DNA-Protein Crosslink Proteolysis Repair

Bruno Vaz*, Marta Popovic* and Kristijan Ramadan

Cancer Research UK and Medical Research Council Oxford Institute for Radiation Oncology, Department of Oncology, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, UK.

*These authors contributed equally to this work.

Correspondence: kristijan.ramadan@oncology.ox.ac.uk

Keywords: DNA-protein crosslink repair, SPARTAN/DVC1 protease, Wss1 protease, proteolysis-dependent DNA-protein crosslink repair, DNA replication, genome stability, ageing, cancer

Abstract

Proteins that are covalently bound to DNA constitute a specific type of DNA lesions known as DNA-protein crosslinks (DPCs). DPCs represent physical obstacles to the progression of DNA replication. If not repaired, DPCs cause stalling of DNA replication forks that consequently leads to DNA double strand breaks, the most cytotoxic DNA lesion. Although DPCs are abundant DNA lesions, the mechanism of DPC repair was unclear until now. Recent work unveiled that DPC repair is orchestrated by proteolysis performed by two distinct metalloproteases, SPARTAN in metazoans and Wss1 in yeast. This review summarises recent discoveries on two proteases in DNA replication-coupled DPC repair and establishes DPC proteolysis repair as a separate DNA repair pathway for genome stability and protection from accelerated ageing and cancer.

Overview of DNA-protein crosslinks

DNA contains all genetic information a cell and organism requires to grow, differentiate, survive and divide. DNA is an extremely fragile molecule embedded in highly reactive environment such as H_2O. DNA is constantly damaged by various endogenous and exogenous agents, such as reactive oxygen species and UV light, respectively. Loss of DNA integrity leads to various defects in cellular physiology and consequently to diseases such as cancer, diabetes, accelerated aging, neurodegeneration or cell death. To preserve DNA integrity all organisms possess an elaborate genome maintenance apparatus, consisting of multiple DNA damage repair (DDR) and DNA damage tolerance (DDT) pathways (see Glossary) (Figure 1). Different DDR pathways are involved in the recognition and repair of specific types of DNA lesions [1, 2]. Currently, most of the known DDR pathways are well-characterised. However, it is still not well understood how DNA-protein crosslinks are repaired [3]. Although DPCs are one of the most abundant DNA lesions and their presence, if not removed, is cytotoxic, the existence of a specialised DPC repair pathway remained elusive. Recently, several research groups identified a unique DNA-protein crosslink repair pathway based on proteolysis, which we have termed here as DNA-protein crosslink proteolysis repair (DPC-PR) [4-8]. DPC-PR pathway is conserved from yeast to humans and is orchestrated by DNA-dependent proteases Wss1 in yeast and SPARTAN (SPRTN), also known as DVC1, in metazoans. The aim of this review is to establish DPC-PR as a unique DNA repair pathway based on DNA replication-coupled proteolysis orchestrated by SPRTN protease in metazoans and Wss1 in yeast.

Origins and chemistry of DNA-protein crosslinks

DPCs are created when proteins covalently and irreversibly bind to DNA. Virtually any protein in close proximity to DNA can be crosslinked to DNA upon exposure to various endogenous or exogenous crosslinking agents. Mass-spectrometry analyses identified numerous DNA binding proteins including histones, transcription factors, DNA repair and replication proteins, as well as non-DNA binding proteins as DPCs [4, 9-11]. DPC classification is explained in Box 1. Given the high concentration of histones in close proximity to DNA, it is not surprising that histones are among the most abundant DPCs [12]. Aldehydes, reactive oxygen (ROS), nitrogen species (NOS) and DNA helical alterations are one of the most common endogenous DPC-inducing sources. Aldehydes are generated by normal cellular metabolic pathways including histone demethylation [13], AlkB-type repair [14] amino acid metabolism [15] and lipid peroxidation [16]. It is estimated that endogenous formaldehyde concentration in human plasma is 100 μM, and in some tissues can reach up to 400 μM [17, 18]. Similarly, ROS and NOS are metabolic side products of cellular respiration, immune responses and inflammation [19]. While aldehydes, ROS and NOS crosslink any protein in the vicinity of DNA, DNA helical alterations such as DNA abasic sites and oxanines (nitric oxide induced guanine lesions) are so far shown to crosslink histones [20], DNA glycosylases, DNA polymerase β [21] and high mobility group (HMG) proteins [22].

Cancer treatment including Ionizing Radiation (IR) and anticancer drugs as well as exposure to UV-light are the main exogenous DPC-inducing factors. IR causes DPCs directly, through the irradiation of the DNA backbone which creates unstable DNA cation radicals or indirectly, through the transfer of radiation energy to water molecules which in turn enhances local concentrations of ROS. Anticancer drugs like

mitomycin C, platinum compounds, nitrogen mustard, DNA methyltransferases inhibitors [23] and Topoisomerase poisons (camptothecin and etoposide) [24, 25] cause DPCs through agent-specific mechanisms. UV-light excites DNA bases, most commonly thymidines, which in turn covalently bind to amino acids, specifically cysteine, lysine, phenylalanine, tryptophan or tyrosine with highest efficiency [26]. For the detailed chemistry of DPCs see Box 2. Altogether, DPCs are constantly formed in our genome and if not repaired cause severe threat to genome integrity.

Involvement of canonical DNA repair pathways in DPC repair

Although DPCs are abundant DNA lesions, the mechanisms of DPC repair were under-investigated and thus poorly understood [12]. A few biochemical and genetic studies in different organisms suggested that two canonical DNA repair pathways, namely nucleotide excision repair (NER) and homologous recombination (HR), orchestrate DPC repair and protect cells from DPC-induced cytotoxicity (Figure 2A) [27]. This concept came from the initial studies in bacteria where NER was found to remove and repair small DPCs (smaller than 16kDa) [28], while HR and subsequent replication restart repaired bulky DPCs [29] [12]. A genome-wide screen in yeast implicated NER in the repair of DPCs after acute exposure to high formaldehyde doses and HR in the repair of DPCs after chronic exposure to low formaldehyde doses [30]. Sensitivity analysis in mutant yeast strains also showed that NER is dominant in the DPC repair after high formaldehyde doses [5]. The coordination of NER and HR in yeast is probably dependent on cell cycle phase (high formaldehyde doses cause cell cycle arrest and thus favour NER), and the size of the DPC, similar as in bacteria, although this has not been shown so far. The analysis of mammalian NER excision capacity *in vitro* and *in vivo* showed that NER is only able to remove DNA-

crosslinked proteins of less than 8 - 10 kDa [28, 29, 31-33]. Correspondingly, cells from *Xeroderma pigmentosum* patients (bearing mutations in different NER factors) are sensitive to different DPC-inducing agents [34, 35].

Unlike NER, the involvement of HR in DPC repair has only been shown indirectly in bacteria and yeast, and more recently in metazoans. Thus, it is still difficult to understand whether HR is directly involved in DPC repair. HR-deficient *E. coli* cells (*ΔrecA ΔrecB*) were sensitive to DPC-inducing agents, formaldehyde and azacytidine [29, 36]. In yeast, deletion of the HR genes *sgs1*, *xrs2*, *mre11*, *rad50* and/or *rad52* sensitised cells to chronic doses of formaldehyde [5, 30]. The role of HR in DPC repair was further supported by the observation that formaldehyde-treated cells showed elevated levels of DSBs and Rad51 foci and an increased rate of sister chromatid exchange (SCE) events [37]. However, involvement of HR is not surprising given that one of the main outcomes of DPC accumulation is the emergence of DSBs [4, 38]. As formaldehyde not only induces DPCs, but also protein-protein crosslinks and, more importantly DNA inter- and intra-strand crosslinks (ICL), it activates the Fanconi anemia pathway as well as HR. Thus, all aforementioned experimental endpoints such as cell survival, SCE, and DSB formation measure the total cellular response to the various types of formaldehyde-induced DNA damage, and not DPCs alone. Therefore, it is hard to conclude to what extent HR is involved, if at all, in DPC repair.

Interestingly, recent work demonstrates that depletion of Mre11, a crucial nuclease in HR, does not impair general DPC removal in human cells [4], thus supporting a hypothesis that HR cannot deal with the majority of diverse DPCs. However, two recent studies in human cells have demonstrated that MRE11, independently of its

function in HR and together with another nuclease, CtIP, removes TOP2-ccs. MRE11 does not directly act on TOP2-ccs, but endonucleolytically cleaves DNA 15-20 bp downstream of TOP2-cc [39, 40]. Thus, by removing TOP2-ccs from the 5' end of DSBs, MRE11 removes DPCs after DSB formation. Similarly, biochemical data in *Xenopus* egg extract demonstrated that TOP2-cc complex is processed by cooperation of MRN (MRE11-RAD50-NBS1) complex, BRCA1 and CtIP [41]. Altogether, these data suggest that MRN complex removes covalently attached TOP2 from DSBs in both HR-dependent and HR-independent manner. However, further studies are needed to clarify a direct role of HR in DPC repair.

Considering that NER can act only on small DPCs, and HR removes only DPCs after DSB formation, it was speculated that the Fanconi Anemia DNA repair pathway might be involved in DPC repair. Fanconi anemia pathway-deficient cells were found to be hypersensitive to formaldehyde and azacytidine but not to the specific TOP1-cc-inducing agent camptothecin [4, 42, 43]. However, similar to HR- and NER-deficient cells, cells deficient of FANCD2, one of the main component of the Fanconi Anemia pathway, do not accumulate DPCs and exhibit normal DPC repair kinetics following treatment with formaldehyde [4, 6]. Furthermore, immunodepletion of FANCD2 from *Xenopus* egg extracts did not affect DNA replication fork progression past DPCs [44]. Therefore, the role of the Fanconi Anemia pathway in DPC repair seems to be strictly associated with the repair of DNA interstrand crosslinks (ICLs), a lesion also induced by formaldehyde, and not with DPC repair.

In conclusion, in eukaryotes canonical DNA repair pathways such as NER and HR remove DPCs via the activity of nucleases that cleave DNA near to where a DPC is formed. While the role of NER in the repair of small DPCs is beneficial to cells, the

role of HR in removing bulky DPCs requires DSB formation. Thus, activating HR-pathway for DPC repair may be deleterious for cells due the potentially cytotoxic consequences of DSBs. However, it is quite clear that the Fanconi Anemia pathway is not involved in DPC repair pathway.

Proteolysis Orchestrated DPC repair

Due to diverse array of DPCs (various proteins attached with different chemistry to DNA) contrasted with the known specificity of the canonical DNA repair pathways postulated to be involved in DPC removal (see above). Thus, it was speculated that cells must contain a specialised DNA repair pathway that is based on direct removal (proteolysis) of DPCs. NER can only excise small DPCs (max size of 8-16 kDa) suggesting that bulky (bigger than 16 kDa) DPCs need to be processed into smaller peptides before the action of NER. Similarly, removal of enzymatic DPCs, TOP1-ccs and TOP2-ccs by Tyrosyl-DNA phosphodiesterase 1 (TDP1) and 2 (TDP2), respectively, requires upstream proteolysis of TOP1 and 2 into smaller peptides [45] as TDP1 and 2 can efficiently process peptides of ~150 amino acid long [46-48] (Box3). Altogether, these data suggest that it must exists a protease that proteolysis large DPCs to small peptide remnants attached to the DNA backbone. These peptide remnants are further processed by NER, TDP1 and TDP2 or bypassed by translesion DNA synthesis (TLS) during DNA synthesis. Without such a protease, bulky DPCs would block the progression of the DNA replication fork and lead to DSBs in proliferative cells. Thus, proteolysis-coupled DPC repair was proposed [49].

Proteasome in DPC repair

The proteasome, being the main protease involved in protein degradation, was

considered to be involved in proteolytic DPC repair. However, the role of the proteasome system in DPC processing remains unclear due to contradictory literature reports. In bacteria, inhibition of ATP-dependent proteases, which function like a proteasome, did not affect cell survival after exposure to DPC inducing agents formaldehyde and azacytidine [29]. In human cells, proteasome inhibition prevented the removal of histone DPCs, TOP1ccs, and TOP2ccs [50-52] and sensitised human cells to low doses of formaldehyde [53]. By contrast, in *Xenopus* egg extracts, DPC proteolysis and bypass during DNA replication was inhibited upon depletion of the free pool of ubiquitin, but not by proteasome inhibition [44]. These discrepancies could be explained by the fact that proteasome inhibitors MG132, lactacystin, and bortezemib used in the aforementioned studies can also deplete the nuclear ubiquitin pool in human cells, thus making it hard to conclude whether proteasome is directly involved in DPC repair [54]. Altogether, the data suggest that DPC proteolysis is an ubiquitin-dependent process but a direct role of proteasome in DPC repair is still not clear. A precise understanding of the role of proteasome in DPC repair requires additional studies and different approaches including usage of specific proteasome inhibitors and *in vitro* studies.

DNA-dependent Proteases in DPC repair

The proteolysis-dependent model of DPC repair [49] was further supported by several studies, which all demonstrated replication-coupled proteolysis of a specific DPC *in vitro* [44, 55, 56]. However, the protease in question remained unknown. This model was further supported by the discovery of Wss1, a protease in yeast found to cleave TOP1, histone H1 and HMG proteins *in vitro* and contribute to the cellular resistance to formaldehyde and camptothecin [5]. Four recent studies identified a new

mammalian protease, SPARTAN (SPRTN), also known as DNA damage valosin containing factor 1 (DVC1), as a crucial component of DPC repair in human cells [4, 6-8]. Interestingly, even though bacterial species possess SPRT-like proteins (Figure 3A), proteolysis-dependent DPC repair in bacteria has not yet been reported. Like SPRTN [4], yeast protease Wss1 [57] binds zinc and possesses an HEXXH active site, thus making both of these enzymes members of the zinzin family of metallopeptidases [4]. SPRTN is a pleiotropic DNA-dependent protease that cleaves several chromatin-associated substrates, including core histones, H2A, H2B, H3, H4, linker histone H1, HMG1, HLTF, Fan1, TOP1 and TOP2 [4, 6-8]. Both proteases, Wss1 and SPRTN, are linked to DNA replication and share some common characteristics [4, 8, 58-60]. Wss1 and SPRTN need DNA to activate their proteolytic activity, cleave DNA binding proteins, physically interact with AAA ATPase p97/VCP, the central component of the ubiquitin-proteasome system, and inactivation of Wss1 and SPRTN hypersensitises yeast and human cells to formaldehyde, respectively.

Although Wss1 and SPRTN share similar proteolytic activity *in vitro, in vivo* they show considerable differences in DPC removal and sensitivity to DPC-inducing agents. SPRTN alone prevents accumulation of endogenous DPCs, as well as formaldehyde-induced DPCs [4, 6]. Concordantly, SPRTN protects cells from formaldehyde-induced DPC toxicity [4, 6, 7]. On the contrary, Wss1 is not involved in the removal of DPCs following formaldehyde treatment, but does partially protect cells from formaldehyde-induced DPC toxicity [5]. Another difference between Wss1 and SPRTN is observed in the repair of TOP1-ccs. SPRTN deficiency results in Top1-cc accumulation and severe sensitivity to CPT [4], while Wss1 depletion does not cause any adverse effects in untreated yeast cells [5]. Only upon co-depletion of

Wss1 with Tdp1 is cell survival affected, while in mammals, TDP1 and SPRTN co-depletion has not an additive effect in comparison to SPRTN depletion alone [4]. The *in vivo* differences between the two proteases are further demonstrated by the inability of ectopic SPRTN over-expression to rescue the phenotypes of Wss1-deficient yeast cells [7]. However, despite numerous functional differences, both proteases are essential for replication fork progression [4, 8, 58, 61], indicating that Wss1, like SPRTN, removes DPC blocks during replication. Unlike SPRTN in mammals [62-64], Wss1 is not an essential gene in yeast [65], indicating that the function of Wss1 can be compensated. Differences in the phenotypes observed in Wss1- and SPRTN-deficient cells could be due to the existence of other proteases that process DPCs in yeast. Considering the high toxicity of DPCs and the high growth rate of yeast cells, it would not be surprising if yeasts possess other DNA dependent-proteases. Accordingly, in *Schizosaccharomyces pombe* a Wss2 protease was found to protect cells from acetaldehyde [66].

SPRTN and Wss1 are Two Distinct Proteases

Published literature indicates that SPRTN and WSS1 diverged through evolution from a common ancestor [67]. However, the functional and cellular differences between Wss1 and SPRTN prompted us to phylogenetically analyse the zinzin metallopeptidase superfamily. Phylogenetic analysis suggests that Wss1 and SPRTN are members of two separate families (Figure 3A). The SPRT family, where SPRTN belongs, consists of five subgroups: bacterial, archaea, cyanobacterial, plant and animal. The prokaryote and animal SPRT families map to the same branch of the phylogenetic tree suggesting they share a common ancestor. WLM family, where Wss1 belongs, is equally distant from SPRT family as it is from another gluzincin

family of alanyl aminopeptidases (Figure 3A). This analysis indicates that WLM and SPRT families do not share a common ancestor as was previously suggested [67]. The differences in our methodology compared to previous phylogenetic studies comparing SPRT and WLM families are: (i) an extended number of species were included in the analysis, most importantly prokaryotes (bacteria, archaea, cyanobacteria) and plants which increases the accuracy of tree topologies; and (ii) our analysis included another gluzincin family which enables comparative perspective to the relationship between SPRT and WLM families (Figures S1, S2 and Table S1). Moreover, the SPRT family is present in bacteria, archaea, cyanobacteria, funghi, plants and animals, but is absent from yeast (Figure 3A, Table S1). Wss1 is part of the WLM protein family which, like the SPRT family, consists of zinzin metalloproteases [68]. WLM proteins are present in yeast, funghi and plants, but are absent from animals and bacteria (Fig 3A) [12, 68]. A conserved feature of all zinzin metalloproteases is a short consensus HEXXH motif in their active centres, which includes two zinc-binding histidines and a glutamic acid. SPRT and WLM domains (Figure 3B) are also very different in terms of amino acid sequence identity (5 % identical, 14% similar) and can only be aligned over a short region around the HEXXH motif (Figure 2S) [4]. In addition, SPRT domains have many highly conserved regions, which are not present in WLM domains (Figure S2). To further strengthen our finding that SPRTN and Wss1 are two independently-evolved enzymes, we modeled the structure of the SPRT domain of SPRTN using the recently solved crystal structure of Wss1b in fission yeast (Schizosaccharomyces pombe) (PDB 5JIG) [6] as a template (homology modelling, SWISS-MODEL) [69]. The only part of SPRTN that modelled to Wss1b with high confidence (Supplementary Methods) was a protease core consisting of two α-helices containing the HEXXH motif and the third zinc binding histidine residue (Figure 3C,

right). The described protein core of SPRTN is shared with other gluzincin metallopeptidases, including Wss1b among others. This is confirmed by a homology model of SPRTN domain using another zinzin protease, abylysin (PDB 4JIU) (Figure 3C, left). Indeed, the region over which the SPRT domain could be modelled with high confidence was longer when aligned to abylysin than to Wss1b. The additional part of SPRTN domain modeled with high confidence includes three β-sheets upstream of the HEXXH active centre.

Apart from the differences among the SPRT and WLM domains in terms of amino acid sequence and structure, both proteins differ distinctly in their C-terminal regions. SPRTN has a long C-terminal arm (276 amino acid long), while Wss1 has a comparatively short C-terminus (48 amino acid long) (Figure 3B). Other than both having a p97/Cdc48-binding motifs, the C-terminal regions of both proteins share no similarities: SPRTN contains a PCNA binding motif (PIP) and a ubiquitin-binding motif whereas Wss1 contains SUMO binding motifs (Figure 3B). The C-terminal part of SPRTN is involved in TLS, where it serves as a platform to recruit p97 at sites of stalled DNA replication fork. For detailed role of SPRTN and p97 in TLS please read the following literature [59, 70-75].

Therefore, we suggest that extrapolating similarities between SPRTN and Wss1 with respect to substrate specificity, affinity, mode of substrate binding and recruitment to chromatin should be done with caution. Most importantly, structural extrapolations should not be made before the crystal structure of SPRTN is solved. We conclude that the WLM and SPRT families are two separate families within the gluzincin subgroup of the zinzin superfamily. Like other gluzincins, they share similar properties such as an HEXXH proteolytic active site. Moreover, SPRTN is evolutionary closer to

bacterial SPRT proteases than to Wss1. Thus, we would like to clear up the confusion in the published literature, which occasionally states that yeast Wss1 is an ortholog of SPRTN. Any similar functional properties shared by these two proteases is a result of a convergent rather than divergent evolution [4].

DNA Replication-coupled DPC Proteolysis Repair

DPCs constitute strong physical blocks for the progression of DNA replication, causing DNA replication fork stalling and, consequently, fork collapse [4, 8]. Using *in vitro* approach in *Xenopus* egg extract, it was recently demonstrated that in order for replication to progress in the presence of DPCs, DPCs have to be cleaved into smaller peptides on both the leading and lagging DNA strand [44]. However, the protease involved in the processing of DPCs remained unknown until SPRTN was identified as the S-phase specific protease responsible for DPC repair [4] (Figure 2B). Recent findings showed that the majority of DPCs are indeed removed specifically during S-phase [4]. As a part of the replisome, SPRTN prevents fork stalling and DSB formation caused by DPC accumulation and protects replicative cells from DPC-induced toxicity [4, 8]. However, the precise orchestration of replication-coupled DPC-PR is still unknown, specifically: (i) which factors act downstream of SPRTN to remove peptide remnants after DPC proteolysis and whether this occurs only in S-phase; (ii) how SPRTN protease is regulated, and (iii) which other factors are involved in SPRTN-dependent DPC-PR. NER, TLS, HR are all possible candidates for action during S-phase, while NER and MRE11 nuclease activity remain active in non-cycling cells too. Furthermore, it is still an open question as to which factors act upstream of TDP1 in non-cycling cells considering that it is known that TDP1 removes TOP1cc peptide remnants at the sites of transcription stalling in post-mitotic

cells [76, 77]. Overall, we cannot exclude that DPC repair also occurs in non-cycling cells, especially in non-proliferative cells such as neurons where DPC accumulation would pose a threat to transcription. However, work presented here strongly suggests that SPRTN protease, as a component of the DNA replication machinery, forms a unique, proteolysis based DNA repair pathway for DPC repair during DNA synthesis.

DPCs and human pathogenesis

The contribution of DPCs to human pathogenesis was proposed by several studies that associated different DPC-inducing agents with ageing and cancer [78-80]. Mice exposed to formaldehyde accumulate DPCs mostly in their bone marrow and develop squamous cell carcinomas in the nasal passages upon formaldehyde inhalation [81, 82]. DPC accumulation was also correlated with ageing in both mice and humans [83, 84]. However, there was no direct evidence that defective DPC repair could be pathological until the recent characterization of SPARTAN syndrome also known as Ruijs-Aalfs Syndrome (RJALS), which is caused by mutations in the *SPRTN* gene [60, 61]. SPARTAN syndrome is characterised by premature ageing and early-onset hepatocellular carcinoma. At the cellular level, cells from SPARTAN syndrome patients accumulate DPCs and are unable to cope with DPCs during DNA replication, leading to DSB formation during S-phase [4, 7, 61]. Therefore, SPARTAN syndrome constitutes the first direct link between a defective DPC repair and the development of cancer and ageing in humans [4].

Concluding remarks and future perspectives

Repair of DPCs was dogmatically considered to be solely under the jurisdiction of canonical DNA repair pathways like NER and HR. However, recent independent work from several laboratories demonstrates that a specialised DNA repair pathway, which strictly depends on proteolysis, repairs DPCs. We named this novel pathway DNA-protein crosslink proteolysis repair (DPC-PR). The proteases involved in DPC-PR repair are Wss1 in yeast and SPARTAN in metazoans. The discoveries of these two proteases and a human syndrome resulting from defective DPC repair establish this new DNA repair pathway as an essential mechanism for genome maintenance and protection from accelerated ageing and cancer in mammals (see Outstanding questions).

Acknowledgments

We wish to thank John Fielden and Joseph A. Newman for critical reading and feedback on this manuscript. The Ramadan laboratory is supported by the Medical Research Council-UK grant.

References

1 Ciccia, A. and Elledge, S.J. (2010) The DNA damage response: making it safe to play with knives. *Molecular cell* 40, 179-204

2 Jeggo, P.A., *et al.* (2016) DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer* 16, 35-42

3 Stingele, J. and Jentsch, S. (2015) DNA–protein crosslink repair. *Nature Reviews Molecular Cell Biology* 16, 455-460

4 Vaz, B., *et al.* (2016) Metalloprotease SPRTN/DVC1 Orchestrates Replication-Coupled DNA-Protein Crosslink Repair. *Molecular cell* 64, 704-719

5 Stingele, J., *et al.* (2014) A DNA-dependent protease involved in DNA-protein crosslink repair. *Cell* 158, 327-338

6 Stingele, J., *et al.* (2016) Mechanism and Regulation of DNA-Protein Crosslink Repair by the DNA-Dependent Metalloprotease SPRTN. *Molecular cell* 64, 688-703

7 Lopez-Mosqueda, J., *et al.* (2016) SPRTN is a mammalian DNA-binding metalloprotease that resolves DNA-protein crosslinks. *Elife* 5

8 Morocz, M., *et al.* (2017) DNA-dependent protease activity of human Spartan facilitates replication of DNA-protein crosslink-containing DNA. *Nucleic acids research*

9 Loeber, R.L., *et al.* (2009) Proteomic analysis of DNA-protein cross-linking by antitumor nitrogen mustards. *Chemical research in toxicology* 22, 1151-1162

10 Michaelson-Richie, E.D., *et al.* (2010) DNA-protein cross-linking by 1,2,3,4-diepoxybutane. *Journal of proteome research* 9, 4356-4367

11 Lai, Y., *et al.* (2016) Measurement of Endogenous versus Exogenous Formaldehyde-Induced DNA-Protein Crosslinks in Animal Tissues by Stable Isotope Labeling and Ultrasensitive Mass Spectrometry. *Cancer research* 76, 2652-2661

12 Ide, H., *et al.* (2015) Formation, Repair, and Biological Effects of DNA-Protein Cross-Link Damage. *Advances in DNA Repair*, 43-80

13 Kooistra, S.M. and Helin, K. (2012) Molecular mechanisms and potential functions of histone demethylases. *Nat Rev Mol Cell Biol* 13, 297-311

14 Trewick, S.C., *et al.* (2002) Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. *Nature* 419, 174-178

15 Swenberg, J.A., *et al.* (2011) Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment. *Toxicol Sci* 120 Suppl 1, S130-145

16 O'Brien, P.J., *et al.* (2005) Aldehyde sources, metabolism, molecular toxicity mechanisms, and possible effects on human health. *Crit Rev Toxicol* 35, 609-662

17 Heck, H. and Casanova, M. (2004) The implausibility of leukemia induction by formaldehyde: a critical review of the biological evidence on distant-site toxicity. *Regulatory toxicology and pharmacology : RTP* 40, 92-106

18 Tong, Z., *et al.* (2013) Accumulated hippocampal formaldehyde induces age-dependent memory decline. *Age* 35, 583-596

19 Halliwell, B. (1994) Free radicals, antioxidants, and human disease: Curiosity, cause, or consequence? *Lancet* 344, 721-724

20 Sczepanski, J.T., *et al.* (2013) Nucleosome core particle-catalyzed strand scission at abasic sites. *Biochemistry* 52, 2157-2164

21 DeMott, M.S., *et al.* (2002) Covalent trapping of human DNA polymerase beta by the oxidative DNA lesion 2-deoxyribonolactone. *The Journal of biological chemistry* 277, 7637-7640

22 Nakano, T., *et al.* (2003) DNA-protein cross-link formation mediated by oxanine: A novel genotoxic mechanism of nitric oxide-induced DNA damage. *Journal of Biological Chemistry* 278, 25264-25272

23 Tretyakova, N.Y., *et al.* (2015) DNA-Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Accounts of Chemical Research* 48, 1631-1644

24 Pommier, Y. (2009) DNA topoisomerase I inhibitors: chemistry, biology, and interfacial inhibition. *Chem Rev* 109, 2894-2902

25 Nitiss, J.L. and Nitiss, K.C. (2013) Tdp2: A Means to Fixing the Ends. *PLoS Genetics* 9, 2-4

26 Zhang, L., *et al.* (2004) Detecting DNA-binding of proteins in vivo by UV-crosslinking and immunoprecipitation. *Biochem Biophys Res Commun* 322, 705-711

27 Ide, H., *et al.* (2011) Repair and biochemical effects of DNA-protein crosslinks. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 711, 113-122

28 Minko, I.G., *et al.* (2002) Incision of DNA-protein crosslinks by UvrABC nuclease suggests a potential repair pathway involving nucleotide excision repair. *Proc Natl Acad Sci U S A* 99, 1905-1909

29 Nakano, T., *et al.* (2007) Nucleotide excision repair and homologous recombination systems commit differentially to the repair of DNA-protein crosslinks. *Molecular cell* 28, 147-158

30 de Graaf, B., *et al.* (2009) Cellular pathways for DNA repair and damage tolerance of formaldehyde-induced DNA-protein crosslinks. *DNA repair* 8, 1207-1214

31 Reardon, J.T. and Sancar, A. (2006) Repair of DNA-polypeptide crosslinks by human excision nuclease. *Proc Natl Acad Sci U S A* 103, 4056-4061

32 Baker, D.J., *et al.* (2007) Nucleotide excision repair eliminates unique DNA-protein cross-links from mammalian cells. *The Journal of biological chemistry* 282, 22592-22604

33 Nakano, T., *et al.* (2009) Homologous recombination but not nucleotide excision repair plays a pivotal role in tolerance of DNA-protein cross-links in mammalian cells. *The Journal of biological chemistry* 284, 27065-27076

34 Kumari, A., *et al.* (2012) Formaldehyde-induced genome instability is suppressed by an XPF-dependent pathway. *DNA repair* 11, 236-246

35 Speit, G., *et al.* (2000) Induction and repair of formaldehyde-induced DNA-protein crosslinks in repair-deficient human cell lines. *Mutagenesis* 15, 85-90

36 Krasich, R., *et al.* (2015) Functions that protect Escherichia coli from DNA-protein crosslinks. *DNA repair* 28, 48-59

37 Shaham, J., *et al.* (1997) DNA-Protein Crosslinks and Sister Chromatid Exchanges as Biomarkers of Exposure to Formaldehyde. *Int J Occup Environ Health* 3, 95-104

38 Nakano, T., *et al.* (2016) Radiation-Induced DNA–Protein Cross-Links: Mechanisms and Biological Significance. *Free Radical Biology and Medicine*

39 Hoa, N.N., *et al.* (2016) Mre11 Is Essential for the Removal of Lethal Topoisomerase 2 Covalent Cleavage Complexes. *Molecular cell* 64, 580-592

40 Deshpande, R.A., *et al.* (2016) Nbs1 Converts the Human Mre11/Rad50 Nuclease Complex into an Endo/Exonuclease Machine Specific for Protein-DNA Adducts. *Molecular cell* 64, 593-606

41 Aparicio, T., *et al.* (2016) MRN, CtIP, and BRCA1 mediate repair of topoisomerase II-DNA adducts. *The Journal of cell biology* 212, 399-408

42 Orta, M.L., *et al.* (2013) 5-Aza-2'-deoxycytidine causes replication lesions that require Fanconi anemia-dependent homologous recombination for repair. *Nucleic acids research* 41, 5827-5836

43 Rosado, I.V., *et al.* (2011) Formaldehyde catabolism is essential in cells deficient for the Fanconi anemia DNA-repair pathway. *Nature structural & molecular biology* 18, 1432-1434

44 Duxin, J.P., *et al.* (2014) Repair of a DNA-protein crosslink by replication-coupled proteolysis. *Cell* 159, 346-357

45 Debethune, L., *et al.* (2002) Processing of nucleopeptides mimicking the topoisomerase I-DNA covalent complex by tyrosyl-DNA phosphodiesterase. *Nucleic acids research* 30, 1198-1204

46 Ashour, M.E., *et al.* (2015) Topoisomerase-mediated chromosomal break repair: an emerging player in many games. *Nat Rev Cancer* 15, 137-151

47 Pommier, Y., *et al.* (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat Rev Mol Cell Biol* 17, 703-721

48 Interthal, H. and Champoux, J.J. (2011) Effects of DNA and protein size on substrate cleavage by human tyrosyl-DNA phosphodiesterase 1. *Biochem J* 436, 559-566

49 Reardon, J.T., *et al.* (2006) Repair of DNA-protein cross-links in mammalian cells. *Cell Cycle* 5, 1366-1370

50 Quievryn, G. and Zhitkovich, A. (2000) Loss of DNA-protein crosslinks from formaldehyde-exposed cells occurs through spontaneous hydrolysis and an active repair process linked to proteosome function. *Carcinogenesis* 21, 1573-1580

51 Mao, Y., *et al.* (2001) 26 S proteasome-mediated degradation of topoisomerase II cleavable complexes. *The Journal of biological chemistry* 276, 40652-40658

52 Lin, C.P., *et al.* (2008) A ubiquitin-proteasome pathway for the repair of topoisomerase I-DNA covalent complexes. *The Journal of biological chemistry* 283, 21074-21083

53 Ortega-Atienza, S., *et al.* (2015) Proteasome activity is important for replication recovery, CHK1 phosphorylation and prevention of G2 arrest after low-dose formaldehyde. *Toxicol Appl Pharmacol* 286, 135-141

54 Takeshita, T., *et al.* (2009) Perturbation of DNA repair pathways by proteasome inhibitors corresponds to enhanced chemosensitivity of cells to DNA damage-inducing agents. *Cancer Chemother Pharmacol* 64, 1039-1046

55 Chválová, K., *et al.* (2007) Mechanism of the formation of DNA-protein cross-links by antitumor cisplatin. *Nucleic acids research* 35, 1812-1821

56 Yeo, J.E., *et al.* (2014) Synthesis of site-specific DNA-protein conjugates and their effects on DNA replication. *ACS Chem Biol* 9, 1860-1868

57 Balakirev, M.Y., *et al.* (2015) Wss1 metalloprotease partners with Cdc48/Doa1 in processing genotoxic SUMO conjugates. *Elife* 4

58 O'Neill, B.M., *et al.* (2004) Coordinated functions of WSS1, PSY2 and TOF1 in the DNA damage response. *Nucleic acids research* 32, 6519-6530

59 Hiom, K. (2014) SPRTN is a new player in an old story. *Nature genetics* 46, 1155-1157

60 Ramadan, K., *et al.* (2016) Strategic role of the ubiquitin-dependent segregase p97 (VCP or Cdc48) in DNA replication. *Chromosoma*

61 Lessel, D., *et al.* (2014) Mutations in SPRTN cause early onset hepatocellular carcinoma, genomic instability and progeroid features. *Nature genetics* 46, 1239-1244

62 Hart, T., *et al.* (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* 163, 1515-1526

63 Wang, T., *et al.* (2015) Identification and characterization of essential genes in the human genome. *Science* 350, 1096-1101

64 Maskey, R.S., *et al.* (2014) Spartan deficiency causes genomic instability and progeroid phenotypes. *Nat Commun* 5, 5744

65 Biggins, S., *et al.* (2001) Genes involved in sister chromatid separation and segregation in the budding yeast Saccharomyces cerevisiae. *Genetics* 159, 453-470

66 Noguchi, C., *et al.* (2016) Genetic controls of DNA damage avoidance in response to acetaldehyde in fission yeast. *Cell Cycle*, 0

67 Stingele, J., *et al.* (2015) DNA-protein crosslink repair: proteases as DNA repair enzymes. *Trends in biochemical sciences* 40, 67-71

68 Iyer, L.M., *et al.* (2004) Novel predicted peptidases with a potential role in the ubiquitin signaling pathway. *Cell Cycle* 3, 1440-1450

69 Arnold, K., *et al.* (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22, 195-201

70 Davis, E.J., *et al.* (2012) DVC1 (C1orf124) recruits the p97 protein segregase to sites of DNA damage. *Nature structural & molecular biology* 19, 1093-1100

71 Ghosal, G., *et al.* (2012) Proliferating cell nuclear antigen (PCNA)-binding protein C1orf124 is a regulator of translesion synthesis. *The Journal of biological chemistry* 287, 34225-34233

72 Centore, R.C., *et al.* (2012) Spartan/C1orf124, a reader of PCNA ubiquitylation and a regulator of UV-induced DNA damage response. *Molecular cell* 46, 625-635

73 Juhasz, S., *et al.* (2012) Characterization of human Spartan/C1orf124, an ubiquitin-PCNA interacting regulator of DNA damage tolerance. *Nucleic acids research* 40, 10795-10808

74 Mosbech, A., *et al.* (2012) DVC1 (C1orf124) is a DNA damage-targeting p97 adaptor that promotes ubiquitin-dependent responses to replication blocks. *Nature structural & molecular biology* 19, 1084-1092

75 Kim, M.S., *et al.* (2013) Regulation of error-prone translesion synthesis by Spartan/C1orf124. *Nucleic acids research* 41, 1661-1668

76 Hudson, J.J., *et al.* (2012) SUMO modification of the neuroprotective protein TDP1 facilitates chromosomal single-strand break repair. *Nat Commun* 3, 733

77 Das, B.B., *et al.* (2014) PARP1-TDP1 coupling for the repair of topoisomerase I-induced DNA damage. *Nucleic acids research* 42, 4435-4449

78 Craft, T.R., *et al.* (1987) Formaldehyde mutagenesis and formation of DNA-protein crosslinks in human lymphoblasts in vitro. *Mutat Res* 176, 147-155

79 Wu, F.Y., *et al.* (2002) Association of DNA-protein crosslinks and breast cancer. *Mutat Res* 501, 69-78

80 Garaycoechea, J.I., *et al.* (2012) Genotoxic consequences of endogenous aldehydes on mouse haematopoietic stem cell function. *Nature* 489, 571-575

81 Conaway, C.C., *et al.* (1996) Formaldehyde mechanistic data and risk assessment: endogenous protection from DNA adduct formation. *Pharmacol Ther* 71, 29-55

82 Ye, X., *et al.* (2013) Inhaled formaldehyde induces DNA-protein crosslinks and oxidative stress in bone marrow and other distant organs of exposed mice. *Environ Mol Mutagen* 54, 705-718

83 Khokhlov, A.N., *et al.* (1986) [Strengthening of the DNA-protein complex during the stationary aging of cultured cells]. *Biull Eksp Biol Med* 101, 416-418

84 Zahn, R.K., *et al.* (1999) Assessment of DNA-protein crosslinks in the course of aging in two mouse strains by use of a modified alkaline filter elution applied to whole tissue samples. *Mech Ageing Dev* 108, 99-112

85 Horton, J.K., *et al.* (2015) DNA polymerase beta-dependent cell survival independent of XRCC1 expression. *DNA repair* 26, 23-29

86 Prasad, R., *et al.* (2014) Suicidal cross-linking of PARP-1 to AP site intermediates in cells undergoing base excision repair. *Nucleic acids research* 42, 6337-6351

87 Costa, M., *et al.* (1997) Dna-Protein Cross-Links Produced By Various Chemicals in Cultured Human Lymphoma Cells. *Journal of Toxicology and Environmental Health* 50, 433-449

88 Barker, S., *et al.* (2005) DNA-protein crosslinks: Their induction, repair, and biological consequences. *Mutation Research - Reviews in Mutation Research* 589, 111-135

89 Voulgaridou, G.-P., *et al.* (2011) DNA damage induced by endogenous aldehydes: current state of knowledge. *Mutation research* 711, 13-27

90 Ide, H., *et al.* (2008) Repair of DNA-protein crosslink damage: coordinated actions of nucleotide excision repair and homologous recombination. *Nucleic Acids Symp Ser (Oxf)*, 57-58

91 Gómez-Herreros, F., *et al.* (2013) TDP2-Dependent Non-Homologous End-Joining Protects against Topoisomerase II-Induced DNA Breaks and Genome Instability in Cells and In Vivo. *PLoS Genetics* 9

Glossary

DNA damage repair (DDR): commonly referred to canonical DDR pathways: base excision repair (BER), nucleotide excision repair (NER), and mismatch repair (MMR) which repair single strand DNA damage and homologous recombination (HR) and non-homologous endjoining (NHEJ) which repair DNA double strand breaks.

DNA damage tolerance (DDT): a mechanism of bypassing DNA lesions during DNA replication, thus allowing replication forks to progress, leaving the lesions to be repaired later.

DNA-protein crosslink (DPC): any protein that is irreversibly covalently attached to the DNA.

Homologous recombination (HR): an error-free canonical double strand break repair pathway which acts during S-phase.

Nucleotide excision repair (NER): a canonical DNA damage repair pathway that excises DNA lesions on one DNA strand and is not cell cycle-specific.

SPARTAN Syndrome also known as Ruijs-Aalfs syndrome (RJALS): a monogenic human syndrome with mutations in *SPRTN* gene characterised by premature aging phenotypes and early-onset hepatocellular carcinoma.

SPRTN/DVC1: a DNA-dependent metalloprotease that proteolytically digests the proteinaceous part of DPCs during DNA replication fork progression in metazoans.

Wss1: a DNA-dependent metalloprotease that proteolytically digests the proteinaceous part of DPCs during S-phase in yeast.

Translesion DNA synthesis (TLS): a DDT pathway involving specialised DNA polymerases that can bypass unrepaired DNA lesions during replication. Some of the TLS polymerases are error-prone and thus TLS can be mutagenic.

Topoisomerase1 and 2 cleavage complexes (TOP1ccs and TOP2ccs): TOP1 or TOP2 which are covalently trapped to the DNA next to a single strand (ss) or a double strand (ds) DNA break, respectively. The strand breaks near TOPccs were created by the enzymes themselves during their normal enzymatic cycles.

Text Boxes

Box 1. Types of DPCs

DPCs are intricately complex DNA lesions. The complexity is a result of the diversity of crosslink formations in terms of protein size and physicochemical properties (charge, level of disorder), type of crosslink (crosslinked to one or both intact DNA strands) and number of covalent bonds (stability and structure of DPC), as well as temporal distribution of crosslink occurrence (cell cycle dependence). DPCs are commonly divided into two groups: (A) general, non-enzymatic DPCs (Type 1) which include any protein found in proximity to DNA at the moment of exposure to endogenous or exogenous DPC-inducing agents and (B) specific, enzymatic DPCs (Type 2) such as Topoisomerase 1 and 2 cleavage complexes (TOP1-ccs and TOP2-ccs), DNA polymerase (Pol) β and poly(ADP-ribose) polymerase 1 (PARP1) crosslinks. During their normal enzymatic cycles, these enzymes reversibly bind to DNA. However, upon treatment with anti-cancer drugs, camptothecin and/or etoposide, TOP1ccs and TOP2ccs are formed [24, 25], while abortive DNA repair can lead to Pol β and PARP1 crosslink formation [38]. Type 1 DPCs are the most prevalent under physiological conditions and they always include proteins crosslinked to an undisrupted DNA strand(s) [12]. Conversely, Type 2 DPCs include proteins crosslinked to a broken DNA backbone. TOP1 is crosslinked to the 3' end of a ssDNA break, TOP2 to the two 5' ends of a dsDNA break [46], while PARP1 is crosslinked to the 3' end of a ssDNA break resulting from deficient DNA repair [85, 86]. In the case of topoisomerases, the breaks are created during the normal enzymatic decatenation reactions of topoisomerases and persist after crosslinking. The chemistry of both types of DPCs is explained in BOX 2. Altogether, this highlights that genomic DNA is constantly exposed to various DPCs, induced either by endogenous and

exogenous sources. Indeed, DPCs are among the most common DNA lesions. Thus, cells have to have specialised DPC recognition and repair mechanisms to prevent DPC-induced genotoxicity.

Box 2. Chemistry of DPC formation

Aldehydes react with proteins in the vicinity of DNA resulting in the formation of DPCs and base adducts [87-89] and to a lesser extent intra- and inter-strand crosslinks (ICLs), SSBs and DSBs [90]. Firstly, aldehydes react with lysine, cysteine and histidine residues on the protein forming a protein adduct which then reacts with the amino group on DNA bases resulting in crosslinks of variable stability. Target sites for protein crosslinking on DNA molecule include: N7 of guanine, C-5 methyl group of thymine and the exocyclic amino groups of guanine, cytosine and adenine [23], thus further expanding the complexity and diversity of crosslinks. Reactive oxygen and nitrogen species (ROS and RNS) can react with DNA (guanine, cytosine and thymine bases) and/or proteins (lysine and tyrosine side chains) resulting in the formation of free radicals or electrophilic lesions, which in turn react with another protein and DNA molecule thus creating the crosslink. For detail review on the chemistry of DPCs we refer the readers to Tretyakova et al. [23].

TOP1 and 2 relax DNA supercoils during DNA transactions (replication, transcription, recombination and repair) by decatenating DNA through single or double strand incisions, respectively, and re-ligating the nicked ends. The chemistry behind it is as follows: the catalytic tyrosine residue in topoisomerase becomes transiently covalently attached to the DNA phosphate at the 3' (TOP1) or 5' (TOP2) end of the broken DNA as a result of nucleophilic attack by the catalytic tyrosine on

the DNA phosphodiester bond. Normally, this topoisomerase-DNA covalent reaction intermediate would then dissociate and topoisomerase would re-ligate the broken DNA ends. However, camptothecin (TOP1) and etoposide (TOP2) bind to the catalytic site at the enzyme-DNA interface, thus preventing re-ligation and leaving the enzyme trapped to DNA and the strand break/s unsealed [24, 25, 91]. Topoisomerases can also become trapped to the DNA backbone during their normal enzymatic cycles if in proximity to oxidative radicals or upon encountering bulky DNA adducts, ribonucleotides or abasic sites [46]. Given the high demand for DNA relaxation by topoisomerases during all DNA transactions, it is expected that the number of topoisomerases bound to DNA is high. Indeed, it was recently confirmed that the incidence of endogenous TOP2ccs is much higher than previously thought and that they cause severe genomic instability if left unrepaired [39].

Box 3. The repair of Topoisomerase 1 and 2 cleavage complexes (TOP1- and 2-ccs)

The repair of TOP-ccs (enzymatic DPCs) has been extensively studied in recent decades due to their importance for cancer therapy. Many therapeutic approaches rely on the inhibiton of topoisomerases through the action of topoisomerase poisons, camptothecin and etoposide, which induce TOPccs with the aim to stop cancer proliferation and/or induce cancer cell death. Indeed, therapy for more than 30-50% of cancers relies on the use of topoisomerase inhibitors [46]. Until very recent advances, it was thought that TOP1cc and TOP2cc repair relies mainly on the action of phosphodiesterase enzymes, tyrosyl-DNA-phosphodiesterase 1 and 2 (TDP1 and TDP2), respectively. These enzymes act on the DNA backbone by excising small

peptide remnants left from TOPccs, thus leaving ss (TDP1) or ds breaks (TDP2) which are subsequently repaired by canonical DDR pathways. More specifically, TDP2 cleaves DNA by hydrolysing 5' tyrosine phosphodiester bonds which converts them into 5'phosphate ends which are subsequently repaired by NHEJ [91]. Although the mechanism of action of TDPs was known, the upstream factors involved in TOPcc repair remained unknown. More specifically, how TOPccs are cleaved into smaller peptides, thus enabling TDPs to act was only resolved recently when the protease SPRTN was found to remove TOPccs in vitro and in vivo [4, 7]. SPRTN proteolytically digests TOP1 and TOP2 thus reducing the size of the DPCs and enabling the action of TDPs. Indeed, it was shown that SPRTN and TDP1 act epistaticaly in the repair of TOP1ccs [4]. Recently, a separate mechanism for TOP2cc repair was discovered, one which relies on the action of the nuclease Mre11, otherwise known as a part of the MRN complex with a main role in HR [39-41]. Mre11 removes TOP2ccs independently of HR, while NHEJ acts downstream of Mre11 activity to repair ds breaks left after nuclease action [39]. Additionally, Mre11-mediated TOP2cc removal is a dominant and separate pathway to TDP2-mediated TOP2cc repair [39]. In light of these recent discoveries it is possible that SPRTN and TDP1 and TDP2 form a separate pathway for TOPcc repair, distinct from Mre11-mediated TOP2cc removal. However, we cannot exclude the possibility that SPRTN is also needed in Mre11-mediated TOP2cc removal to reduce the size of the proteinacious part of the DPC. The precise coordination of these two pathways remains to be determined.

Figure legends

Figure 1. DNA repair pathways. Schematic model of various DNA lesions caused by different genotoxic agents and respective DNA repair pathways. Specialised DNA repair pathways cope with specific type of DNA lesions. DNA lesions caused by the covalent attachment of bulky protein, DNA-protein crosslink (DPC) are predominantly repaired by DNA-protein crosslink proteolysis repair (DPC-PR). DDR; DNA damage response, DDT; DNA damage tolerance, BER; base excision repair, NER; nucleotide excision repair, HR; homologous recombination, NHEJ; non-homologous end joining, MMR; mismatch repair, FA; Fanconi Anemia, TLS; translesion DNA synthesis, DPC; DNA-protein crosslink.

Figure 2. (A) Comparison of NER, HR and DPC-PR in respect to repair of DPCs. Upper panel; Colour code represents activity of three known repair pathways for DPC repair during the cell cycle. Green = NER, Blue = HR and Violet = DPC-PR. Lower panel: DPC repair pathways and their cellular characteristics and consequences. Mutations in these pathways cause several inborn diseases; XP; Xeroderma Pigmentosum, TTD; Trichothiodystrophy, CS; Cockayne Syndrome, ATLD; Ataxia-teleangiectasia-like disorder, NBS; Nijmegen breakage syndrome, SPARTAN syndrome (also known as Ruijs-Aalfs) **(B) SPRTN is a central player in replication-coupled proteolysis repair pathway of DPCs.** SPRTN proteolytically digests the proteinaceous part of bulky DPCs which block DNA replication fork progression. After digesting the DPC, SPRTN inactivates itself by self-cleavage.

Figure 3. (A) Evolutionary analysis of SPRT and WLM families. Alanyl aminopeptidase family, another member of the gluzincins subgroup of zinzin

metalloproteases, was used to compare the evolutionary relationship between the SPRT and WLM families. SPRT is equally distant to the WLM group as it is to Alanyl aminopeptidase group, thus there is no indication of common ancestry between the SPRT and WLM families. The SPRT family is present in bacteria, archaea, cyanobacteria, plants and animals, while the WLM family emerged in yeast, fungi and plants. Accession numbers of sequences used in the analysis, methodology for phylogenetic analysis as well as an expanded phylogenetic tree including branch support values are available in the supplementary material. **(B) Domain organization of SPRTN and Wss1.** SPRTN and Wss1 share a short consensus HEXXH motifs in the active centre of the protease, a common property of all zinzin metalloproteases, and a p97/Cdc48 binding motif (SHP and VIM, in blue). C-terminal regions (downstream of SPRT and/or WLM domain) differ substantially between the proteins, with the long C-terminal arm of SPRTN (276 amino acids) bearing a PCNA binding (PIP, in light blue) and ubiquitin binding motifs (UBZ, in orange) and the short C-terminal arm of Wss1 (48 amino acids) bearing sumo interaction motifs (SIM, in red). **(C) Homology model of human SPRT domain.** The human SPRT domain was modelled according to yeast Wss1b (left panel) (5JIG) and another gluzincin member, abylysin (right panel) (4JIU) using the SWISS-MODEL workspace (see supplementary methods). Both models show a conserved protease core of SPRTN with two α-helices (in green) that contain the catalytic active centre, including three zinc binding histidines (in red) and a glutamic acid (in blue). Abylysin-based model gave broader coverage of SPRTN domain (66. – 142. amino acids of full length SPRTN) and higher model confidence (see supplementary methods) compared to the Wss1b-based model coverage (103.-144. amino acids), thus confirming that SPRTN is equally structurally similar to Wss1 as to any other gluzincin protein. Three β-

sheets (orange) upstream of the catalytic core were modeled with high confidence according to the abylysin structure (right panel), while two β-sheets between the two α-helices in the Wss1b-based model are modeled with to low confidence in order to be considered reliable (see supplementary material).

Outstanding questions

Which factors act downstream of SPRTN proteolysis? How is the whole DPC repair pathway orchestrated and the choice between different downstream pathways made, e.g. NER, HR. Mre11, TDPs and TLS?

How are DPCs removed in non-cycling, lowly proliferative cells, where they pose a threat to transcription progression?

How are TOP1ccs and TOP2ccs predominantly repaired, via SPRTN cleavage followed by TDP-mediated peptide excision or via Mre11 through the excision of ssDNA overhang bearing the TOP2cc? Can Mre11 also act in the removal of other DPCs in vivo?

How is SPRTN protease regulated? What is the signal to trigger SPRTN-mediated proteolysis and how is its pleiotropic protease activity kept in check to prevent unspecific substrate cleavage?

What is the exact mechanism behind the link between DPC accumulation and the carcinogenesis and premature aging observed in SPARTAN syndrome patients?

Figure 1

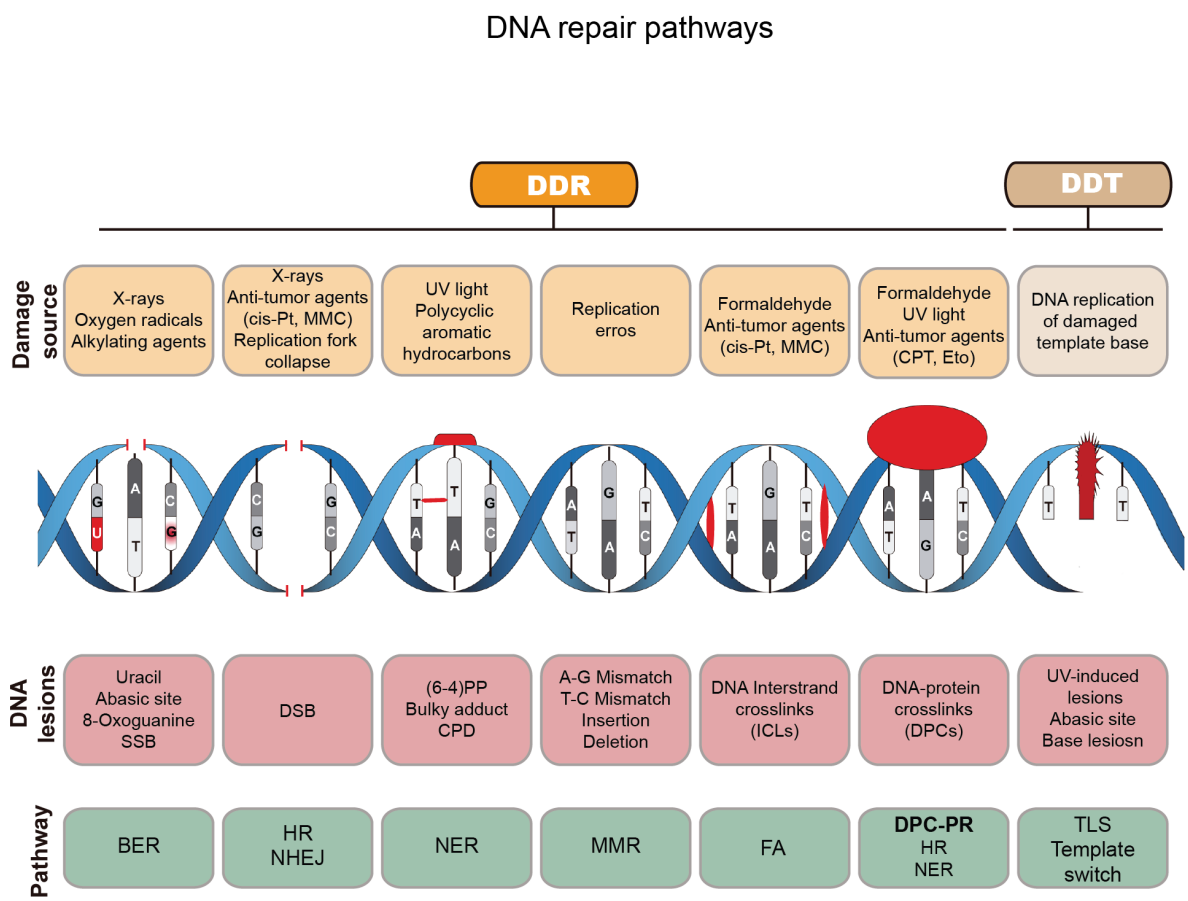


Figure 1

Figure 2

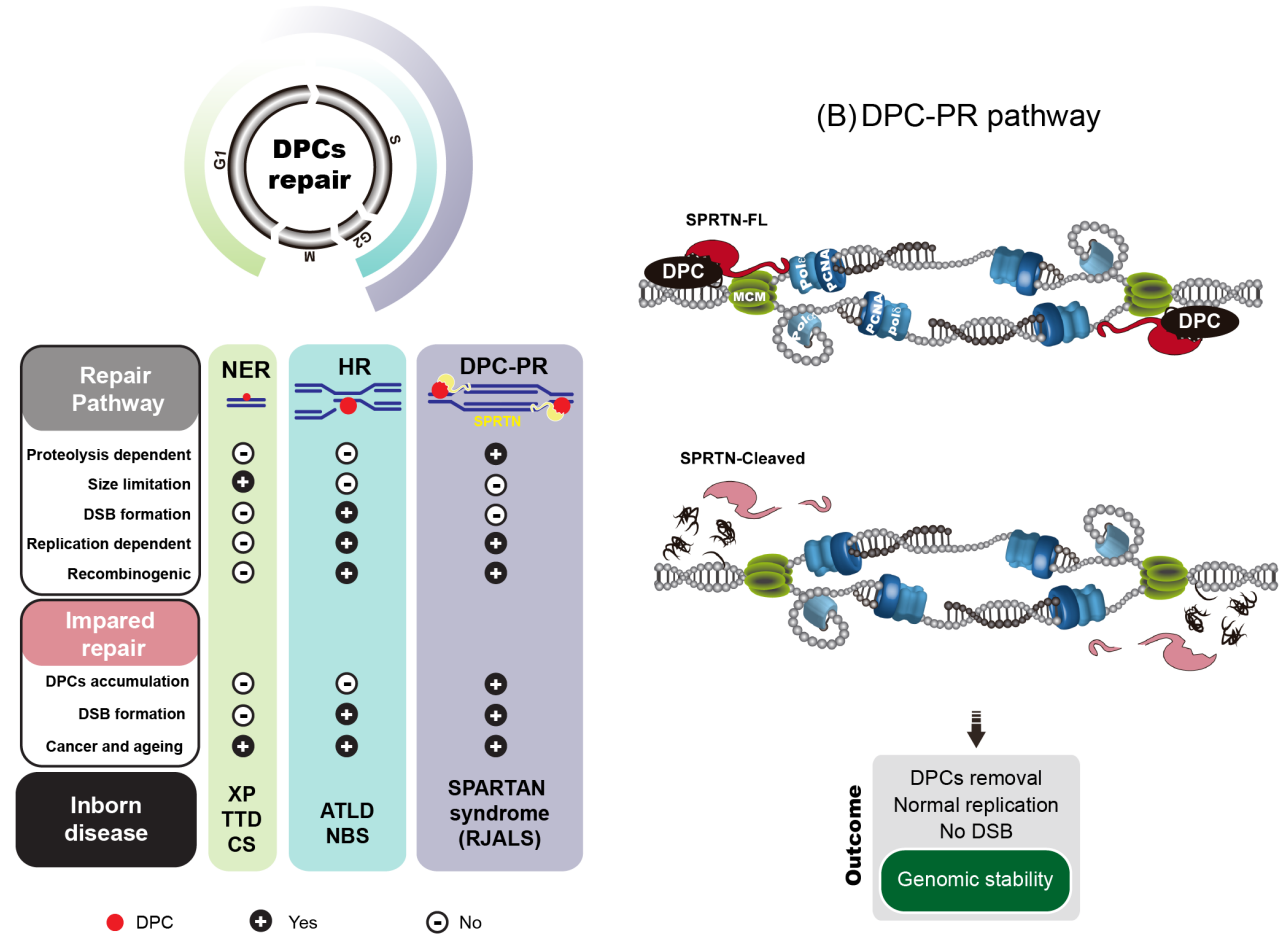
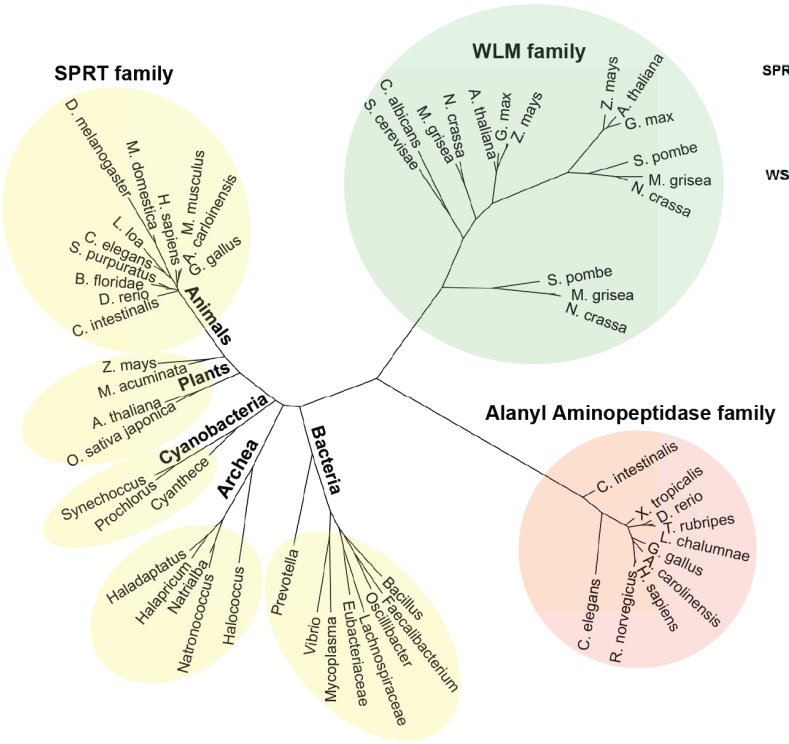


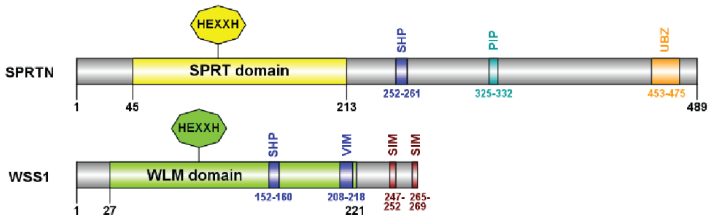
Figure 2

Figure 3

(A) Evolutionary tree



(B) Domain organization



(C) Homology models of SPRT domain

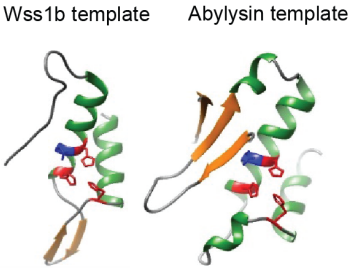


Figure 3

Inventory of Supplementary Information

1. Supplementary methods include methodology used to construct phylogenetic tree in Figure 3A and to perform homology modelling of the SPRT domain of human SPRTN based on the newly published structure of yeast Wss1b and zinzin metallopeptidase abylysin shown in Figure 3C.

2. Figure S1 contains phylogenetic tree with branch support values and full protein names which is complementary to phylogenetic tree shown in Figure 3A. Related to Figure 3A.

3. Table S1 contains protein accession codes from NCBI database of sequences used to construct phylogenetic tree in Figure 3A. Related to Figure 3A.

4. Figure S2 contains multiple sequence alignment of SPRT and WLM domains across species. Related to Figure 3C.

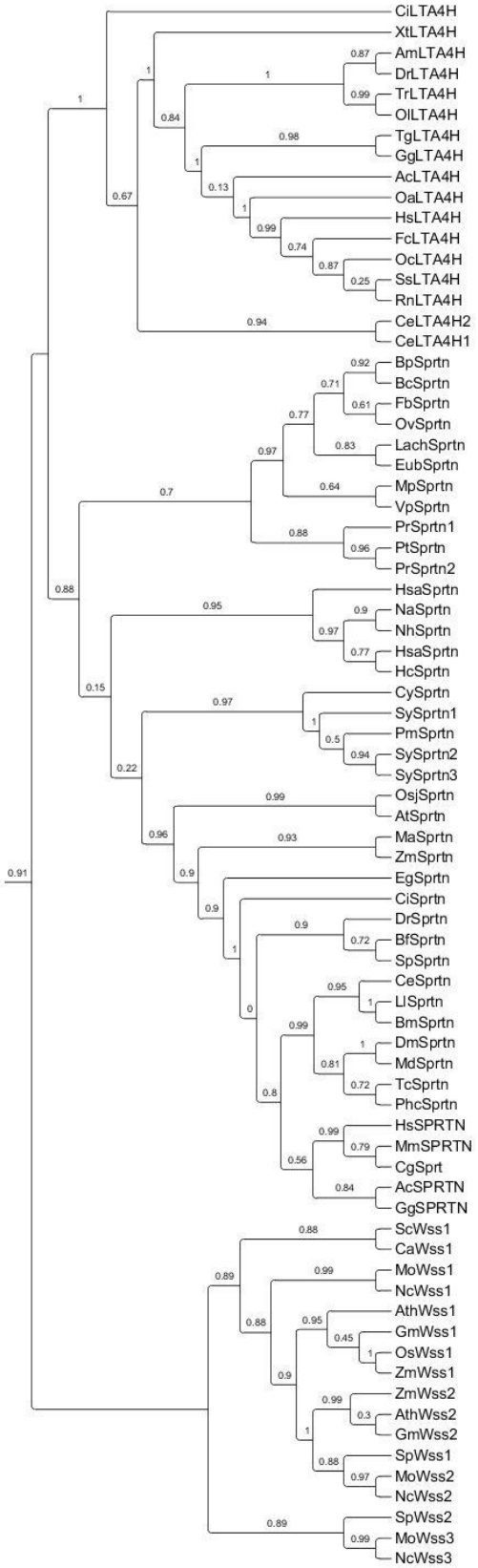
5. Supplementary references

1. Supplementary methods

PSI-BLAST (Position-Specific Iterated Blast) was used to identify orthologs of SRTN in bacteria, archaea, cyanobacteria, yeast, fungi, plants and animals by blasting the SPRT domain protein sequence of human SPRTN through the NCBI database (National Center for Biotechnology Information) [1]. The same approach was used to identify WLM domain-containing proteins using the WLM domain protein sequence from S. cerevisiae Wss1. Alanyl aminopeptidases (leukotriene A-4 hydrolase) protein sequences were downloaded from NCBI, and M1-LTA4H domains containing a HEXXH protease core were used for alignment and tree construction. Multiple sequence alignments were done using MAFFT [2]. Quality of alignment was estimated with Guidance software (alignment score was 0.532896) [3]. Phylogenetic tree was constructed using Maximum Likelihood method in PhyML 3.0.1 software (LG model, 10 rate categories, best of NNI and SPR for tree searching operations) [4]. Confidence of nodes was estimated by approximate likelihood ratio test (aLRT) [5]. Homology modelling of human SPRT domain was done using the crystal structure of Wss1b in fission yeast (Schizosaccharomyces pombe) (PDB: 5JIG) as a

template or using zinzin protease, abylysin (PDB: 4JIU) in the SWISS-MODEL workspace [6, 7]. Confidence of the models was estimated using the QMEAN scoring function [8].

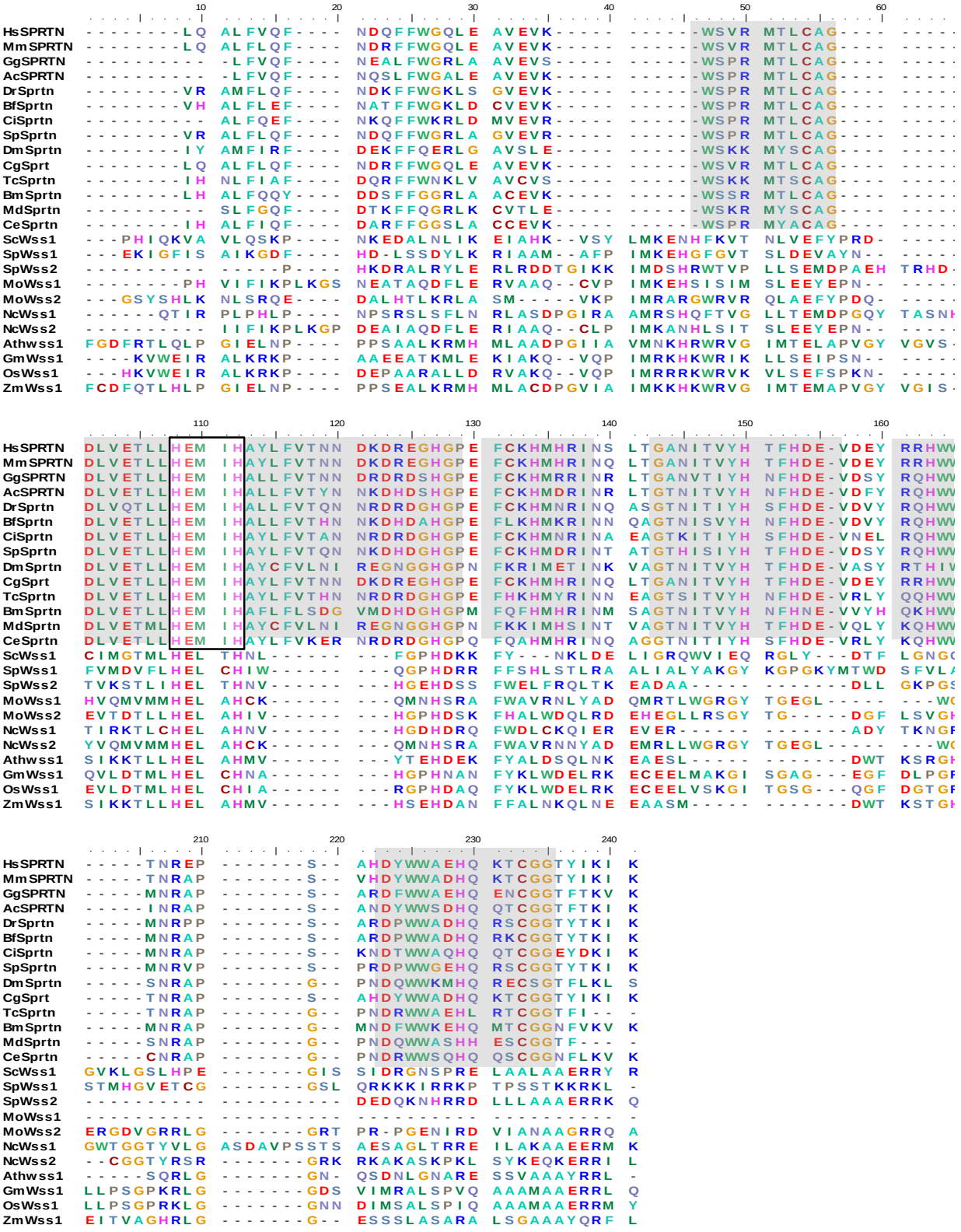
2. Figure S1. Phylogenetic tree with branch support values and full protein names which is complementary to phylogenetic tree shown in Figure 3A.



3. Table S1. Protein accession codes for sequences used to construct phylogenetic tree in Figure 3A.

Abbreviation	Accession	Species
NP_000886.1	HsLTA4H	Homo sapiens
XP_002711258.1	OcLTA4H	Oryctolagus cuniculus
NP_001025202.1	RnLTA4H	Rattus norvegicus
XP_011282463.1	FcLTA4H	Felis catus
XP_005664289.1	SsLTA4H	Sus scrofa
XP_001509819.2	OaLTA4H	Ornithorhynchus anatinus
NP_001006234.1	GgLTA4H	Gallus gallus
XP_002189060.1	TgLTA4H	Taeniopygia guttata
XP_003221111.1	AcLTA4H	Anolis carolinensis
NP_001006898.1	XtLTA4H	Xenopus tropicalis
NP_998451.1	DrLTA4H	Danio rerio
XP_004084996.1	OlLTA4H	Oryzias latipes
XP_003976424.1	TrLTA4H	Takifugu rubripes
XP_007232319.1	AmLTA4H	Astyanax mexicanus
XP_002123481.1	CiLTA4H	Ciona intestinalis
NP_001023056.1	CeLTA4H1	Caenorhabditis elegans
NP_500385.1	CeLTA4H2	Caenorhabditis elegans
NP_114407.3	HsSPRTN	Homo sapiens
NP_001104611.1	MmSPRTN	Mus musculus
XP_003496855	CgSprt	Cricetulus griseus
XP_419571.4	GgSPRTN	Gallus gallus
XP_008122534.1	AcSPRTN	Anolis carolinensis
XP_005173863.1	DrSprtn	Danio rerio
XP_002610684	BfSprtn	Branchiostoma floridae
XP_002125958.1	CiSprtn	Ciona intestinalis
XP_786958	SpSprtn	Strongylocentrotus purpuratus
NP_573032	DmSprtn	Drosophila melanogaster
XP_972573	TcSprtn	Tribolium castaneum
XP_002428589	PhcSprtn	Pediculus humanus corporis
XP_00314602	LlSprtn	Loa loa
XP_001896364	BmSprtn	Brugia malayi
XP_011295796	MdSprtn	Musca domestica
XP_006846650	AtSprtn1	Amborella trichopoda
NP_505853.1	CeSprtn	Caenorhabditis elegans
XP_010909703	EgSprtn	Elaeis guineensis
XP_009388648.1	MaSprtn	Musa acuminata
XP_015611787.1	OsjSprtn	Oryza sativa Japonica
XP_008652777.1	ZmSprtn	Zea mays
WP038014267.1	SySprtn1	Synechoccus sp.
WP038024487.1	SySprtn2	Synechoccus sp.
WP038542109.1	SySprtn3	Synechoccus sp.
WP_011827472.1	PmSprtn	Prochlorococcus marinus
WP_012626670.1	CySprtn	Cyanthece sp.
CDE33259.1	PrSprtn1	Prevotella sp.
CDB05322	PrSprtn2	Prevotella sp.
WP_028900949.1	PtSprt	Prevotella timonensis
WP_025810455.1	BpSprtn	Bacillus paralicheniformis
CUB08940	BcSprtn	Bacillus cereus
CZT57581.1	EubSprtn	EubacteriaceaeDUF45
WP_014116131.1	OvSprtn	Oscillibacter valericigenes
CBL02948.1	FpSprtn	Faecalibacterium prausnitzii
WP_051605842.1	LachSprtn	LachnospiraceaeDUF45
WP_027123856.1	MpSprtn	Mycoplasma pirum
WP_053809052.1	VpSprtn	Vibrio parahaemolyticus
WP_005041369.1	HsaSprtn	Halococcus salifodinae
WP_049972315.1	HcSprtn	Haladaptatus cibarius
WP_005553854.1	NaSprtn	Natronococcus amylolyticus
WP_049993186.1	HsalSprtn	Halapricum salinum
WP_006653888.1	NhSprtn	Natrialba hulunbeirensis
NP_593097	SpWss2	Saccharomyces pombe
NP_588321	SpWss1	Saccharomyces pombe
EDN62373	ScWss1	Saccharomyces cerevisiae
XP003710670	MoWss3	Magnaporthe oryzae
XP_003714265	MoWss1	Magnaporthe oryzae
XP_003709211	MoWss2	Magnaporthe oryzae
NcXP_964016	NcWss1	Neurospora crassa
XP_963623	NcWss2	Neurospora crassa
XP_959170	NcWss3	Neurospora crassa
CaXP_720120	CaWss1	Candida albicans
AtBAB09266	AtWss2	Arabidopsis thaliana
AtNP_001321636	AtWss1	Arabidopsis thaliana
XP_014626678	GmWss1	Glycine max
XP_014628122	GmWss2	Glycine max
XP_015643443	OsWss1	Oryza sativa
NP_001143131	ZmWss1	Zea mays
XP_008659182	ZmWss2	Zea mays

4. Figure S2. Multiple sequence alignment of SPRT domain and WLM domain across species Conserved regions specific for SPRT domains are shown in grey, active site motif HEXXH is shown in a black frame.



5. Supplementary References

1 Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, 3389-3402

2 Katoh, K., *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30, 3059-3066

3 Penn, O., *et al.* (2010) GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic acids research* 38, W23-28

4 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52, 696-704

5 Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* 55, 539-552

6 Arnold, K., *et al.* (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22, 195-201

7 Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714-2723

8 Benkert, P., *et al.* (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27, 343-350