ULA1	NEDD8-activating enzyme E1 regulatory subunit	*A. thaliana*
P13051	UNG	Uracil-DNA glycosylase	*A. thaliana*
Q14191	WRN	Werner syndrome ATP-dependent helicase	*E. coli*
O43543	XRCC2	DNA repair protein XRCC2	*A. thaliana*
O43542	XRCC3	DNA repair protein XRCC3	*A. thaliana*
B4DM52	B4DM52	DNA ligase	*A. thaliana*

ST4. Orthologs domain composition

Protein name	Domain architecture	Conservation	Variations	Species/phylogenetic group with variation
1433E	14-3-3	Yes		
ATM	PI3 PI4 kinase - FATC	No	**PWWP** - PI3 PI4 kinase - FATC	*A. thaliana*
ATM	PI3 PI4 kinase - FATC	No *	**TAN** - PI3 PI4 kinase - FATC	*C. neoformans* and *S. cerevisiae*
ATM	PI3 PI4 kinase - FATC	No *	**FAT** - PI3 PI4 kinase - FATC	Most species
ATM	PI3 PI4 kinase - FATC	No *	**TAN - FAT** - PI3 PI4 kinase - FATC	*S. pombe*, *B. floridae* and from *X. tropicalis*
ATR	PI3_PI4_kinase - FATC	Yes		
BARD1	Ank - BRCT	No	**zf-C3HC4** - Ank - BRCT	No phylogenetic trend
BLM	DEAD - Helicase_C - RQC - HRDC	No	DEAD - Helicase_C - RQC - **Helicase_Sgs1**	*S. cerevsiae* (Note: HRDC and Helicase_Sgs1 are members of the HRDC-like (CL0426) clan)
BLM	DEAD - Helicase C - RQC - HRDC	No	DEAD - Helicase C - RQC - HRDC - **HRDC - HRDC**	*D. radiodurans*
BLM	DEAD - Helicase_C - RQC - HRDC	No	DEAD - Helicase_C - RQC - HRDC - **GerE**	*A. variabilis*
BLM	DEAD - Helicase C - RQC - HRDC	No	**BDHCT** - DEAD - Helicase C - RQC - HRDC	Vertebrata
BRCA1	zf-C3HC4 - BRCT	Yes		
BRCA2	BRCA2 - BRCA-2 helical - BRCA-2 OB1 - BRCA-2 OB3	No **	BRCA2 - BRCA-2 helical - BRCA-2 OB1 - **Tower** - BRCA-2 OB3	Vertebrata
BRCC3	Mov34	Yes		
BRE	BRE	Yes		
CDT1	CDT1	Yes		
CHK1	Pkinase	Yes		
CHK2	FHA - Pkinase	Yes		
CUL1	Cullin - Cullin_Nedd8	Yes		
CUL4	Cullin - Cullin Nedd8	Yes		
DCR1B	Lactamase B - DRMBL	Yes		
DCR1C	Lactamase B - DRMBL	Yes		
DDB1	CPSF A	Yes *		
DNA2L	Dna2	Yes		
DNLI4	DNA_ligase_A_N - DNA_ligase_A_M - DNA_ligase_A_C - BRCT - BRCT	No	DNA_ligase_A_N - DNA_ligase_A_M - DNA_ligase_A_C - BRCT - **DNA_ligase_IV** - BRCT	*D. discoideum* and Metazoa (except *S. japonicum*)
DTL	WD40	Yes **		
EME1	ERCC4	Yes		
ERCC1	Rad10 - HHH	Yes		
ERCC2	DEAD 2 - DUF1227	Yes		
ERCC3	ResIII - Helicase_C	Yes		
ERCC5	XPG N - XPG I	Yes		
ERCC6	SNF2_N - Helicase_C	Yes		
ERCC8	WD40	Yes		
EXO1	XPG_N - XPG_I	Yes		
FANCM	ResIII - Helicase C	No	**DEAD** - Helicase C	No phylogenetic trend (Note: ResIII and DEAD are members of the P-loop_NTPase (CL0023) clan)
FBW1A	F-box - WD40 - WD40 - WD40 - WD40 - WD40 - WD40 - WD40	No	**Beta-TrCP_D** - F-box - WD40 - WD40 - WD40 - WD40 - WD40 - WD40 - WD40	Metazoa
FBX31	F-box - DUF3506	Yes		
H2AX	Histone	Yes		
HERC2	RCC1 - Cyt-b5 - MIB_HERC2 - Cul7 - ZZ - APC10 - RCC1 - HECT	Yes **		In human 5 N-terminal y 14 C-terminal RCC1 repeats
HNRPK	KH 1 - KH 1 - KH 1	No	**ROKNT** - KH 1 - KH 1 - KH 1	Vertebrata
HUS1	Hus1	Yes		
KAT5	Tudor-knot - MOZ SAS	Yes		
MAPK2	Pkinase	Yes		
MDC1	FHA - BRCT - BRCT	Yes		
MDM2	SWIB - zf-RanBP	Yes		
MDM4	SWIB - zf-RanBP	Yes		
MK03	Pkinase	Yes		
MLH1	HATPase_c - DNA_mis_repair - MutL_C	No	HATPase_c - DNA_mis_repair - **MutL_C**	Prokaryotes
MPIP1	Rhodanese	No	**M-inducer_phosp** - Rhodanese	From *X. tropicalis*
MPIP3	Rhodanese	No	**M-inducer_phosp** - Rhodanese	From *X. tropicalis*
MRE11	Metallophos - Mre11 DNA bind	Yes		
MSH2	MutS I - MutS II - MutS III - MutS IV - MutS V	Yes		
MSH3	MutS_I - MutS_II - MutS_III - MutS_IV - MutS_V	Yes		
MSH6	MutS I - MutS II - MutS III - MutS IV - MutS V	No	**PWWP** - MutS I - MutS II - MutS III - MutS IV - MutS V	*O. sativa*, and from *N. vectensis* to human, but not in fungi, arthropoda and *C. elegans*
MTA2	BAH - ELM2 - Myb_DNA-binding - GATA	Yes		
MUS81	ERCC4	Yes		
MYST1	Tudor-knot - MOZ_SAS	Yes		
NBN	FHA - BRCT - Nbs1 C	Yes		
NR4A2	zf-C4 - Hormone recep	Yes		
NSE2	zf-Nse	Yes		
PARP1	zf-PARP - PADR1 - BRCT - WGR - PARP reg - PARP	No	**zf-PARP -** zf-PARP - PADR1 - BRCT - WGR - PARP reg - PARP	Plants and Metazoa
PARP2	WGR - PARP_reg - PARP	No	**SAP - SAP** - WGR - PARP_reg - PARP	Plants
PAXI1	BRCT	Yes		
PCNA	PCNA_N - PCNA_C	Yes		
PIAS1	SAP - zf-MIZ	No	SAP - **PHD** - zf-MIZ	Plants
PIAS4	SAP - zf-MIZ	Yes		
PLK1	Pkinase	No	Pkinase - **POLO_box - POLO_box**	Eukaryotes
PMS2	HATPase c - DNA mis repair - MutL C	Yes		
PRKDC	NUC194 - FAT - PI3 PI4 kinase - FATC	Yes		
RAD1	Rad1	Yes		
RAD17	Rad17	Yes		
RAD18	zf-C3HC4 - SAP	Yes		
RAD50	SMC_N - Rad50_zn_hook - SMC_N	No	Rad50_zn_hook	From *S. japonicum* to human, except arthopoda, the SMC_N domains were not identified
RAD51	Rad51	No	**RecA**	Bacteria
RAD51	Rad51	No	**Cdd1** - Rad51	Archaea
RAD51	Rad51	No	Rad51 - **Lon_C**	*B. subtilis*
RAD51	Rad51	No	**HHH** - Rad51	From *T. brucei* to human
RAD9	Rad9	Yes		
RBBP8	CtIP N - SAE2	Yes		
RBX1	zf-C3HC4	Yes		
RD23B	ubiquitin - UBA - XPC-binding - UBA	Yes		
RFA1	Rep-A_N - tRNA_anti - Rep_fac-A_C	Yes		
RFA2	tRNA anti - RPA C	Yes		
RFA3	Rep_fac-A_3	Yes		
RMI1	DUF1767	Yes		
RN168	zf-C3HC4	Yes		
RNF8	FHA - zf-C3HC4	Yes		
SIRT1	DUF592 - SIR2	Yes		
SKP1	Skp1_POZ - Skp1	Yes		
SLX1	GIY-YIG (FANCL C / zf-RING-like)	Yes		
SLX4	BTB	Yes		
SMAL1	SNF2 N - Helicase C	No	**HARP** - SNF2 N - Helicase C	From N. vectensis
SMC1A	SMC_N	Yes		
SMC5	SMC N	Yes		
SMC6	SMC N	Yes		
SOX4	HMG box	Yes		
TAOK1	Pkinase	Yes		

TDP1	Tyr-DNA phospho	Yes *		
TDP2	Exo endo phos	No	**zf-RanBP** - Exo endo phos	O. sativa and P. patens
TERF2	TRF - Myb_DNA-binding	Yes		
TIF1B	zf-B box - zf-B box - PHD	Yes		
TIM	TIMELESS - TIMELESS_C	Yes		
TIPIN	Swi3	No	**zf-CCHC** - Swi3	Plants
TOP3A	Toprim - Topoisom bac - zf-C4 Topoisom - zf-GRF - zf-GRF	Yes *		
TOPB1	BRCT	Yes **		
TP53B	53-BP1 Tudor - BRCT - BRCT	Yes		
TRIPC	WWE - HECT	Yes *		
UBE2N	UQ con	Yes		
UBE2T	UQ_con	Yes		
UBP11	DUSP - UCH	Yes *		
UBR5	E3_UbLigase_EDD - zf-UBR - PABP - HECT	Yes		
UIMC1	UIM	Yes		
WEE1	Pkinase	Yes		
XLF	XLF	Yes		
XPA	XPA N - XPA C	Yes		
XPC	Rad4 - BHD_1 - BHD_2 - BHD_3	Yes		
XPF	ERCC4	Yes		
XRCC1	XRCC1_N - BRCT - BRCT	Yes		
XRCC4	XRCC4	Yes		
XRCC5	Ku N - Ku - Ku C - Ku PK bind	Yes		
XRCC6	Ku N - Ku - Ku C - SAP	Yes		
*	Some orthologs with slightly different domain architecture, most probably due to incorrect gene prediction or domains with bad scores.			
**	The number of repeats varies in different species			

ST5: Domain enrichment (Results 4.4.5)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
BRCT	PF00533	9	29	3.54E-15	6.16E-13
DNA_photolyase	PF00875	5	9	1.67E-10	2.91E-08
FAD_binding_7	PF03441	5	9	1.67E-10	2.91E-08
Helicase_C	PF00271	9	174	6.81E-08	1.19E-05
PARP_reg	PF02877	3	3	7.69E-08	1.34E-05
WGR	PF05406	3	3	7.69E-08	1.34E-05
UQ_con	PF00179	6	52	9.64E-08	1.68E-05
MutS_I	PF01624	3	5	7.64E-07	1.33E-04
Rep_fac_C	PF08542	3	5	7.64E-07	1.33E-04
HhH-GPD	PF00730	4	20	1.48E-06	2.58E-04
ResIII	PF04851	4	20	1.48E-06	2.58E-04
SNF2_N	PF00176	5	50	2.44E-06	4.25E-04
MutS_II	PF05188	3	7	2.66E-06	4.62E-04
MutS_III	PF05192	3	7	2.66E-06	4.62E-04
Rad51	PF08423	3	7	2.66E-06	4.62E-04
AAA	PF00004	7	151	4.23E-06	7.35E-04
HHH	PF00633	3	8	4.24E-06	7.38E-04
ERCC4	PF02732	3	9	6.34E-06	1.10E-03
SMC_N	PF02463	3	9	6.34E-06	1.10E-03
MutS_V	PF00488	3	10	9.03E-06	1.57E-03
PARP	PF00644	3	10	9.03E-06	1.57E-03
SAP	PF02037	3	11	1.24E-05	2.15E-03
Ku	PF02735	2	2	1.82E-05	3.17E-03
Ku_C	PF03730	2	2	1.82E-05	3.17E-03
Ku_N	PF03731	2	2	1.82E-05	3.17E-03
PADR1	PF08063	2	2	1.82E-05	3.17E-03
zf-PARP	PF00645	2	3	5.45E-05	9.48E-03

Table ST5a: DDR domains enriched in *A. thaliana* (only those with a Bonferroni adjusted p-value < 0.01 are shown)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
AAA	PF00004	7	34	5.54E-06	7.81E-04
SMC_N	PF02463	4	7	6.49E-06	9.15E-04
Rad51	PF08423	3	3	9.38E-06	1.32E-03
Rep_fac_C	PF08542	3	3	9.38E-06	1.32E-03
FHA	PF00498	5	15	1.01E-05	1.43E-03
zf-C3HC4	PF00097	6	27	1.68E-05	2.37E-03
Pkinase	PF00069	11	116	3.06E-05	4.31E-03
MutS_I	PF01624	3	4	3.69E-05	5.21E-03
MutS_II	PF05188	3	5	9.09E-05	1.28E-02
UQ_con	PF00179	4	14	1.65E-04	2.33E-02
MutS_III	PF05192	3	6	1.79E-04	2.52E-02
MutS_V	PF00488	3	6	1.79E-04	2.52E-02

Table ST5b: DDR domains enriched in *S. cerevisiae* (only those with a Bonferroni adjusted p-value < 0.05 are shown)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
HHH	PF00633	3	3	1.86E-06	1.47E-04
HhH-GPD	PF00730	3	3	1.86E-06	1.47E-04
SMC_N	PF02463	3	4	7.37E-06	5.82E-04
UvrD-helicase	PF00580	3	4	7.37E-06	5.82E-04
EndIII_4Fe-2S	PF10576	2	2	1.54E-04	1.22E-02
IMS	PF00817	2	2	1.54E-04	1.22E-02
IMS_C	PF11799	2	2	1.54E-04	1.22E-02
IMS_HHH	PF11798	2	2	1.54E-04	1.22E-02
UVR	PF02151	2	2	1.54E-04	1.22E-02
GIY-YIG	PF01541	2	3	4.59E-04	3.62E-02

Table ST5c: DDR domains enriched in *E. coli* (only those with a Bonferroni adjusted p-value < 0.05 are shown)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
AAA	PF00004	7	49	2.60E-09	3.64E-07
Rad17	PF03215	3	3	9.43E-08	1.32E-05
BRCT	PF00533	4	27	6.82E-06	9.54E-04
UQ_con	PF00179	4	28	7.92E-06	1.11E-03
DNA_ligase_A_M	PF01068	2	2	2.09E-05	2.93E-03
DNA_mis_repair	PF01119	2	2	2.09E-05	2.93E-03
Ku	PF02735	2	2	2.09E-05	2.93E-03
Ku_N	PF03731	2	2	2.09E-05	2.93E-03
MutS_I	PF01624	2	2	2.09E-05	2.93E-03
Rep_fac_C	PF08542	2	2	2.09E-05	2.93E-03
SMC_N	PF02463	3	13	2.61E-05	3.65E-03
HHH	PF00633	2	3	6.26E-05	8.76E-03

Table ST5d: DDR domains enriched in *C. elegans* (only those with a Bonferroni adjusted p-value < 0.01 are shown)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
BRCT	PF00533	6	13	6.75E-10	1.10E-07
SMC_N	PF02463	4	7	1.97E-07	3.21E-05
ERCC4	PF02732	3	3	6.68E-07	1.09E-04
Rep_fac_C	PF08542	3	3	6.68E-07	1.09E-04
Rad17	PF03215	3	4	2.65E-06	4.33E-04
Pkinase	PF00069	10	174	3.11E-06	5.07E-04
AAA	PF00004	6	49	4.25E-06	6.92E-04
tRNA_anti	PF01336	3	8	3.62E-05	5.90E-03
ResIII	PF04851	3	10	7.66E-05	1.25E-02
DNA_mis_repair	PF01119	2	2	7.71E-05	1.26E-02
Ku	PF02735	2	2	7.71E-05	1.26E-02
Ku_N	PF03731	2	2	7.71E-05	1.26E-02
MutS_I	PF01624	2	2	7.71E-05	1.26E-02
MutS_II	PF05188	2	2	7.71E-05	1.26E-02
MutS_III	PF05192	2	2	7.71E-05	1.26E-02

MutS_V	PF00488	2	2	7.71E-05	1.26E-02
zf-MIZ	PF02891	2	2	7.71E-05	1.26E-02
UQ_con	PF00179	4	32	1.71E-04	2.78E-02
HECT	PF00632	3	14	2.26E-04	3.69E-02
DNA_ligase_A_C	PF04679	2	3	2.30E-04	3.75E-02
DNA_ligase_A_M	PF01068	2	3	2.30E-04	3.75E-02
DNA_ligase_A_N	PF04675	2	3	2.30E-04	3.75E-02
HHH	PF00633	2	3	2.30E-04	3.75E-02

Table ST5e: DDR domains enriched in *D. melanogaster* (only those with a Bonferroni adjusted p-value < 0.05 are shown)

Domain	Pfam ID	DDR proteins with domain	Proteins with domain in proteome	Fisher p-value	Bonferroni adjusted p-value
Rad51	PF08423	3	3	1.38E-05	1.99E-03
Rep_fac_C	PF08542	3	3	1.38E-05	1.99E-03
AAA	PF00004	7	35	1.56E-05	2.25E-03
zf-C3HC4	PF00097	7	38	2.76E-05	3.97E-03
HhH-GPD	PF00730	3	4	5.42E-05	7.81E-03
MutS_I	PF01624	3	4	5.42E-05	7.81E-03
MutS_II	PF05188	3	4	5.42E-05	7.81E-03
MutS_III	PF05192	3	4	5.42E-05	7.81E-03
MutS_V	PF00488	3	4	5.42E-05	7.81E-03
SAP	PF02037	3	6	2.62E-04	3.77E-02
SMC_N	PF02463	3	6	2.62E-04	3.77E-02
UQ_con	PF00179	4	14	2.70E-04	3.89E-02

Table ST5f: DDR domains enriched in *S. pombe* (only those with a Bonferroni adjusted p-value < 0.05 are shown)

ST6. GO functional enrichment of 118 human DDR proteins

Category: BiologicalProcess (GOTERM_BP_FAT)

Term	Count	%	PValue	Genes	Pop Hits	Pop Total	Fold Enrichment	Bonferroni
GO:0006974~response to DNA damage stimulus	88	73.95	1.88E-113	DDB1, DNA2L, SLX1, FACD2, MRE11, MSH6, LIG4, RFA3, XPC, TDP2, SLX4, ATRIP, MRI40, FANCM, XRCC4, UBP11, CUL4, CHK1, RAD50, PARP2, EXO1, MDM2, SMC6, BARD1, RFA2, ERCC3, RAD23B, FBX31, MUS81, RAD9A, SMC5, ERCC5, SMC1A, XLF, CLSPN, RAD18, PMS2, XPF, TERF2, XRCC6, BLM, MLH1, NBN, PARP1, F175A, PRKDC, MSH3, TOPB1, ATM, BRCC3, BRCA1, RBX1, ERCC1, XPA, RAD1, ERCC8, MMS21, PCNA, TIPIN, DCR1C, H2AX, RFA1, SIR1, BRE, TP53B, RN168, RBBP8, TDP1, UBE2N, XRCC5, TIM, CHK2, UBR5, KAT5, DCR1B, RNF8, ATR, DTL, HUS1, ERCC2, ERCC6, XRCC1, BRCA2, MSH2, EME1, RAD51, RAD17	373	13528	27.047394	1.77E-110
GO:0006281~DNA repair	81	68.067	7.39E-110	DDB1, DNA2L, SLX1, MRE11, FACD2, MSH6, LIG4, RFA3, XPC, TDP2, SLX4, ATRIP, MRI40, FANCM, XRCC4, CUL4, CHK1, RAD50, PARP2, EXO1, MDM2, SMC6, BARD1, RFA2, ERCC3, RAD23B, MUS81, RAD9A, SMC5, ERCC5, SMC1A, XLF, CLSPN, RAD18, PMS2, XPF, XRCC6, BLM, MLH1, NBN, PARP1, F175A, PRKDC, MSH3, TOPB1, ATM, BRCC3, BRCA1, RBX1, ERCC1, XPA, RAD1, ERCC8, MMS21, PCNA, DCR1C, H2AX, RFA1, SIR1, BRE, TP53B, RN168, RBBP8, TDP1, UBE2N, XRCC5, UBR5, KAT5, DCR1B, RNF8, ATR, HUS1, ERCC2, ERCC6, XRCC1, BRCA2, MSH2, EME1, RAD51, RAD17	284	13528	32.69778	6.97E-107
GO:0006259~DNA metabolic process	90	75.63	1.13E-104	DDB1, DNA2L, SLX1, FACD2, MRE11, MSH6, LIG4, TOP3A, RFA3, XPC, TDP2, SLX4, ATRIP, MRI40, FANCM, XRCC4, MPIP1, CUL4, MPIP3, CHK1, RAD50, PARP2, EXO1, MDM2, SMC6, BARD1, RFA2, SMAL1, ERCC3, RAD23B, MUS81, RAD9A, SMC5, ERCC5, SMC1A, XLF, CLSPN, RAD18, PMS2, XPF, TERF2, XRCC6, BLM, MLH1, NBN, PARP1, F175A, PRKDC, MSH3, TOPB1, ATM, RMI1, BRCC3, BRCA1, RBX1, ERCC1, XPA, RAD1, ERCC8, MMS21, PCNA, TIPIN, DCR1C, H2AX, RFA1, SIR1, BRE, TP53B, RN168, RBBP8, TDP1, UBE2N, XRCC5, UBR5, KAT5, DCR1B, RNF8, ATR, CDT1, DTL, HUS1, ERCC2, ERCC6, XRCC1, BRCA2, MSH2, EME1, RAD51, RAD17	506	13528	20.391237	1.06E-101
GO:0033554~cellular response to stress	88	73.95	3.52E-96	DDB1, DNA2L, SLX1, FACD2, MRE11, MSH6, LIG4, RFA3, XPC, TDP2, SLX4, ATRIP, MRI40, FANCM, XRCC4, UBP11, CUL4, CHK1, RAD50, PARP2, EXO1, MDM2, SMC6, BARD1, RFA2, ERCC3, RAD23B, FBX31, MUS81, RAD9A, SMC5, ERCC5, SMC1A, XLF, CLSPN, RAD18, PMS2, XPF, TERF2, XRCC6, BLM, MLH1, NBN, PARP1, F175A, PRKDC, MSH3, TOPB1, ATM, BRCC3, BRCA1, RBX1, ERCC1, XPA, RAD1, ERCC8, MMS21, PCNA, TIPIN, DCR1C, H2AX, RFA1, SIR1, BRE, TP53B, RN168, RBBP8, TDP1, UBE2N, XRCC5, TIM, CHK2, UBR5, KAT5, DCR1B, RNF8, ATR, DTL, HUS1, ERCC2, ERCC6, XRCC1, BRCA2, MSH2, EME1, RAD51, RAD17	566	13528	17.824519	3.32E-93
GO:0006302~double-strand break repair	34	28.571	1.99E-53	SLX1, XPF, MRE11, XRCC6, LIG4, BLM, NBN, MLH1, TDP2, F175A, PRKDC, SLX4, BRCC3, MRI40, BRCA1, ERCC1, XRCC4, DCR1C, H2AX, RFA1, BRE, RAD50, TDP1, RN168, UBE2N, XRCC5, KAT5, UBR5, RNF8, HUS1, XLF, BRCA2, MSH2, RAD51	62	13528	62.869328	1.88E-50
GO:0006310~DNA recombination	34	28.571	7.54E-44	PMS2, SLX1, XPF, MRE11, MSH6, XRCC6, LIG4, BLM, NBN, MLH1, PRKDC, SLX4, MSH3, ATM, BRCA1, ERCC1, XRCC4, MMS21, DCR1C, H2AX, RFA1, CHK1, RAD50, EXO1, UBE2N, XRCC5, SMC6, HUS1, MUS81, SMC5, BRCA2, EME1, MSH2, RAD51	105	13528	37.122841	7.11E-41
GO:0042770~DNA damage response, signal transduction	30	25.21	1.29E-40	MSH6, BLM, NBN, MLH1, XPC, F175A, ATRIP, ATM, BRCC3, MRI40, BRCA1, RAD1, XPA, TIPIN, H2AX, CHK1, BRE, MDM2, CHK2, KAT5, UBR5, ATR, FBX31, ERCC6, HUS1, RAD9A, SMC1A, BRCA2, MSH2, RAD17	80	13528	42.991525	1.22E-37
GO:0051052~regulation of DNA metabolic process	32	26.891	6.67E-39	DNA2L, TERF2, XPF, MRE11, MSH6, BLM, NBN, MLH1, F175A, MSH3, BRCC3, MRI40, BRCA1, ERCC1, ERCC8, TIPIN, PCNA, H2AX, BRE, RAD50, RN168, UBE2N, UBR5, RNF8, CDT1, ATR, HUS1, RAD9A, BRCA2, MSH2, RAD51, RAD17	114	13528	32.180791	6.29E-36
GO:0000075~cell cycle checkpoint	30	25.21	1.12E-38	DDB1, BLM, NBN, XPC, F175A, ATRIP, ATM, BRCC3, MRI40, BRCA1, RAD1, TIPIN, H2AX, CHK1, BRE, RBBP8, MDM2, CHK2, UBR5, ERCC3, FBX31, CDT1, ATR, HUS1, ERCC2, RAD9A, SMC1A, MSH2, RAD17	91	13528	37.794748	1.05E-35
GO:0031570~DNA integrity checkpoint	24	20.168	1.03E-34	CHK1, BRE, BLM, NBN, MDM2, CHK2, UBR5, XPC, F175A, ATRIP, ATM, ATR, FBX31, BRCC3, CDT1, HUS1, MRI40, BRCA1, RAD9A, RAD1, MSH2, TIPIN, RAD17, H2AX	52	13528	52.912647	9.69E-32
GO:0000077~DNA damage checkpoint	23	19.328	1.16E-33	CHK1, BRE, BLM, NBN, MDM2, CHK2, UBR5, XPC, F175A, ATRIP, ATM, ATR, FBX31, BRCC3, HUS1, MRI40, BRCA1, RAD9A, RAD1, MSH2, TIPIN, RAD17, H2AX	48	13528	54.933616	1.09E-30
GO:0010212~response to ionizing radiation	24	20.168	6.44E-33	BRE, FACD2, XRCC6, LIG4, BLM, RN168, UBR5, F175A, PRKDC, TOPB1, ATM, RNF8, BRCC3, ERCC6, MRI40, BRCA1, XRCC4, ERCC1, XLF, BRCA2, MSH2, ERCC8, DCR1C, H2AX	60	13528	45.857627	6.07E-30
GO:0009314~response to radiation	33	27.731	3.77E-32	FACD2, XPF, MSH6, XRCC6, LIG4, BLM, XPC, F175A, PRKDC, ATM, TOPB1, BRCC3, MRI40, BRCA1, ERCC1, XRCC4, XPA, ERCC8, DCR1C, H2AX, BRE, RN168, UBR5, ERCC3, RNF8, ERCC6, HUS1, ERCC2, ERCC5, SMC1A, XLF, BRCA2, MSH2	200	13528	18.916271	3.55E-29
GO:0051726~regulation of cell cycle	36	30.252	8.43E-29	DDB1, BLM, NBN, XPC, F175A, ATRIP, ATM, BRCC3, MDM4, MRI40, BRCA1, RAD1, MPIP1, HERC2, TIPIN, MPIP3, H2AX, CHK1, BRE, RBBP8, TIM, CHK2, MDM2, UBR5, ERCC3, CDT1, FBX31, ATR, HUS1, ERCC2, RAD9A, SMC1A, BRCA2, MSH2, RAD17	331	13528	12.468841	7.95E-26
GO:0006260~DNA replication	29	24.37	4.60E-27	DNA2L, MRE11, TERF2, LIG4, BLM, RFA3, TOP3A, ATRIP, RMI1, BRCA1, RAD1, MPIP1, TIPIN, PCNA, MPIP3, RFA1, SIR1, CHK1, RAD50, RFA2, ATR, CDT1, DTL, HUS1, RAD9A, CLSPN, BRCA2, RAD51, RAD17	190	13528	17.498305	4.33E-24

ST6. GO functional enrichment of 118 human DDR proteins

Category: CellularComponent (GOTERM_CC_FAT)

Term	Count	%	PValue	Genes	Pop Hits	Pop Total	Fold Enrichment	Bonferroni
GO:0005654~nucleoplasm	59	49.58	2.54E-42	DDB1, DNA2L, MRE11, FACD2, TOP3A, RFA3, PAXI1, XPC, TDP2, ATRIP, PLK1, MDM4, CUL1, MPIP1, HNRPK, MPIP3, CHK1, PARP2, MDM2, RFA2, ERCC3, RAD23B, ERCC5, CLSPN, PIAS1, TERF2, XPF, XRCC6, BLM, NBN, PARP1, PRKDC, ATM, TOPB1, BRCA1, SKP1, ERCC1, XPA, ERCC8, PCNA, TIF1B, H2AX, SIR1, RFA1, TP53B, XRCC5, CHK2, MDC1, KAT5, CDT1, ATR, ERCC6, ERCC2, MTA2, BRCA2, XRCC1, RAD51, WEE1	882	12782	8.6366843	4.14E-40
GO:0031981~nuclear lumen	64	53.782	6.74E-36	DDB1, DNA2L, MRE11, FACD2, TOP3A, PAXI1, RFA3, XPC, TDP2, ATRIP, PLK1, MDM4, CUL1, MPIP1, HNRPK, MPIP3, CHK1, PARP2, PIAS4, MDM2, RFA2, ERCC3, RAD23B, RAD9A, MUS81, ERCC5, CLSPN, PIAS1, TERF2, XPF, XRCC6, BLM, NBN, PARP1, PRKDC, TOPB1, ATM, BRCA1, SKP1, ERCC1, XPA, ERCC8, PCNA, TIF1B, H2AX, RFA1, SIR1, TP53B, XRCC5, CHK2, MDC1, KAT5, CDT1, ATR, ERCC6, ERCC2, MTA2, XRCC1, BRCA2, EME1, RAD51, WEE1, RAD17	1450	12782	5.6986973	1.10E-33
GO:0070013~intracellular organelle lumen	64	53.782	1.22E-30	DDB1, DNA2L, MRE11, FACD2, TOP3A, PAXI1, RFA3, XPC, TDP2, ATRIP, PLK1, MDM4, CUL1, MPIP1, HNRPK, MPIP3, CHK1, PARP2, PIAS4, MDM2, RFA2, ERCC3, RAD23B, RAD9A, MUS81, ERCC5, CLSPN, PIAS1, TERF2, XPF, XRCC6, BLM, NBN, PARP1, PRKDC, TOPB1, ATM, BRCA1, SKP1, ERCC1, XPA, ERCC8, PCNA, TIF1B, H2AX, RFA1, SIR1, TP53B, XRCC5, CHK2, MDC1, KAT5, CDT1, ATR, ERCC6, ERCC2, MTA2, XRCC1, BRCA2, EME1, RAD51, WEE1, RAD17	1779	12782	4.6448067	1.99E-28
GO:0043233~organelle lumen	64	53.782	4.61E-30	DDB1, DNA2L, MRE11, FACD2, TOP3A, PAXI1, RFA3, XPC, TDP2, ATRIP, PLK1, MDM4, CUL1, MPIP1, HNRPK, MPIP3, CHK1, PARP2, PIAS4, MDM2, RFA2, ERCC3, RAD23B, RAD9A, MUS81, ERCC5, CLSPN, PIAS1, TERF2, XPF, XRCC6, BLM, NBN, PARP1, PRKDC, TOPB1, ATM, BRCA1, SKP1, ERCC1, XPA, ERCC8, PCNA, TIF1B, H2AX, RFA1, SIR1, TP53B, XRCC5, CHK2, MDC1, KAT5, CDT1, ATR, ERCC6, ERCC2, MTA2, XRCC1, BRCA2, EME1, RAD51, WEE1, RAD17	1820	12782	4.5401709	7.51E-28
GO:0005694~chromosome	39	32.773	8.54E-30	SLX1, TERF2, XPF, FACD2, MSH6, XRCC6, BLM, LIG4, NBN, RFA3, MLH1, TOP3A, SLX4, MYST1, TOPB1, BRCA1, ERCC1, XRCC4, TIPIN, PCNA, TIF1B, H2AX, RFA1, CHK1, RAD50, TP53B, RN168, XRCC5, TIM, KAT5, SMC6, RFA2, RNF8, ATR, SMC5, SMC1A, CLSPN, RAD51, RAD18	460	12782	10.946377	1.39E-27
GO:0031974~membrane-enclosed lumen	64	53.782	1.44E-29	DDB1, DNA2L, MRE11, FACD2, TOP3A, PAXI1, RFA3, XPC, TDP2, ATRIP, PLK1, MDM4, CUL1, MPIP1, HNRPK, MPIP3, CHK1, PARP2, PIAS4, MDM2, RFA2, ERCC3, RAD23B, RAD9A, MUS81, ERCC5, CLSPN, PIAS1, TERF2, XPF, XRCC6, BLM, NBN, PARP1, PRKDC, TOPB1, ATM, BRCA1, SKP1, ERCC1, XPA, ERCC8, PCNA, TIF1B, H2AX, RFA1, SIR1, TP53B, XRCC5, CHK2, MDC1, KAT5, CDT1, ATR, ERCC6, ERCC2, MTA2, XRCC1, BRCA2, EME1, RAD51, WEE1, RAD17	1856	12782	4.4521073	2.35E-27
GO:0000228~nuclear chromosome	24	20.168	2.97E-23	RFA1, CHK1, SLX1, TERF2, XPF, MSH6, RAD50, XRCC6, BLM, MLH1, TIM, XRCC5, RFA3, NBN, RFA2, SLX4, TOPB1, ATR, SMC1A, ERCC1, RAD51, TIPIN, TIF1B, H2AX	162	12782	19.127572	4.84E-21
GO:0044454~nuclear chromosome part	20	16.807	4.34E-20	RFA1, SLX1, XPF, TERF2, MSH6, RAD50, XRCC6, BLM, MLH1, TIM, RFA3, NBN, XRCC5, RFA2, SLX4, ATR, ERCC1, TIPIN, TIF1B, H2AX	122	12782	21.165756	7.08E-18
GO:0044427~chromosomal part	27	22.689	6.32E-18	SLX1, XPF, TERF2, MSH6, XRCC6, BLM, NBN, RFA3, MLH1, SLX4, MYST1, ERCC1, TIPIN, PCNA, TIF1B, H2AX, RFA1, CHK1, RAD50, TP53B, TIM, XRCC5, KAT5, RFA2, ATR, SMC1A, RAD18	386	12782	9.0310881	1.03E-15
GO:0043232~intracellular non-membrane-bounded organelle	56	47.059	5.43E-15	SLX1, MRE11, FACD2, MSH6, LIG4, RFA3, TOP3A, SLX4, MYST1, PLK1, MDM4, XRCC4, HNRPK, CHK1, RAD50, PARP2, MDM2, SMC6, RFA2, RAD9A, MUS81, SMC5, SMC1A, CLSPN, RAD18, TERF2, XPF, XRCC6, BLM, MLH1, NBN, PARP1, TOPB1, ATM, BRCA1, ERCC1, PCNA, TIPIN, TIF1B, H2AX, SIR1, RFA1, TP53B, RN168, XRCC5, TIM, MDC1, KAT5, ATR, RNF8, ERCC6, MTA2, BRCA2, EME1, RAD51, RAD17	2596	12782	2.7851395	8.87E-13
GO:0043228~non-membrane-bounded organelle	56	47.059	5.43E-15	SLX1, MRE11, FACD2, MSH6, LIG4, RFA3, TOP3A, SLX4, MYST1, PLK1, MDM4, XRCC4, HNRPK, CHK1, RAD50, PARP2, MDM2, SMC6, RFA2, RAD9A, MUS81, SMC5, SMC1A, CLSPN, RAD18, TERF2, XPF, XRCC6, BLM, MLH1, NBN, PARP1, TOPB1, ATM, BRCA1, ERCC1, PCNA, TIPIN, TIF1B, H2AX, SIR1, RFA1, TP53B, RN168, XRCC5, TIM, MDC1, KAT5, ATR, RNF8, ERCC6, MTA2, BRCA2, EME1, RAD51, RAD17	2596	12782	2.7851395	8.87E-13
GO:0070531~BRCA1-A complex	7	5.8824	1.21E-12	BRE, BRCC3, MRI40, BRCA1, UBR5, F175A, BARD1	7	12782	129.11111	1.97E-10
GO:0005657~replication fork	10	8.4034	1.53E-12	RFA1, CHK1, TP53B, BLM, NBN, RFA3, PCNA, RFA2, RAD18, H2AX	32	12782	40.347222	2.49E-10
GO:0000781~chromosome, telomeric region	9	7.563	3.36E-11	XPF, TERF2, RAD50, XRCC6, BLM, ERCC1, NBN, XRCC5, SLX4	29	12782	40.068966	5.47E-09
GO:0000151~ubiquitin ligase complex	12	10.084	7.65E-11	BRE, BRCC3, FBX31, RNF8, BRCA1, SKP1, RBX1, RN168, CUL1, CUL4, HERC2, BARD1	90	12782	17.214815	1.25E-08

ST6. GO functional enrichment of 118 human DDR proteins

Category: MolecularFunction (GOTERM_MF_FAT)

Term	Count	%	PValue	Genes	Pop Hits	Pop Total	Fold Enrichment	Bonferroni
GO:0003684~damaged DNA binding	19	15.966	9.91E-26	DDB1, XPF, MSH6, TP53B, NBN, XPC, MSH3, ERCC3, RAD23B, BRCA1, ERCC1, RAD1, XPA, XRCC1, MSH2, RAD51, RAD18, H2AX	50	12983	45.680926	2.30E-23
GO:0043566~structure-specific DNA binding	24	20.168	1.32E-23	RFA1, PMS2, TERF2, MRE11, XPF, MSH6, XRCC6, EXO1, TDP1, BLM, MLH1, XRCC5, RFA3, XPC, RFA2, MSH3, RAD23B, ERCC5, ERCC1, BRCA2, MSH2, RAD51, PCNA, RAD18	145	12983	19.897318	3.07E-21
GO:0003697~single-stranded DNA binding	16	13.445	1.77E-19	RFA1, PMS2, XPF, BLM, TDP1, MLH1, RFA3, XPC, RFA2, MSH3, RAD23B, ERCC5, ERCC1, BRCA2, MSH2, RAD51	55	12983	34.971044	4.11E-17
GO:0003677~DNA binding	61	51.261	6.03E-19	DDB1, DNA2L, MRE11, MSH6, LIG4, RFA3, TOP3A, XPC, FANCM, XRCC4, HNRPK, RAD50, PARP2, EXO1, PIAS4, RFA2, SMAL1, ERCC3, RAD23B, MUS81, ERCC5, CLSPN, XLF, PIAS1, RAD18, PMS2, TERF2, XPF, XRCC6, BLM, NR4A2, MLH1, NBN, PARP1, PRKDC, MSH3, ATM, TOPB1, BRCA1, ERCC1, XPA, RAD1, SOX4, PCNA, TIF1B, H2AX, RFA1, TP53B, TDP1, XRCC5, CDT1, ATR, ERCC6, ERCC2, MTA2, BRCA2, XRCC1, EME1, MSH2, RAD51	2331	12983	3.1458562	1.40E-16
GO:0004536~deoxyribonuclease activity	12	10.084	1.70E-15	DNA2L, SLX1, XPF, MRE11, MUS81, RAD50, RAD9A, ERCC5, EXO1, ERCC1, RAD1, SLX4	34	12983	42.428105	3.86E-13
GO:0004520~endodeoxyribonuclease activity	10	8.4034	5.67E-14	DNA2L, SLX1, XPF, MRE11, MUS81, RAD50, ERCC5, EXO1, ERCC1, SLX4	22	12983	54.642256	1.31E-11
GO:0004518~nuclease activity	17	14.286	1.80E-13	DNA2L, SLX1, PMS2, MRE11, XPF, RAD50, EXO1, TDP1, SLX4, RAD9A, MUS81, ERCC5, FANCM, ERCC1, RAD1, EME1, DCR1C	158	12983	12.934306	4.17E-11
GO:0008022~protein C-terminus binding	16	13.445	5.22E-13	SIR1, MRE11, XPF, TERF2, XRCC6, LIG4, XRCC5, MDM2, ERCC3, TOPB1, ERCC6, ERCC2, XRCC4, ERCC1, MSH2, RAD51	141	12983	13.641187	1.21E-10
GO:0003690~double-stranded DNA binding	14	11.765	1.01E-12	PMS2, MRE11, TERF2, MSH6, XRCC6, BLM, TDP1, MLH1, XRCC5, MSH3, ERCC5, MSH2, RAD51, PCNA	97	12983	17.350325	2.33E-10
GO:0004519~endonuclease activity	13	10.924	3.01E-11	DNA2L, SLX1, PMS2, MRE11, XPF, RAD50, EXO1, SLX4, MUS81, ERCC5, ERCC1, EME1, DCR1C	100	12983	15.627685	6.97E-09
GO:0008094~DNA-dependent ATPase activity	10	8.4034	8.07E-10	DNA2L, ERCC6, ERCC2, XRCC6, BLM, XRCC5, RAD51, ERCC8, SMAL1, ERCC3	57	12983	21.089994	1.87E-07
GO:0019787~small conjugating protein ligase activity	14	11.765	9.75E-10	UBE2T, RN168, UBE2N, PIAS4, TRIPC, BARD1, RNF8, MDM4, BRCA1, RBX1, UBP11, HERC2, ERCC8, PIAS1	166	12983	10.138443	2.26E-07
GO:0003678~DNA helicase activity	9	7.563	1.01E-09	DNA2L, ERCC6, ERCC2, XRCC6, BLM, XRCC5, ERCC8, SMAL1, ERCC3	40	12983	27.047917	2.34E-07
GO:0032404~mismatch repair complex binding	6	5.042	1.90E-09	PMS2, ATR, MSH6, MLH1, MSH2, PCNA	8	12983	90.159722	4.41E-07
GO:0016881~acid-amino acid ligase activity	14	11.765	1.01E-08	UBE2T, RN168, UBE2N, PIAS4, TRIPC, BARD1, RNF8, MDM4, BRCA1, RBX1, UBP11, HERC2, ERCC8, PIAS1	201	12983	8.3730422	2.34E-06

ST7. Classification of DDR proteins according to their role

Protein	UniProt ID	Ensemble Gene ID	Category	Pathway	Emergence in evolution	PMID reference
1433E	P62258	ENSG00000108953	Mediator	Checkpoint	Ancient Eukaryotes	21945648
1433E	P62258	ENSG00000108953	Effector	Checkpoint	Ancient Eukaryotes	21945648
ATM	Q13315	ENSG00000149311	Sensor	DSB repair	Plants	20674189
ATM	Q13315	ENSG00000149311	Transducer	DSB repair	Plants	21363960
ATR	Q13535	ENSG00000175054	Sensor	Replication stress	Ancient Eukaryotes	20674189
ATR	Q13535	ENSG00000175054	Transducer	Replication stress	Ancient Eukaryotes	21363960
ATRIP	Q8WXE1	ENSG00000164053	Sensor	Replication stress	Plants	21211780
ATRIP	Q8WXE1	ENSG00000164053	Mediator	Replication stress	Plants	20947357
BARD1	Q99728	ENSG00000138376	Mediator	DSB repair	Plants	19261746
BLM	P54132	ENSG00000197299	Effector	HR	Bacteria	21325134
BRCA1	P38398	ENSG00000012048	Sensor	DSB repair	Ancient Eukaryotes	21203981
BRCA1	P38398	ENSG00000012048	Mediator	DSB repair	Ancient Eukaryotes	21363960
BRCA2	P51587	ENSG00000139618	Mediator	DSB repair	Ancient Eukaryotes	21203981
BRCC3/BRCC36	P46736	ENSG00000185515	Mediator	DSB repair	Ancient Eukaryotes	19261746
BRE/BRCC45	Q9NXR7	ENSG00000158019	Mediator	DSB repair	Ancient Eukaryotes	19261746
CDT1	Q9H211	ENSG00000167513	Effector	Checkpoint	Ancient Eukaryotes	18082599
CHK1	O14757	ENSG00000149554	Transducer	Checkpoint	Fungi	12781359
CHK2	O96017	ENSG00000183765	Transducer	Checkpoint	Unikonta	12781359
CLSPN	Q9HAW4	ENSG00000092853	Sensor	Replication stress	Fungi	21633183
CLSPN	Q9HAW4	ENSG00000092853	Mediator	Replication stress	Fungi	21363960
CUL1	Q13616	ENSG00000055130	Effector	Checkpoint	Ancient Eukaryotes	19231300
CUL4	Q13619	ENSG00000139842	Effector	Checkpoint, NER	Ancient Eukaryotes	19231300
DCR1B/Apollo	Q9H816	ENSG00000118655	Effector	ICL, checkpoint	Ancient Eukaryotes	18469862
DCR1C/Artemis	Q96SD1	ENSG00000152457	Effector	NHEJ	Ancient Eukaryotes	20543526
DDB1	Q16531	ENSG00000167986	Mediator	Checkpoint, NER	Ancient Eukaryotes	18082599
DNA2L	P51530	ENSG00000138346	Effector	HR	Plants	21325134
DTL/CDT2	Q9NZJ0	ENSG00000143476	Mediator	Checkpoint	Ancient Eukaryotes	18082599
EME1	Q96AY2	ENSG00000154920	Effector	HR	Plants	21859861
ERCC1	P07992	ENSG00000012061	Effector	NER	Ancient Eukaryotes	18166977
ERCC2/XPD	P18074	ENSG00000104884	Effector	NER	Ancient Eukaryotes	18166977
ERCC3/XPB	P19447	ENSG00000163161	Effector	NER	Bacteria	18166977
ERCC5/XPG	P28715	ENSG00000134899	Effector	NER	Ancient Eukaryotes	18166977
ERCC6/CSB	Q03468	ENSG00000032514	Sensor	NER	Ancient Eukaryotes	18166977
ERCC6/CSB	Q03468	ENSG00000032514	Mediator	NER	Ancient Eukaryotes	18166977
ERCC8/CSA	Q13216	ENSG00000049167	Effector	NER	Ancient Eukaryotes	18166977
EXO1	Q9UQ84	ENSG00000174371	Effector	HR, MMR, NER	Ancient Eukaryotes	14676842, 21808022, 22326273
F175A/Abraxas	Q6UWZ7	ENSG00000163322	Mediator	DSB repair	Vertebrata	19261746
FACD2	Q9BXW9	ENSG00000144554	Mediator	DSB repair	Ancient Eukaryotes	20676667
FACD2	Q9BXW9	ENSG00000144554	Effector	DSB repair	Ancient Eukaryotes	20676667
FANCM	Q8IYD8	ENSG00000187790	Sensor	FA pathway, NER	Archaea	21975120, 22615860
FBW1A/BTRCP	Q9Y297	ENSG00000166167	Effector	DSB repair	Fungi	22099186
FBX31	Q5XUX0	ENSG00000103264	Effector	Checkpoint	Metazoa	19412162
H2AX	P16104	ENSG00000188486	Mediator	DSB repair	Ancient Eukaryotes	20860841
HERC2	O95714	ENSG00000128731	Transducer	DSB repair	Metazoa	20023648
HERC2	O95714	ENSG00000128731	Mediator	DSB repair	Metazoa	20023648
HNRPK	P61978	ENSG00000165119	Effector	Checkpoint	Metazoa	19579069
HUS1	O60921	ENSG00000136273	Sensor	Replication stress	Ancient Eukaryotes	20860841
KAT5/TIP60	Q92993	ENSG00000172977	Effector	Checkpoint, DSB repair	Ancient Eukaryotes	17923702
LIG4	P49917	ENSG00000174405	Effector	NHEJ	Ancient Eukaryotes	21329706
MAPK2	P49137	ENSG00000162889	Transducer	Checkpoint	Fungi	19230643
MDC1	Q14676	ENSG00000206481	Sensor	DSB repair	Vertebrata	21326949
MDC1	Q14676	ENSG00000206481	Mediator	DSB repair	Vertebrata	21363960
MDM2	Q00987	ENSG00000135679	Effector	Checkpoint	Metazoa	21541195
MDM4/MDMX	O15151	ENSG00000198625	Effector	Checkpoint	Vertebrata	21541195
MK03/ERK1	P27361	ENSG00000102882	Transducer	Checkpoint	Ancient Eukaryotes	16186792
MLH1	P40692	ENSG00000076242	Sensor	MMR	Bacteria	16612326
MMS21	Q96MF7	ENSG00000156831	Transducer	HR	Ancient Eukaryotes	22369641
MPIP1/CDC25A	P30304	ENSG00000164045	Effector	Checkpoint	Ancient Eukaryotes	20860841
MPIP3/CDC25C	P30307	ENSG00000158402	Effector	Checkpoint	Fungi	20860841
MRE11	P49959	ENSG00000020922	Sensor	DSB repair	Ancient Eukaryotes	21363960
MRI40/BABAM1	Q9NWV8	ENSG00000105393	Mediator	DSB repair	Plants	19261748
MSH2	P43246	ENSG00000095002	Sensor	MMR, FA pathway	Ancient Eukaryotes	16612326, 21975120
MSH3	P20585	ENSG00000113318	Sensor	MMR, FA pathway	Bacteria	16612326, 21975120
MSH6	P52701	ENSG00000116062	Sensor	MMR, FA pathway	Bacteria	16612326, 21975120
MSH6	P52701	ENSG00000116062	Transducer	NHEJ	Bacteria	21075794
MTA2	O94776	ENSG00000149480	Sensor	Replication stress	Metazoa	20805320
MUS81	Q96NY9	ENSG00000172732	Effector	HR	Ancient Eukaryotes	21859861
MYST1	Q9H7Z6	ENSG00000103510	Effector	NHEJ	Ancient Eukaryotes	20479123
NBN/NBS1	O60934	ENSG00000104320	Transducer	DSB repair	Plants	18082599
NBN/NBS1	O60934	ENSG00000104320	Sensor	DSB repair	Plants	18082599
NBN/NBS1	O60934	ENSG00000104320	Mediator	DSB repair	Plants	21252998
NBN/NBS1	O60934	ENSG00000104320	Effector	DSB repair	Plants	21252998
NR4A2	P43354	ENSG00000153234	Mediator	Replication stress	Bilateria	21979916
PALB2	Q86YC2	ENSG00000083093	Mediator	BSB repair	Vertebrata	21203981
PARP1	P09874	ENSG00000143799	Sensor	BER	Ancient Eukaryotes	20965415
PARP2	Q9UGN5	ENSG00000129484	Sensor	BER	Ancient Eukaryotes	20965415
PAXI1/PTIP	Q6ZW49	ENSG00000157212	Mediator	DSB repair	Ancient Eukaryotes	21363960
PCNA	P12004	ENSG00000132646	Mediator	Replication stress, NER, BER	Archaea	17512402
PCNA	P12004	ENSG00000132646	Effector	Replication stress, NER, BER	Archaea	20068082
PIAS1	O75925	ENSG00000033800	Effector	DSB repair	Plants	20016603
PIAS4	Q8N2W9	ENSG00000105229	Effector	DSB repair	Fungi	20016603
PLK1	P53350	ENSG00000166851	Transducer	Checkpoint	Bacteria	20126263
PMS2	P54278	ENSG00000122512	Sensor	MMR	Ancient Eukaryotes	16612326
PMS2	P54278	ENSG00000122512	Mediator	MMR	Ancient Eukaryotes	16612326
PRKDC	P78527	ENSG00000121031	Transducer	NHEJ	Ancient Eukaryotes	21211780
RAD1	O60671	ENSG00000113456	Sensor	Replication stress	Ancient Eukaryotes	21363960
RAD17	O75943	ENSG00000152942	Sensor	Replication stress	Ancient Eukaryotes	20068082
RAD18	Q9NS91	ENSG00000070950	Transducer	DSB repair	Ancient Eukaryotes	19396164
RAD23B	P54727	ENSG00000119318	Sensor	NER	Ancient Eukaryotes	18166977
RAD50	Q92878	ENSG00000113522	Sensor	DSB repair	Archaea	20860841
RAD51	Q06609	ENSG00000051180	Sensor	DSB repair	Bacteria	21252998
RAD9A	Q99638	ENSG00000172613	Sensor	Replication stress	Ancient Eukaryotes	20860841
RBBP8/CTIP	Q99708	ENSG00000101773	Effector	DSB repair	Bilateria	20029420
RBX1	P62877	ENSG00000100387	Effector	Checkpoint	Ancient Eukaryotes	19231300

RFA1	P27694	ENSG00000132383	Sensor	Replication stress, DSB repair, NER	Ancient Eukaryotes	12791985
RFA2	P15927	ENSG00000117748	Sensor	Replication stress, DSB repair, NER	Ancient Eukaryotes	12791985
RFA2	P15927	ENSG00000117748	Mediator	Replication stress, DSB repair, NER	Ancient Eukaryotes	17531546
RFA3	P35244	ENSG00000106399	Sensor	Replication stress, DSB repair, NER	Ancient Eukaryotes	12791985
RFA3	P35244	ENSG00000106399	Mediator	Replication stress, DSB repair, NER	Ancient Eukaryotes	19843584
RMI1	Q9H9A7	ENSG00000178966	Transducer	HR	Plants	15775963
RMI1	Q9H9A7	ENSG00000178966	Mediator	HR	Plants	15775963
RN168	Q8IYW5	ENSG00000163961	Transducer	DSB repair	Chordata	19875294
RN168	Q8IYW5	ENSG00000163961	Mediator	DSB repair	Chordata	19203579
RNF8	O76064	ENSG00000112130	Transducer	DSB repair	Ancient Eukaryotes	18550271
RNF8	O76064	ENSG00000112130	Mediator	DSB repair	Ancient Eukaryotes	18001825
SIR1	Q96EB6	ENSG00000096717	Transducer	NHEJ, NER	Unikonta	20097625, 20670893
SKP1	P63208	ENSG00000113558	Effector	Checkpoint	Ancient Eukaryotes	22099186
SLX1	Q9BQ83	ENSG00000181625	Effector	HR,MMR	Ancient Eukaryotes	19596236
SLX4	Q8IY92	ENSG00000188827	Mediator	DSB repair	Metazoa	19596236
SLX4	Q8IY92	ENSG00000188827	Effector	DSB repair, ICL	Metazoa	19596236
SMAL1	Q9NZC9	ENSG00000138375	Effector	Replication stress	Ancient Eukaryotes	19841479
SMC1A	Q14683	ENSG00000072501	Effector	DSB repair	Bacteria	19842212
SMC5	Q8IY18	ENSG00000198887	Mediator	HR	Ancient Eukaryotes	16810316
SMC6	Q96SB8	ENSG00000163029	Mediator	HR	Ancient Eukaryotes	16810316
SOX4	Q06945	ENSG00000124766	Effector	Checkpoint	Chordata	19234109
TAOK1	Q7L7X3	ENSG00000160551	Transducer	Checkpoint	Bacteria	18082599
TDP1	Q9NUW8	ENSG00000042088	Effector	Adducts removal	Ancient Eukaryotes	16141202
TDP2	O95551	ENSG00000111802	Effector	Adducts removal	Ancient Eukaryotes	22740648
TERF2	Q15554	ENSG00000132604	Mediator	Telomere maintenance	Vertebrata	19287395
TIF1B/KAP1	Q13263	ENSG00000130726	Transducer	DSB repair	Bilateria	18082607
TIF1B/KAP1	Q13263	ENSG00000130726	Mediator	DSB repair, Checkpoint	Bilateria	17056014
TIM	Q9UNS1	ENSG00000111602	Mediator	Replication stress, DSB repair, Circadian Clock	Ancient Eukaryotes	20068082
TIM	Q9UNS1	ENSG00000111602	Effector	Replication stress, DSB repair, Circadian Clock	Ancient Eukaryotes	17296725
TIPIN	Q9BVW5	ENSG00000075131	Mediator	Replication stress	Plants	20068082
TOP3A	Q13472	ENSG00000177302	Effector	HR	Bacteria	16595695
TOPB1	Q92547	ENSG00000163781	Mediator	Replication stress	Ancient Eukaryotes	21363960
TP53B	Q12888	ENSG00000067369	Sensor	DSB repair	Metazoa	21633183
TP53B	Q12888	ENSG00000067369	Mediator	DSB repair	Metazoa	20724228
TRIPC	Q14669	ENSG00000153827	Transducer	DSB repair	Ancient Eukaryotes	17525332
UBE2N/UBC13	P61088	ENSG00000177889	Transducer	DSB repair	Bacteria	18082599
UBE2T	Q9NPD8	ENSG00000077152	Transducer	FA pathway, NER	Bacteria	22615860
UBP11/USP11	P51784	ENSG00000102226	Effector	DSB repair	Ancient Eukaryotes	15314155
UBR5	O95071	ENSG00000104517	Transducer	Replication stress	Metazoa	11714696
UIMC1/Rap80	Q96RL1	ENSG00000087206	Mediator	DSB repair	Vertebrata	19261746
WEE1	P30291	ENSG00000166483	Transducer	Checkpoint	Ancient Eukaryotes	21859861
WEE1	P30291	ENSG00000166483	Effector	Checkpoint	Ancient Eukaryotes	19230643
XLF/NHEJ1	Q9H9Q4	ENSG00000187736	Effector	NHEJ	Fungi	16439205
XPA	P23025	ENSG00000136936	Effector	NER	Ancient Eukaryotes	18166977
XPC	Q01831	ENSG00000154767	Sensor	NER	Ancient Eukaryotes	18166977
XPF	Q92889	ENSG00000175595	Effector	NER	Ancient Eukaryotes	18166977
XRCC1	P18887	ENSG00000073050	Mediator	BER	Ancient Eukaryotes	19497792
XRCC4	Q13426	ENSG00000152422	Mediator	NHEJ	Plants	17241822
XRCC5/Ku80	P13010	ENSG00000079246	Sensor	NHEJ	Ancient Eukaryotes	21211780
XRCC6/Ku70	P12956	ENSG00000196419	Sensor	NHEJ	Ancient Eukaryotes	21211780

ST8. PTMs emergence

Name	ID	PTM in Hsa	Residue	PTM by	PMID	From
ATM	Q13315	phosphorylation	S367	ATM (auto)	16858402	Mdo
ATM	Q13316	phosphorylation	S1893	ATM (auto)	16858402	Xtr
ATM	Q13317	phosphorylation	S1981	ATM (auto)	16858402	Ppa (plants)/Nve
ATM	Q13317	acetylation	K3016	KAT5	17923702	Ppa
BRCA1	P38398	phosphorylation	S1524	ATM	17525332	Ddi/Nve/Mdo
BRCA1	P38398	phosphorylation	S1387	ATM and ATR	11114888	Osa/Tad/Cin
BRCA1	P38398	phosphorylation	S1423	ATM and ATR	11114888	Gga
BRCA1	P38398	phosphorylation	S1143	ATR	11114888	Hsa
BRCA1	P38398	phosphorylation	S1280	ATR	11114888	Gga
BRCA1	P38398	phosphorylation	T1394	ATR	11114888	Ddi/Xtr
BRCA1	P38398	phosphorylation	S1457	ATR	11114888	Mus
BRCA1	P38398	phosphorylation	S988	CHK2	20364141	Xtr
BRCA1	P38398	sumoylation	K109	PIAS1	20016594	Gga
BRCA1	P38398	sumoylation	K119	PIAS1	20016594	Cin
BRCA1	P38398	ubiquitination	N/A	UBE2T	19887602	N/A
BRCA2	P51587	phosphorylation	T3387	CHK1 and CHK2	18317453	Hsa
BRCA2	P51587	phosphorylation	S683	ATM or ATR	17525332	Mdo
BRCA2	P51587	phosphorylation	S755	ATM or ATR	17525332	Oan
CDT1	Q9H211	ubiquitination	QXRVTDF-motif	CUL4	16482215	Xtr
CHK1	O14757	phosphorylation	S317	ATR	11390642	Ecu
CHK1	O14757	phosphorylation	S345	ATR	11390642	Spo/Tad
CHK2	O96017	phosphorylation	T68	ATM	16481012	Ddi/Spo/Cte
CLSPN	Q9HAW4	phosphorylation	T916	CHK1 ? (1)	16963448/19556879	Mdo
CLSPN	Q9HAW4	phosphorylation	S945	ATR ? (1)	18331829/19556879	Tca
CLSPN	Q9HAW4	phosphorylation	S30	PLK1 (2)	16885022	Cel
CLSPN	Q9HAW4	ubiquitination	N/A	CUL1 - RBX1	19231300	N/A
DCR1B (Apollo)	Q9H816	phosphorylation	S444	ATM or ATR	17525332	Hsa
DCR1C (Artemis)	Q96SD1	phosphorylation	S645	ATM	15723659	Ngr/Bde/Bfl
DCR1C (Artemis)	Q96SD1	phosphorylation	N/A	PRKDC	15456891	N/A
EXO1	Q9UQ84	phosphorylation	S714	ATR	18048416	Hsa
F175A (Abraxas)	Q6UWZ7	phosphorylation	S406	ATM or ATR	17525340	Dre
FACD2	Q9BXW9	phosphorylation	S222	ATM	12086603	Ngr/Mbr/Tca
FACD2	Q9BXW9	phosphorylation	S1404	ATM	12086603	Osa/Cin/Mdo
FACD2	Q9BXW9	ubiquitination	K561	FANCL+UBE2T	11239454	Ehu
FBX31	Q5XUX0	phosphorylation	S278	ATM	19412162	Dre
H2AX	P16104	phosphorylation	S140	ATM and PRKDC	9488723/14627815	Ath
H2AX	P16104	ubiquitination	K119	UBE2N (UBC13)	19230794	Ehu
H2AX	P16104	ubiquitination	K120	RN168 ? (3)	19230794	Ehu
H2AX	P16104	ubiquitination	K120	RNF8 ? (3)	19230794	Ehu
H2AX	P16104	acetylation	K6	KAT5 (Tip60) ?	20703100	Ptr
H2AX	P16104	acetylation	K10	KAT5 (Tip60) ?	20703100	Ath
HERC2	O95715	phosphorylation	T4827	After IR (ATM ?)	20023648	Dre
HNRPK	P61978	ubiquitination	K422	MDM2 ?	18655026	N/A
MDM2	Q00987	phosphorylation	S395	ATM	11331603	Xtr
MDM4	O15151	phosphorylation	S367	CHK1 and CHK2	16163388	Dre
MDM4	O15151	phosphorylation	S342	CHK2	16163388	Dre
MDM4	O15152	phosphorylation	S403	ATM	16163388	Dre
MK03 (ERK1)	P27361	phosphorylation	T202	MAP2K1 and MAP2K2	17081983	Pfa
MK03 (ERK1)	P27361	phosphorylation	Y204	MAP2K1 and MAP2K2	17081983	Tbr
MPIP1 (Cdc25A)	P30304	phosphorylation	S76	CHK1	14681206	Ngr/Ddi/Xtr
MPIP1 (Cdc25A)	P30304	phosphorylation	S178	CHK1	12676583	Ddi/Tad/Xtr
MPIP1 (Cdc25A)	P30304	phosphorylation	T507	CHK1	14559997	Spo/Tad/Tca/Xtr
MPIP1 (Cdc25A)	P30304	phosphorylation	S124	CHK1 and CHK2	12676583	Ddi/Xtr
MPIP1 (Cdc25A)	P30304	phosphorylation	S279	CHK1 and CHK2	12676583	Gga
MPIP1 (Cdc25A)	P30304	phosphorylation	S293	CHK1 and CHK2	12676583	Ddi
MPIP3 (Cdc25C)	P30307	phosphorylation	S216	CHK1 and CHK2	15629715	Hsa
MPIP3 (Cdc25C)	P30307	phosphorylation	S216	MAP2K2	15629715	Hsa
MSH2	P43246	phosphorylation	S860	ATM or ATR	17525332	Ath/Cin
MSH3	P20585	phosphorylation	S201	ATM or ATR	17525332	Mus
MSH6	P52701	phosphorylation	S348	ATR ?	17525332	Ame
MYST1	Q9H7Z6	acetylation	K274	MYST1 (auto)	22020126	Ehu
NBN	O60934	phosphorylation	S343	ATM	10839545	Ppa
NBN	O60934	phosphorylation	S397	ATM	10839545	Dre
NBN	O60934	phosphorylation	S615	ATM	10839545	Hsa
NR4A2	P43354	phosphorylation	S181	PRKDC ?	17081983	Cte
NR4A2	P43354	phosphorylation	S337	PRKDC	21979916	Cte
PALB2	Q86YC2	phosphorylation	S59	ATM or ATR	17525332	Hsa
PALB2	Q86YC2	phosphorylation	S64	ATM or ATR	17525332	Mdo
PALB2	Q86YC2	phosphorylation	S157	ATM or ATR	17525332	Mdo
PALB2	Q86YC2	phosphorylation	S376	ATM or ATR	17525332	Oan
PRKDC	P78527	phosphorylation	T2609	auto	12186630	Tad
RAD17	O75943	phosphorylation	S656	ATM and ATR	11418864	Cpa/Bfl
RAD17	O75943	phosphorylation	S646	ATM and ATR	11418864	Cpa/Dre
RAD50	Q92878	phosphorylation	S635	ATM or ATR	17525332	Osa/Ecu/Dme
RAD51	Q06609	phosphorylation	T309	CHK1	15665856	Bna,Cpa/Ecu/Sja
RAD9A	Q99638	phosphorylation	S272	ATM and ATR	22453082	Tad
RBBP8 (CTIP)	Q99708	phosphorylation	S664	ATM	10910365	Cte
RBBP8 (CTIP)	Q99708	phosphorylation	S745	ATM	10910365	Xtr

RBBP8 (CTIP)	Q99708	ubiquitination	N/A	BRCA1 (4)	16818604	N/A
RFA1	P27694	phosphorylation	T180	ATM or ATR	17525332	Mus
RFA2	P15927	phosphorylation	S12	PRKDC	9139719	Bde/Xtr
RFA2	P15927	phosphorylation	S33	PRKDC	9139719	Ptr/Ath/Cel/Bfl
RN168	Q8IYW5	sumoylation	N/A	PIAS4	20016603	N/A
RNF8	O76064	sumoylation	N/A	PIAS4	20016603	N/A
SMC1A	Q14683	phosphorylation	S957	ATM	11877377	Spo/Nve, Cte/Cin
SMC1A	Q14683	phosphorylation	S966	ATM and ATR	11877377	Tbr/Osa/Nve
SMC5	Q8IY18	sumoylation	N/A	MMS21	18086888	N/A
SMC6	Q96SB8	sumoylation	N/A	MMS21	16055714	N/A
TIF1B (KAP1)	Q13263	phosphorylation	S824	ATM	17942393	Sja/Xtr
TIF1B (KAP1)	Q13263	phosphorylation	S473	CHK1 and CHK2	21851590	Xtr
TIF1B (KAP1)	Q13263	sumoylation	K554	auto	17079232/18082607	Xtr
TIF1B (KAP1)	Q13263	sumoylation	K779	auto	17079232/18082607	Xtr
TIF1B (KAP1)	Q13263	sumoylation	K804	auto	17079232/18082607	Xtr
TIPIN	Q9BVW5	phosphorylation	S222	ATM or ATR	17525332	Tca
TOPB1	Q92547	ubiquitination	N/A	UBR5	11714696	N/A
TRIPC	Q14669	phosphorylation	S1577	ATM or ATR	17525332	Ehu/Ath/Spo/Tad/Tca
UBE2T	Q9NPD8	ubiquitination	K91	auto	19111657	Ddi
UBE2T	Q9NPD8	ubiquitination	K182	auto	19111657	Xtr
UIMC1 (RAP80)	Q96RL1	phosphorylation	S140	ATM or ATR	17525332	Gga
UIMC1 (RAP80)	Q96RL1	phosphorylation	S402	ATM or ATR	17525332/17525340	Gga
UIMC1 (RAP80)	Q96RL1	phosphorylation	S419	ATM or ATR	17525332/17525340	Hsa
UIMC1 (RAP80)	Q96RL1	sumoylation	N/A	PIAS1	20016594	N/A
WEE1	P30291	phosphorylation	S53	PLK1	15070733	Spo/Dme/Dre
WEE1	P30291	ubiquitination	N/A	CUL1 - RBX1	15070733	N/A
XPA	P23025	phosphorylation	S196	ATM or ATR	17525332	Ehu
XPA	P23025	deacetylation	K63	SIR1	20670893	Cte
XPA	P23025	deacetylation	K67	SIR1	20670893	Cpa
XRCC1	P18887	phosphorylation	S371	PRKDC	16397295	Dre
XRCC4	Q13426	phosphorylation	S260	PRKDC	15177042	Gga
XRCC4	Q13426	phosphorylation	S320	PRKDC	15177042	Mdo
XRCC5 (Ku80)	P13010	phosphorylation	S577	PRKDC	10026262	Ppa/Ddi/Tad,Nve/Gga
XRCC5 (Ku80)	P13010	phosphorylation	S580	PRKDC	10026262	Ppa/Tad/Mdo
XRCC6 (Ku70)	P12956	phosphorylation	S6	PRKDC	10026262	Sja
XRCC6 (Ku70)	P12956	phosphorylation	S51	PRKDC	9362500	Ath/Mbr/Xtr

Observations

(1) Probably the phosphorylation is carried by other kinase not yet identified
(2) DSGxxS degron sequence (D29-S34) conserved from Xtr
(3) By similarity with ubiquitination sites in other histones
(4) It appears that CtIP can be ubiquitinated by BRCA1 interchangeably at multiple lysine residues.
? No certainty of PTM actually being exterted by the modifier

ST9. PTMs ages emergence analysis

Target	Target age code	Modifier	Modifier age code	Conservation code	Pair conservation	Ages
ATM	3	ATM (auto)	3	C	Ppa-Ppa	Plants
ATM	3	KAT5	2	B	Ppa-Ptr	Plants-Early Euks
BRCA1	2	ATM	3	A	Ngr-Ppa	Early Euks-Plants
BRCA1	2	ATR	2	B	Ngr-Ehu	Early Euks
BRCA1	2	CHK2	4	A	Ngr-Ddi	Early Euks-Unikonta
BRCA1	2	PIAS1	3	A	Ngr-Cre	Early Euks-Plants
BRCA1	2	UBE2T	1	B	Ngr-Pst	Early Euks-Bacteria
BRCA2	2	CHK1	5	A	Cpa-Ecu	Early Euks-Opisthokonta
BRCA2	2	CHK2	4	A	Cpa-Ddi	Early Euks-Unikonta
BRCA2	2	ATM	3	A	Cpa-Ppa	Early Euks-Plants
BRCA2	2	ATR	2	B	Cpa-Ehu	Early Euks
CDT1	2	CUL4	2	B	Cpa-Ehu	Early Euks
CHK1	5	ATR	2	B	Ecu-Ehu	Opisthokonta-Early Euks
CHK2	2	ATM	3	A	Cpa-Ppa	Early Euks-Plants
CLSPN	5	CHK1 ? (1)	5	B	Spo-Ecu	Opisthokonta
CLSPN	5	ATR ? (1)	2	B	Spo-Ehu	Opisthokonta-Early Euks
CLSPN	5	PLK1 (2)	1	B	Spo-Pst	Opisthokonta-Bacteria
CLSPN	5	CUL1	2	B	Spo-Cpa	Opisthokonta-Early Euks
CLSPN	5	RBX1	2	B	Spo-Gth	Opisthokonta-Early Euks
DCR1B (Apollo)	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
DCR1B (Apollo)	2	ATR	2	C	Ehu-Ehu	Early Euks
DCR1C (Artemis)	2	ATM	3	A	Ngr-Ppa	Early Euks-Plants
DCR1C (Artemis)	2	PRKDC	2	C	Ngr-Ngr	Early Euks
EXO1	2	ATR	2	B	Cpa-Ehu	Early Euks
F175A (Abraxas)	10	ATM	3	B	Dre-Ppa	Vertebrata-Plants
F175A (Abraxas)	10	ATR	2	B	Dre-Ehu	Vertebrata-Early Euks
FACD2	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
FACD2	2	UBE2T	1	B	Ehu-Pst	Early Euks-Bacteria
FBX31	6	ATM	3	B	Tad-Ppa	Metazoa-Plants
H2AX	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
H2AX	2	PRKDC	2	A	Ehu-Ngr	Early Euks
H2AX	2	UBE2N (UBC13)	1	B	Ehu-Pst	Early Euks-Bacteria
H2AX	2	RN168 ? (3)	9	A	Ehu-Cin	Early Euks-Chordata
H2AX	2	RNF8 ? (3)	2	A	Ehu-Ngr	Early Euks
H2AX	2	KAT5 (Tip60) ?	2	A	Ehu-Ptr	Early Euks
HERC2	6	After IR (ATM ?)	3	B	Tad-Ppa	Metazoa-Plants
HNRPK	5	MDM2 ?	6	A	Mbr-Tad	Opisthokonta-Metazoa
MDM2	6	ATM	3	B	Tad-Ppa	Metazoa-Plants
MDM4	10	CHK1	5	B	Dre-Ecu	Vertebrata-Opisthokonta
MDM4	10	CHK2	4	B	Dre-Ddi	Vertebrata-Unikonta
MDM4	10	ATM	3	B	Dre-Ppa	Vertebrata-Plants
MK03 (ERK1)	2	MAP2K2	5	A	Pfa-Spo	Early Euks-Opisthokonta
MPIP1 (Cdc25A)	4	CHK1	5	A	Ddi-Ecu	Unikonta-Opisthokonta
MPIP1 (Cdc25A)	4	CHK2	4	C	Ddi-Ddi	Unikonta
MPIP3 (Cdc25C)	5	CHK1	5	C	Ecu-Ecu	Opisthokonta
MPIP3 (Cdc25C)	5	CHK2	4	B	Ecu-Ddi	Opisthokonta-Unikonta
MPIP3 (Cdc25C)	5	MAP2K2	5	A	Ecu-Spo	Opisthokonta
MSH2	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
MSH2	2	ATR	2	C	Ehu-Ehu	Early Euks
MSH3	1	ATM	3	A	Bap-Ppa	Bacteria-Plants
MSH3	1	ATR	2	A	Bap-Ehu	Bacteria-Early Euks
MSH6	1	ATR ?	2	A	Eco-Ehu	Bacteria-Early Euks
MYST1	2	MYST1 (auto)	2	C	Ehu-Ehu	Early Euks
NBN	3	ATM	3	C	Ppa-Ppa	Plants
NR4A2	8	PRKDC	2	B	Cte-Ngr	Bilateria-Early Euks
PALB2	10	ATM	3	B	Dre-Ppa	Vertebrata-Plants
PALB2	10	ATR	2	B	Dre-Ehu	Vertebrata-Early Euks
PRKDC	2	PRKDC (auto)	2	C	Ngr-Ngr	Early Euks
RAD17	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
RAD17	2	ATR	2	C	Ehu-Ehu	Early Euks
RAD50	1	ATM	3	A	Eco-Ppa	Bacteria-Plants
RAD50	1	ATR	2	A	Eco-Ehu	Bacteria-Early Euks
RAD51	1	CHK1	5	A	Eco-Ecu	Bacteria-Opisthokonta
RAD9A	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
RAD9A	2	ATR	2	C	Ehu-Ehu	Early Euks
RBBP8 (CTIP)	8	ATM	3	B	Cte-Ppa	Bilateria-Plants
RBBP8 (CTIP)	8	BRCA1 (4)	2	B	Cte-Ngr	Bilateria-Early Euks
RFA1	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
RFA1	2	ATR	2	C	Ehu-Ehu	Early Euks
RFA2	2	PRKDC	2	A	Ehu-Ngr	Early Euks
RN168	9	PIAS4	5	B	Cin-Sce	Chordata-Opisthokonta
RNF8	2	PIAS4	5	A	Ngr-Sce	Early Euks-Opisthokonta
SMC1A	1	ATM	3	A	Eco-Ppa	Bacteria-Plants
SMC1A	1	ATR	2	A	Eco-Ehu	Bacteria-Early Euks
SMC5	2	MMS21	2	A	Ehu-Ptr	Early Euks
SMC6	2	MMS21	2	A	Ehu-Ptr	Early Euks
TIF1B (KAP1)	8	ATM	3	B	Cte-Ppa	Bilateria-Plants
TIF1B (KAP1)	8	CHK1	5	B	Cte-Ecu	Bilateria-Opisthokonta
TIF1B (KAP1)	8	CHK2	4	B	Cte-Ddi	Bilateria-Unikonta
TIF1B (KAP1)	8	TIF1B (auto)	8	C	Cte-Cte	Bilateria
TIPIN	3	ATM	3	C	Ppa-Ppa	Plants
TIPIN	3	ATR	2	B	Ppa-Ehu	Plants-Early Euks
TOPB1	2	UBR5	6	A	Ehu-Tad	Early Euks-Metazoa
TRIPC	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
TRIPC	2	ATR	2	C	Ehu-Ehu	Early Euks
UBE2T	1	UBE2T (auto)	1	C	Pst-Pst	Bacteria
UIMC1 (RAP80)	10	ATM	3	B	Dre-Ppa	Vertebrata-Plants
UIMC1 (RAP80)	10	ATR	2	B	Dre-Ehu	Vertebrata-Early Euks
UIMC1 (RAP80)	10	PIAS1	3	B	Dre-Cre	Vertebrata-Plants
WEE1	2	PLK1	1	B	Tbr-Pst	Early Euks-Bacteria
WEE1	2	CUL1	2	B	Tbr-Cpa	Early Euks
WEE1	2	RBX1	2	B	Tbr-Gth	Early Euks
XPA	2	ATM	3	A	Ehu-Ppa	Early Euks-Plants
XPA	2	ATR	2	C	Ehu-Ehu	Early Euks
XPA	2	SIR1	4	A	Ehu-Ddi	Early Euks-Unikonta
XRCC1	2	PRKDC	2	C	Ngr-Ngr	Early Euks
XRCC4	3	PRKDC	2	B	Osa-Ngr	Plants-Early Euks
XRCC5 (Ku80)	2	PRKDC	2	A	Ehu-Ngr	Early Euks
XRCC6 (Ku70)	2	PRKDC	2	A	Ehu-Ngr	Early Euks

Age codes are referred to Figure 10 (see Methods 3.3).
Conservation codes: A, target older than modifier; B, modifier older than target; C, target and modifier from the same age.

Observations
(1) Probably the phosphorylation is carried by other kinase not yet identified
(2) DSGxxS degron sequence (D29-S34) conserved from Xtr
(3) By similarity with ubiquitination sites in other histones
(4) It appears that CtIP can be ubiquitinated by BRCA1 interchangeably at multiple lysine residues.
? No certainty of PTM actually being exterted by the modifier

ST10. DDR proteins and disease

Protein	Disease	OMIM	Subcellular location
ATM	Ataxia telangiectasia (AT)	208900	Nucleus. Cytoplasmic vesicle.
ATR	Seckel syndrome type 1 (SCKL1)	210600	Nucleus.
BLM	Bloom syndrome (BLM)	210900	Nucleus.
BRCA1	Breast cancer (BC)	114480	Nucleus
BRCA2	Breast cancer (BC); Pancreatic cancer type 2 (PNCA2); Breast-ovarian cancer familial type 2 (BROVCA2); Fanconi anemia complementation group D type 1 (FANCD1); Glioma type 3 (GLM3)	114480; 613347; 612555; 605724; 613029	Nucleus.
CDT1	Meier-Gorlin syndrome type 4 (MGORS4)	613804	Nucleus
CHK2	Li-Fraumeni syndrome 2 (LFS2)	609265	Nucleus
DCR1B (Apollo)	Hoyeraal-Hreidarsson syndrome (HHS)	300240	Chromosome › telomere. Nucleus. Cytoplasm › cytoskeleton › centrosome.
DCR1C (Artemis)	Omenn syndrome (OS); Severe combined immunodeficiency autosomal recessive T-cell-negative/B-cell-negative/NK-cell-positive with sensitivity to ionizing radiation (RS-SCID)	603554; 602450	Nucleus.
ERCC1	Cerebro-oculo-facio-skeletal syndrome type 4 (COFS4)	610758	Nucleus
ERCC2 (XPD)	Xeroderma pigmentosum complementation group D (XP-D); Trichothiodystrophy photosensitive (TTDP); Cerebro-oculo-facio-skeletal syndrome type 2 (COFS2)	278730; 601675; 610756	Nucleus. Cytoplasm (spindle)
ERCC3 (XPB)	Xeroderma pigmentosum complementation group B (XP-B); Trichothiodystrophy photosensitive (TTDP)	610651; 601675	Nucleus
ERCC5 (XPG)	Xeroderma pigmentosum complementation group G (XP-G)	278780	Nucleus

ERCC6 (CSB)	Cockayne syndrome type B (CSB);
	Cerebro-oculo-facio-skeletal syndrome type 1 (COFS1);
	De Sanctis-Cacchione syndrome (DSC);
	Susceptibility to age-related macular degeneration type 5 (ARMD5);
	UV-sensitive syndrome (UVS)
		133540; 214150; 278800; 613761; 600630	Nucleus

ERCC8 (CSA)	Cockayne syndrome type A (CSA)	216400	Nucleus
FACD2	Fanconi anemia complementation group D type 2 (FANCD2)	227646	Nucleus
FANCM	Fanconi anemia complementation group M (FANCM)	614087	Nucleus
HERC2	Associated with skin/hair/eye pigmentation variability type 1 (SHEP1)	227220	Cytoplasm. Nucleus.
LIG4	LIG4 syndrome (LIG4S); Severe combined immunodeficiency autosomal recessive T-cell-negative/B-cell-negative/NK-cell-positive with sensitivity to ionizing radiation (RS-SCID)	606593; 602450	Nucleus
MLH1	Hereditary non-polyposis colorectal cancer type 2 (HNPCC2)/Lynch syndrome; Mismatch repair cancer syndrome (MMRCS)/Turcot syndrome/Brain tumor-polyposis syndrome 1 (BTPS1); Muir-Torre syndrome (MuToS/MTS); Susceptibility to endometrial cancer (ENDMC)	609310; 276300; 158320; 608089	Nucleus
MRE11	Ataxia telangiectasia-like disorder (ATLD)	604391	Nucleus
MSH2	Hereditary non-polyposis colorectal cancer type 1 (HNPCC1); Hereditary non-polyposis colorectal cancer type 8 (HNPCC8); Muir-Torre syndrome (MuToS/MTS); Susceptibility to endometrial cancer (ENDMC)	120435; 613244; 158320; 608089	Nucleus
MSH3	Susceptibility to endometrial cancer (ENDMC)	608089	Nucleus
MSH6	Hereditary non-polyposis colorectal cancer type 5 (HNPCC5); Susceptibility to endometrial cancer (ENDMC)	600678; 608089	Nucleus
NBN	Nijmegen breakage syndrome (NBS); Susceptibility to breast cancer (BC)	251260; 114480	Nucleus
PALB2	Fanconi anemia complementation group N (FANCN); Pancreatic cancer type 3 (PNCA3)	610832; 613348	Nucleus
PMS2	Hereditary non-polyposis colorectal cancer type 4 (HNPCC4); Mismatch repair cancer syndrome (MMRCS)	600259; 276300	Nucleus
RAD50	Nijmegen breakage syndrome-like disorder (NBSLD)	613078	Nucleus. Chromosome › telomere.
RAD51	Susceptibility to breast cancer (BC)	114480	Nucleus. Cytoplasm. Mitochondrion matrix.
RN168	Riddle syndrome (RIDDLES)	611943	Nucleus.
SLX4	Fanconi anemia complementation group P (FANCP)	613951	Nucleus
SMAL1	Schimke immuno-osseous dysplasia (SIOD)	242900	Nucleus.
SMC1A	Cornelia de Lange syndrome type 2 (CDLS2)	300590	Nucleus. Chromosome › kinetochore
TDP1	Spinocerebellar ataxia autosomal recessive with axonal neuropathy (SCAN1)	607250	Nucleus. Cytoplasm
XLF	Severe combined immunodeficiency due to NHEJ1 deficiency (NHEJ1-SCID)	611291	Nucleus
XPA	Xeroderma pigmentosum complementation group A (XP-A)	278700	Nucleus
XPC	Xeroderma pigmentosum complementation group C	278720	Nucleus. Cytoplasm
XPF	Xeroderma pigmentosum complementation group F (XP-F)	610965	Nucleus